HHS Public Access

Author manuscript

Nat Struct Mol Biol. Author manuscript; available in PMC 2018 April 02.

Published in final edited form as:

Nat Struct Mol Biol. 2017 November ; 24(11): 1000–1006. doi:10.1038/nsmb.3474.

Nuclear topology modulates the mutational landscapes of cancer genomes

Kyle S. Smith[1,2,a], **Lin L. Liu**[3,a], **Shridar Ganesan**[1], **Franziska Michor**[3,*], and **Subhajyoti De**[1,*]

[1]Rutgers Cancer Institute of New Jersey, New Brunswick, NJ 08901, USA

[2]Department of Pharmacology, University of Colorado-Denver, Aurora 80025, 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, MA 02115, USA

Abstract

Nuclear organization of genomic DNA affects DNA damage and repair processes, and yet its impact on mutational landscapes in cancer genomes remains unclear. Here we analyzed genome-wide somatic mutations from 366 samples of 6 cancer types. We found that lamina-associated regions, which are typically localized at the nuclear periphery, displayed higher somatic mutation frequencies compared to the inter-lamina regions at the nuclear core. This effect remained even after adjusting for features such as GC%, chromatin, and replication timing. Furthermore, mutational signatures differed between the nuclear core and periphery, indicating differences in the patterns of DNA damage and/or DNA repair processes. For instance, smoking and UV-related signatures were more enriched in the nuclear periphery. Substitutions at certain motifs were also more common in the nuclear periphery. Taken together, we found that the nuclear architecture influences mutational landscapes in cancer genomes beyond the effects already captured by chromatin and replication timing.

Emerging evidence indicates that somatic mutations in cancer genomes are non-randomly distributed, 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 DNA damage and/or repair processes[3,5,12,13], and has implications for identification of potential cancer driver genes[14] and clinical management of cancer patients, e.g. radio-sensitivity and immunotherapy[15]. However, the factors identified so far do not explain the entire extent of regional variation of the mutational burden in cancer genomes, suggesting that other factors are yet to be identified.

*Authors for correspondence. michor@jimmy.harvard.edu, subhajyoti.de@rutgers.edu.
[a]These authors have contributed equally to this work.

AUTHOR CONTRIBUTIONS

SD conceived the project with FM. KS, LLL, FM, SD designed the experiments. KS, LLL, SD performed the experiments. KS, LLL, SG, FM, SD interpreted the results. FM, SD wrote the manuscript with input from other authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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 DNA damage and repair processes[22–24]. For instance, the nuclear lamina-associated regions are refractory to homologous recombination-mediated repair and utilize an error-prone alternative end-joining mechanism to repair DNA double strand breaks[25]. Oct-1 and p53 dependent pathways link lamin functions to oxidative stress response[26]. Indeed, a previous multivariate analysis suggests that nuclear lamina association significantly contributes to germ line mutation rate variation[27]. Furthermore, it was recently 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 modulates the somatic mutational landscapes in cancer genomes, and that its effects might go beyond the variations due to known covariates such as chromatin domains and DNA replication timing[4,6].

To test these hypotheses, we obtained somatic point mutation data from 366 completely sequenced genomes of 6 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], enabling us to identify effects of nuclear localization on somatic mutational patterns across diverse cancer types. We focused on the noncoding, non-repetitive, non-conserved 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 (see **Online Methods** for details). We denoted the mutation detection frequency per base pair in these regions, when normalized by the mutation detection frequency per base pair in the genome, as adjusted mutation rate (AMR).

First, we investigated whether nuclear localization of chromosomes correlates with their AMR. We used chr18 and chr19 as classic examples since it has long been known that human chr18 is preferentially localized close to the nuclear periphery, while chr19 is primarily at the nuclear core[38] (Figure 1A). Indeed, the AMR for chr18 was significantly higher compared to that for chr19 across all 6 cancer types analyzed (Figure 1B; Mann Whitney U test p-value < 1e-02 for all cohorts). Integrating paired copy number data when available (e.g. 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 using 3D FISH (fluorescence in situ hybridization), we observed a similar association between the overall nuclear positioning of chromosomes and their AMR – those that are predominantly in the nuclear periphery have a higher AMR compared to those in the core (Figure 1C). The

coefficient of determination was weak (< 0.1) in all cohorts, which was, at least partly, due to the fact that chromosomes are large nuclear entities that typically span multiple nuclear domains; i.e. some parts of the same chromosome could be localized at the periphery while other parts at the relative interior of the nucleus[38]. Therefore, we chose to investigate whether more precise measures of the nuclear localization of genomic regions within and across chromosomes would be able to explain to the observed differences in chromosome-level variations in AMR.

We obtained chromatin immuno-precipitation data for the lamin-family proteins Lamin A and B1 (Figure 2A), and classified a region as constitutively in the nuclear periphery if the region was associated with lamins in all cell types examined; conversely, a region was categorized as constitutively in the nuclear core if it did not overlap with lamin-associated domains in any of the cell types analyzed (Figure 2B). As before, we prioritized non-coding, non-repetitive, non-conserved segments of genomic regions, which are constitutively at the periphery (constitutive lamina-associated domains; *cLAD*) and core (inter-lamina-associated domains; *iLAD*), respectively. We then integrated somatic mutation data from each cancer cohort and calculated the AMR for these two types of regions for each sample. We found that the AMR for *cLADs* was significantly higher compared to that for *iLAD*s, and once again, this observation was consistent across all 6 cancer cohorts (Figure 2C; Mann Whitney U test p-value < 1e-05 for all cohorts). Within respective chromosomes, cLAD and iLAD regions displayed a systematic difference in their AMR, 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 *cLAD*s and *iLAD*s that have evolutionarily conserved patterns of nuclear localization, after integrating data on lamina-associated regions from multiple cell types in mouse (see **Online Methods** for details), and found similar results (Supplementary Fig. 3). Therefore, our findings are not sensitive to our choice of definition for *cLAD*s and *iLAD*s, and indicate that lamina-associated regions localized at the periphery have higher somatic mutation frequencies compared to the inter-lamina 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>T substitutions, including those in the pi-pyrimidine context, were proportionally more common in the cLADs compared to the iLADs (Mann Whitney U test p-value < 1e-08) (Figure 2D; Supplementary Fig. 4). Similarly, C>A substitutions displayed a higher enrichment in the cLADs in the LUSC cohort, indicating that smoking-associated oxidative DNA damage was greater in the nuclear periphery compared to the nuclear core (Mann Whitney U test p-value < 1e-02, Figure 2D; Supplementary Fig. 4). For LUSC patients, data on their smoking history and the number of pack years was available. We calculated the AMR for cLAD and iLAD considering only C:G>A:T mutations, and plotted AMR(cLAD)/AMR(iLAD) against the number of pack years smoked. Indeed, we found that the number of pack years smoked was weakly correlated with the AMR(cLAD)/AMR(iLAD) ratio (Spearman correlation coefficient: 0.29; Supplementary Fig. 4), suggesting that the relative strength of the signature of oxidative

damage induced by smoking in the nuclear periphery was higher for heavy smokers compared to light smokers. Therefore, in cancer types driven by external carcinogens, the nuclear periphery had a proportionally higher burden of corresponding mutation signatures.

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 (Figure 2D; Supplementary Fig. 4). For instance, when we summarized the tri-nucleotide substitution patterns into mutational signatures using non-negative matrix factorization[40], mutation signatures 3 and 5 had a proportionally larger contribution in the iLAD and cLADs, respectively, in most cancer types as compared to the other signatures. Translating the mutational signatures into substitution patterns, it became clear that a majority of the cancer types had a proportional increase in the contribution of mutations in the WNW context (W: A or T; N: A, G, C, or T) in the cLADs at the periphery compared to the iLADs in the core. Different cancer types, however, showed subtle variation in the preference for specific sub-motifs; for instance, in the DLBCL and CLL cohorts, W[T>G]W and also W[T>C]W mutations were relatively more common in the cLADs than in the iLADs (Mann Whitney U test p-value < 1e-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>T]W substitutions were more common in the cLADs relative to the iLADs (Mann Whitney U test p-value < 1e-08; Supplementary Fig. 4).

Nuclear localization of genomic DNA is coupled with many genomic and epigenomic features: regions in the nuclear periphery tend to be, on average, AT-rich, gene-poor, more heterochromatic, and have late replication timing compared to genomic regions in the core[16,18–20]. Features such as replication timing and chromatin influence DNA damage and repair processes, affecting mutational frequencies and signatures[4,6,41–43](Figure 3A). However, not all point mutations arise during replication, and nuclear lamins play a key role in DNA double strand break repair, such that 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 landscapes 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 adjusting for other features in all cancer type tested (Supplementary Fig. 5, Supplementary Tab. 1–2). After normalizing all features to zero mean and unitary variance, we also computed the variable importance metrics using random forest regression (Fig. 3B) and the effect sizes using multiple linear regression (Supplementary Fig. 5) for all features including the cLAD density in each 1MB bin. In general, the variable importance metrics of the cLAD density computed from the random forest regression are of similar magnitudes as that for the 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 as H3K9me3 and replication timing in most cases (**Online methods**). We also

ascertained that the influence of the sample size in the collected cohort on the results was not significant based on a sub-sampling analysis of the lymphoma cohort.

Key differences between the nuclear core and periphery in detected mutational signatures also persisted even when we adjusted for both chromatin (Figure 3C) and replication timing (Supplementary Note; **Methods**). In the SKCA cohort, we found a proportionally higher burden of UV-mediated DNA damage and trans-lesion 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>T]W substitutions, while iLADs had a relative enrichment of mutational signature $S_{SKCA}2$, representing C[C>T]Y; these preferences were observed even after adjusting for both chromatin and replication timing. Indeed, there is evidence that nuclear lamin B1 is critical for the nucleotide excision repair (NER) pathway for effective repair of DNA damage response to UV irradiation[44]. The preference for C[C>T]N (where N: A, T, G, or C) in iLADs over cLADs was detectable in other cancer types including LUSC (signature $S_{LUSC}1$). Moreover, in the LUSC cohort, the signature of oxidative DNA damage marked by C>A substitutions, especially W[C>A]W, was more common in the cLADs even after adjusting for chromatin and replication timing (Mann Whitney U test p-value < 1e-10). Therefore, a higher burden of mutation signatures arising due to external mutagens in the nuclear periphery was, at least partly, attributable to nuclear localization even when adjusting 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>T transitions in iLADs and an increase in T>G transversions in the WTN tri-nucleotide context in cLADs (Mann Whitney U test p-value < 1e-05) (Supplementary Note). The former signature is similar to COSMIC signature 2 and therefore might be due to deamination of cytosine 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>G]T, was also observed in cLADs in the STAD cohort (Mann Whitney U test p-value < 1e-07; Supplementary Note). Based on the interpretation of COSMIC signature 9, we suspect that the signature arises primarily due to mutations attributed to polymerase η[37], but other factors could also play a role.

Nuclear pores are large multi-protein channels that are conduits for nuclear transport of many small molecules and proteins, including DNA damage response and repair factors, and nuclear pores play key role in DNA repair[24,48]. Extending our analysis further, we investigated whether nuclear pore proximal regions (Figure 4A) display mutational patterns different from those observed for nuclear core and periphery regions. Nup98 is a component of the nuclear pore complex (Figure 4B); it is predominantly localized in the nuclear periphery, but can also be detected in the nuclear interior, and its dynamics of interaction with genomic regions depend on the developmental trajectory of the cell[49]. Using Nup98 ChIP-Seq data from multiple cell types[49], we identified genomic regions that bind to Nup98 in one or more cell types. Accordingly we identified cLAD and iLADs that are localized in

the neighborhood of Nup98-bound regions (NBR) or distal from it in a cell type-invariant manner (see Methods for details). cLADs at the nuclear periphery that are also close to NBRs in a cell-type invariant manner are likely to be nuclear pore-proximal. Unfortunately the number of mutations in these sub-regions was small; nonetheless, cLADs that were nuclear pore-proximal had a relatively lower AMR compared to those that were distal (Figure 4C) in the STAD, lymphoma, and CLL cohorts (FDR adjusted Mann Whitney U test p-value < 5e-02). The tri-nucleotide 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 just NBRs. Nonetheless, DNA lesions in the NBRs could be relocated to the nuclear pore complex more quickly for repair, which might play a role in lowering AMR in NBRs; further evidence is required to establish this conjecture conclusively.

Taken together, our mutational signatures and multivariate analyses indicate that the nuclear localization of genomic DNA could 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 the already captured features such as chromatin and replication timing. This fact probably arises because a subset of mutations do not emerge during replication, and nuclear lamina plays a role in DNA damage recognition and repair[21–24]. Our observations are consistent with the reported effects of nuclear lamina on the germline mutation rate variation[27]. Even benign somatic tissue samples, albeit having considerably fewer somatic mutations, show similar patterns as well (Supplementary Fig. 6; p-value > 5e-02). However, our results should be interpreted with caution: (1) the LAD information used in our paper does not match with the (potentially unknown) cell type of origin of the six cancer types studied in this paper. To identify the effect of cell type-specific LADs on mutation frequencies requires matched data which is not yet available; (2) the multicollinearities among features such as replication timing, chromatin, and nuclear localization pose a statistical challenge in order to dissect their individual effects. Here, we performed our analyses from multiple angles – only looking at ‘neutrally’ evolving genomic regions and investigating the data using different multivariate models (Supplementary Fig. 7–8, Supplementary Tab. 3). Even though the results from different analyses are in general consistent with each other, further experiments/analyses are still needed to confirm the effect 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 had a larger mutation burden and also displayed mutation signatures consistent with greater exposure to external mutagens. In addition, some of the DNA damage recognition and repair processes also 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], which

also leads to slow repair of DNA lesions. Furthermore, competing DNA repair mechanisms may recruit different DNA polymerases or their co-factors 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 of the nuclear core, but in condensed heterochromatin domains only accumulation of XPC was observed[54]. Another possibility could be that competing DNA repair mechanisms recruit different DNA polymerases or their co-factors 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 occur through the classical homologous recombination and non-homologous end-joining-mediated repair pathways, but the nuclear lamina-proximal regions tend to be refractory to HR and to allow repair primarily by the error-prone alternative end-joining mechanism[25], which could be a source of point mutations in the nuclear periphery. In any case, our findings advocate for analyzing somatic mutations in tumor and benign tissues the context of their 3D nuclear architecture.

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 and other data types were mapped to the human reference genome (hg19). Mutation frequencies for the samples in these cohorts were comparable to published literature[14], and there were no outlier subsets of samples with excessive mutations and skewed mutational signatures that dominated the overall patterns observed in our analyses.

Annotation of non-coding, non-repetitive, non-conserved regions

Since the mutational landscape of cancer genomes is shaped by both the incidence of mutations as well as natural selection during clonal evolution acting on the variability thus generated[55,56], and since variant calling is technically challenging in some genomic regions (e.g. centromere, telomere, and repetitive region), we focused only on the non-coding, non-repetitive, non-conserved regions (tier III annotation obtained from Mardis et al.[57]). In brief, such regions were identified after excluding repeat-masked regions, coding regions of annotated exons, canonical splice sites, and RNA genes, conserved genomic elements (cutoff: conservation score greater than or equal to 500 based on either the phastConsElements28way table or the phastConsElements17way table from UCSC genome browser), and regions with regulatory potential (Regulatory annotations included are targetScanS, ORegAnno, tfbsConsSites, vistaEnhancers, eponine, firstEF, L1 TAF1 Valid, Poly(A), switchDbTss, encodeUViennaRnaz, cpgIslandExt)[57]. Such regions are generally

expected to evolve in the absence of strong (positive or negative) selective pressures[58], and should have no major issues with next generation sequencing or mappability.

Annotation of nuclear core and periphery regions

Data on nuclear localization of human chromosomes was obtained from Bolzer et al.[38]. We obtained genome-wide data on lamina-associated domains for multiple human and mouse cell types[19,20]. In these datasets, lamina-associated domains were identified using DamID treatment by 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 (ESCs), astrocytes (ACs), neuronal precursor cells (NPCs), and embryonic fibroblasts in mouse (MEFs) were obtained from Peric-Hupkes et al.[20], (i) of lamin B1 in human Tig3 fibroblasts from Guelen et al[19], and (iii) of lamin B1 in human ESCs and HT1080 cells and in mouse *POU2F1*$^{-/-}$ and matching wild-type MEFs; and of lamin A in human HT1080 cells and in mouse NPCs and ACs from Mueleman et al.[59] Genomic regions associated with lamins are predominantly at the nuclear periphery, although some nucleoplasmic lamina-associated domains accumulate around nucleoli in the interior[19,20,39], while 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 overlap lamin-associated regions in (i) all the human cell line tested, and (ii) none of the human cell line tested, and denoted them as being constitutively at the nuclear periphery (dubbed constitutive lamina-associated domains, cLAD) and core (dubbed constitutive inter-lamina-associated domains, iLAD), respectively, in a cell-type invariant manner (Figure 2B). Genomic regions in the nuclear core and periphery have difference in gene density, repetitive elements, and evolutionarily conserved elements, and those features can influence selection on the somatic mutations (e.g. gene region), mutation calling (e.g. repetitive regions). Therefore, to minimize biases in our analysis, for all analyses presented in Figure 1, 2, and 4, we only considered tier-III segments[57] (i.e. noncoding, non-repetitive, non-conserved 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 lamina-associated domains data in the similar manner, we also identified tier-III segments of cLAD and iLADs that have evolutionarily conserved patterns of localization in the nuclear periphery (denoted as conserved and constitutive cLAD regions, cLAD^{c}) and nuclear core (denoted as conserved and constitutive iLAD regions, iLAD^{c}; Figure 2B), respectively, and compared AMR between them (Supplementary Fig. 3).

Annotation of nuclear pore proximal regions

Nucleoporins are key components of nuclear pore complexes that control nucleo-cytoplasmic trafficking. Liang et al. examined genomic regions bound to NUP98, a nucleoporin family nuclear pore protein, by chromatin immuno-precipitation (ChIP) using

multiple Nup98 antibodies in four cell types, three of which are related by direct lineage[49]. In tissue stem and progenitor cell populations, NUP98 bound regions (NBR) are predominantly at the nuclear periphery, but some NUP98 bound regions 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 those were within 50kb of NUP58 ChIP peaks in all cell types examined. We observed similar results using 20kb and 100kb 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 Tab. 1) and, following the approach used in a previous study[61], we only kept one GM12878 cell line dataset to decrease the bias towards blood. Similarly, H3K9me3 histone modification marks across different tissue types were obtained from the Epigenomic Roadmap project[62], including tissues such as liver, lung, and etc (Supplementary Tab. 3). The transcripts, GC% information and phastCons conservation scores for the human genome (hg19) calculated from multiple alignments with other 99 vertebrates were extracted from UCSC genome browser database[63]. For each 1MB bin, the GC%, number of genes overlapping with the bin, the proportion of nucleotides located in gene region, the average phastCons conservation scores were computed. For replication timing and H3K9me3 signals, we first calculated the average signal for each 1MB within each cell type, and then averaged across different cell types. Since in general the cell of origin of different cancer types are 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 effect of lamina-associated domains on the average mutation frequency over different tumors within a certain cancer type adjusting for conservation score, GC%, gene density, average replication timing signals (higher indicating more enriched with early replication timing on average), and the heterochromatin mark H3K9me3 average signal across multiple cell lines (Supplementary Fig. 7–8, Supplementary Tab. 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 Tab. 3. The use of linear regression was justified using the residual plots and central limit theorem when averaging the mutation frequencies of each 1Mb bin over different tumors (Supplementary Fig. 7). To account for potential correlation among 1MB 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 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 based on

permutation methods using the *varimp()* function in the same package (Fig. 3B). The same set of features was included when performing random forest regression, again with average mutation frequencies in 1MB windows across samples as the dependent variable. The goodness of fit of random forest regression was again justified using the residual plots (Supplementary Fig. 7). Since the genomic features analyzed are in general 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. As an alternative approach, we randomly divided the genome into 10 groups 50 times, computed the metrics within each group, calculated the median metrics across groups, and eventually plotted the distribution of these median scores across 50 randomizations. However, as outlined in Strobl et al. 2008[65], such an attempt cannot guarantee the complete removal of the multicollinearity bias. Therefore, even though Supplementary Fig. 8 shows that LAD has a similar, and sometimes even stronger, conditional variable importance metric compared to H3K9me3 and replication timing, this does not necessarily mean that we can interpret such results as “LAD is a more important factor than H3K9me3 in DLBCL”.

Finally, since different cancer cohorts have different sample sizes, it is worth exploring how the sample size influences our key results. To test the robustness of our findings over different sample sizes, we computed the variable importance metrics for the genomic features based on sample size equal to 10, 20, 30, and 40, respectively, in the lymphoma cohort, and found that the patterns are very 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 (NMF) and principal component analysis (PCA) to define mutation signatures, and then evaluated their contribution to each sample’s mutational spectrum using somaticSignature R package[40]. To examine significance of nuclear localization for mutagenic and repair processes, we partitioned the genome according to their chromatin or replication timing context, and then analyzed difference in mutation signatures between cLAD and iLAD regions within respective context. P-values for respective cohorts were calculated by way of Mann Whitney U tests. COSMIC mutational signatures were obtained from the COSMIC: Catalogue of Somatic Mutations in Cancer (http://cancer.sanger.ac.uk/cosmic), and were based on published report[37].

DATA AVAILABILITY

Publicly available datasets were used for this analysis, as mentioned in above sections. Nonetheless, all data will be made available upon request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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

References

1. De S, Michor F. DNA secondary structures and epigenetic determinants of cancer genome evolution. Nat. Struct. Mol. Biol. 2011; 18:950–955. [PubMed: 21725294]
2. De S, Michor F. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. Nat. Biotechnol. 2011; 29:1103–1108. [PubMed: 22101487]
3. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet. 2014; 15:585–598. [PubMed: 24981601]
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. 2013; 4:1502. [PubMed: 23422670]
5. Roberts SA, Gordenin DA. Hypermutation in human cancer genomes: footprints and mechanisms. Nat Rev Cancer. 2014; 14:786–800. [PubMed: 25568919]
6. Schuster-Bockler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature. 2012; 488:504–507. [PubMed: 22820252]
7. Polak P, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature. 2015; 518:360–364. [PubMed: 25693567]
8. Perera D, et al. Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. Nature. 2016; 532:259–263. [PubMed: 27075100]
9. Sabarinathan R, Mularoni L, Deu-Pons J, Gonzalez-Perez A, Lopez-Bigas N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. Nature. 2016; 532:264–267. [PubMed: 27075101]
10. Smith KS, et al. Signatures of accelerated somatic evolution in gene promoters in multiple cancer types. Nucleic Acids Res. 2015; 43:5307–17. [PubMed: 25934800]
11. Pedersen BS, De S. Loss of heterozygosity preferentially occurs in early replicating regions in cancer genomes. Nucleic Acids Res. 2013; 41:7615–24. [PubMed: 23793816]
12. Watson IR, Takahashi K, Futreal PA, Chin L. Emerging patterns of somatic mutations in cancer. Nat Rev Genet. 2013; 14:703–718. [PubMed: 24022702]
13. Alexandrov LB, Stratton MR. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Curr Opin Genet Dev. 2014; 24:52–60. [PubMed: 24657537]
14. Lawrence MS, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013; 499:214–218. [PubMed: 23770567]
15. De S, Ganesan S. Looking beyond drivers and passengers in cancer genome sequencing data. Ann. Oncol. Off. J. Eur. Soc. Med. Oncol. mdw677. 2016; doi: 10.1093/annonc/mdw677
16. Bickmore WA. The spatial organization of the human genome. Annu Rev Genomics Hum Genet. 2013; 14:67–84. [PubMed: 23875797]
17. Gibcus JH, Dekker J. The hierarchy of the 3D genome. Mol Cell. 2013; 49:773–782. [PubMed: 23473598]
18. Cavalli G, Misteli T. Functional implications of genome topology. Nat Struct Mol Biol. 2013; 20:290–299. [PubMed: 23463314]
19. Guelen L, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature. 2008; 453:948–951. [PubMed: 18463634]
20. Peric-Hupkes D, et al. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Mol Cell. 2010; 38:603–613. [PubMed: 20513434]
21. Meister P, Taddei A, Gasser SM. In and out of the replication factory. Cell. 2006; 125:1233–1235. [PubMed: 16814710]

22. Ball AR Jr, Yokomori K. Damage site chromatin: open or closed? Curr Opin Cell Biol. 2011; 23:277–283. [PubMed: 21489773]
23. Bell O, Tiwari VK, Thoma NH, Schubeler D. Determinants and dynamics of genome accessibility. Nat Rev Genet. 2011; 12:554–564. [PubMed: 21747402]
24. Lemaître C, Bickmore WA. Chromatin at the nuclear periphery and the regulation of genome functions. Histochem. Cell Biol. 2015; 144:111–122. [PubMed: 26170147]
25. Lemaître C, et al. Nuclear position dictates DNA repair pathway choice. Genes Dev. 2014; 28:2450–63. [PubMed: 25366693]
26. Shimi T, Goldman RD. Nuclear lamins and oxidative stress in cell proliferation and longevity. Adv. Exp. Med. Biol. 2014; 773:415–30. [PubMed: 24563359]
27. Ananda G, Chiaromonte F, Makova KD. A genome-wide view of mutation rate co-variation using multivariate analyses. Genome Biol. 2011; 12:R27. [PubMed: 21426544]
28. Weischenfeldt J, et al. Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. Nat. Genet. 2016; 49:65–74. [PubMed: 27869826]
29. Kaiser VB, Taylor MS, Semple CA. Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types. PLoS Genet. 2016; 12:e1006207. [PubMed: 27490693]
30. Berger MF, et al. Melanoma genome sequencing reveals frequent PREX2 mutations. Nature. 2012; 485:502–506. [PubMed: 22622578]
31. Cancer Genome Atlas Research, N. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012; 489:519–525. [PubMed: 22960745]
32. Wang K, et al. Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. Nat Genet. 2014; 46:573–582. [PubMed: 24816253]
33. Morin RD, et al. Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing. Blood. 2013; 122:1256–1265. [PubMed: 23699601]
34. Puente XS, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature. 2015; 526:519–24. [PubMed: 26200345]
35. Abeshouse A, et al. The Molecular Taxonomy of Primary Prostate Cancer. Cell. 2015; 163:1011–1025. [PubMed: 26544944]
36. Berger MF, et al. The genomic complexity of primary human prostate cancer. Nature. 2011; 470:214–220. [PubMed: 21307934]
37. Alexandrov LB, et al. Signatures of mutational processes in human cancer. Nature. 2013; 500:415–421. [PubMed: 23945592]
38. Bolzer A, et al. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. PLoS Biol. 2005; 3:e157. [PubMed: 15839726]
39. Nemeth A, et al. Initial genomics of the human nucleolus. PLoS Genet. 2010; 6:e1000889. [PubMed: 20361057]
40. Gehring JS, Fischer B, Lawrence M, Huber W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. Bioinformatics. 2015; 31:3673–3675. [PubMed: 26163694]
41. Kazanov MD, et al. APOBEC-Induced Cancer Mutations Are Uniquely Enriched in Early-Replicating, Gene-Dense, and Active Chromatin Regions. Cell Rep. 2015; 13:1103–1109. [PubMed: 26527001]
42. Morganella S, et al. The topography of mutational processes in breast cancer genomes. Nat Commun. 2016; 7:11383. [PubMed: 27136393]
43. Woo YH, Li WH. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. Nat Commun. 2012; 3:1004. [PubMed: 22893128]
44. Butin-Israeli V, et al. Regulation of Nucleotide Excision Repair by Nuclear Lamin B1. PLoS One. 2013; 8:e69169. [PubMed: 23894423]
45. Di Noia JM, Neuberger MS. Molecular mechanisms of antibody somatic hypermutation. Annu Rev Biochem. 2007; 76:1–22. [PubMed: 17328676]
46. Puente XS, et al. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature. 2011; 475:101–105. [PubMed: 21642962]
47. Hu Y, et al. Activation-induced cytidine deaminase (AID) is localized to subnuclear domains enriched in splicing factors. Exp. Cell Res. 2014; 322:178–192. [PubMed: 24434356]

48. Misteli T, Soutoglou E. The emerging role of nuclear architecture in DNA repair and genome maintenance. Nat Rev Mol Cell Biol. 2009; 10:243–254. [PubMed: 19277046]
49. Liang Y, Franks TM, Marchetto MC, Gage FH, Hetzer MW. Dynamic association of NUP98 with the human genome. PLoS Genet. 2013; 9:e1003308. [PubMed: 23468646]
50. Nagai S, et al. Functional Targeting of DNA Damage to a Nuclear Pore-Associated SUMO-Dependent Ubiquitin Ligase. Science (80-.). 2008; 322
51. Hsu TC. A possible function of constitutive heterochromatin: the bodyguard hypothesis. Genetics. 1975; 79(Suppl):137–150. [PubMed: 1150080]
52. Malhas AN, Lee CF, Vaux DJ. Lamin B1 controls oxidative stress responses via Oct-1. J Cell Biol. 2009; 184:45–55. [PubMed: 19139261]
53. Lange SS, Takata K, Wood RD. DNA polymerases and cancer. Nat Rev Cancer. 2011; 11:96–110. [PubMed: 21258395]
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. 2009; 122:83–91. [PubMed: 19066286]
55. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011; 144:646–674. [PubMed: 21376230]
56. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009; 458:719–24. [PubMed: 19360079]
57. Mardis ER, et al. Recurring mutations found by sequencing an acute myeloid leukemia genome. N Engl J Med. 2009; 361:1058–1066. [PubMed: 19657110]
58. Ohta T. The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. 1992; 23:263–286.
59. Meuleman W, et al. Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. Genome Res. 2013; 23:270–80. [PubMed: 23124521]
60. Hansen RS, et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc Natl Acad Sci U S A. 2010; 107:139–144. [PubMed: 19966280]
61. Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature. 2015; 521:81–84. [PubMed: 25707793]
62. Bernstein BE, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol. 2010; 28:1045–1048. [PubMed: 20944595]
63. Speir ML, et al. The UCSC Genome Browser database: 2016 update. Nucleic Acids Res. 2016; 44:D717–25. [PubMed: 26590259]
64. Freedman DA. On the so-called ‘Huber sandwich estimator’ and ‘robust standarderrors’. Am. Stat. 2006; 60:299–302.
65. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional Variable Importance for Random Forests. BMC Bioinformatics. 2008; 9:307. [PubMed: 18620558]

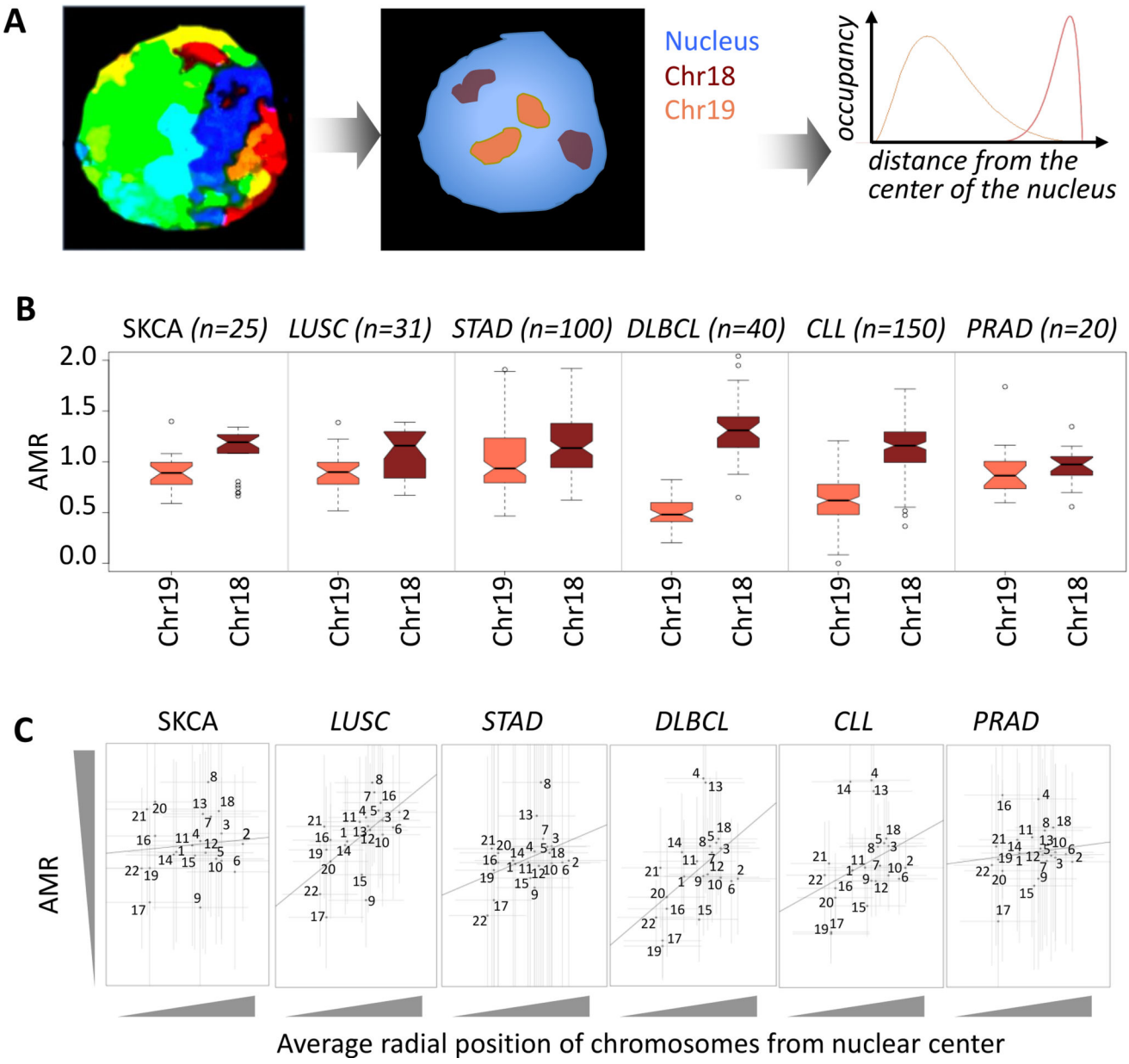

Figure 1. Somatic mutation frequencies differ between chromosomes that are at the nuclear core versus periphery

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 that for chr19. Cancer types compared are as follows: melanoma (SKCA, n = 25), lung squamous cell carcinoma (LUSC, n = 31), gastric cancer (STAD, n = 100), diffuse large B cell lymphoma (DLBCL, n = 40), chronic lymphocytic leukemia (CLL, n = 150), and prostate cancer (PRAD, n = 20). Mann Whitney U test p-value < 1e-02 for all cohorts. In the boxplots, upper whisker is defined to be 1.5×IQR more than the third quartile or the maximal value of the adjusted mutation rate (depending on which value is greater) and the lower whisker is defined to be 1.5×IQR lower than the first quartile or the minimum

value of the adjusted mutation rate (depending on which value is smaller) respectively, where IQR is the difference between the third quartile and the first quartile, i.e. the box length. C) AMR for chr1 to chr22 is plotted against their average normalized radial distances from the center of the nucleus. Average and standard deviation of normalized radial distances of chromosomes from the center of the nucleus were estimated from 54 measurements, as described in[38]. The number of samples used for AMR estimation is identical to panel b. Standard deviations of AMR and radial positions are shown with vertical and horizontal error bars, respectively. Coefficient of determination was < 0.1 in all cohorts.

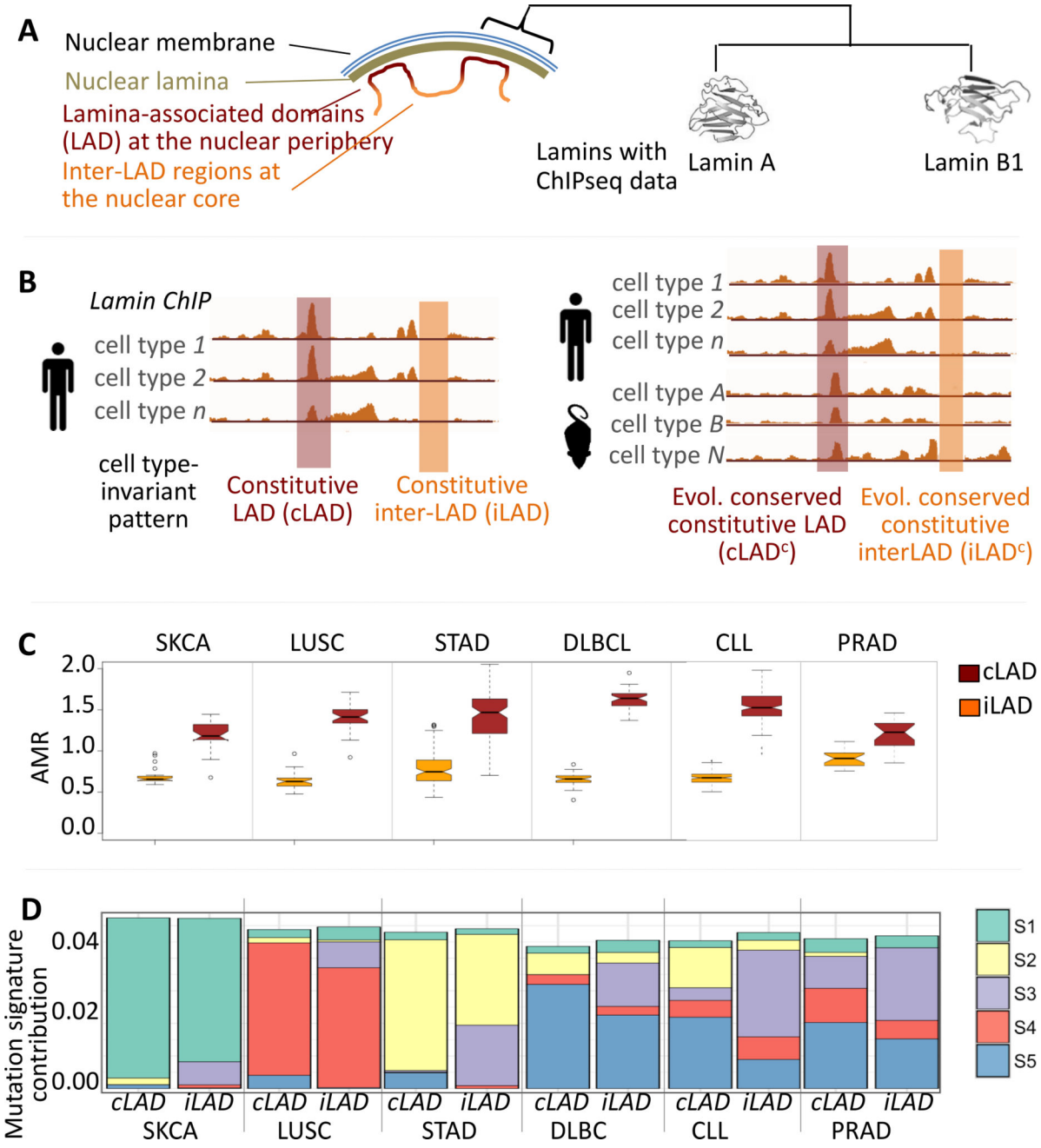

Figure 2. Somatic mutation patterns differ between genomic regions that are at the nuclear core versus periphery

A) Genomic regions interacting with lamina proteins such as Lamin 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 chromatin immuno-precipitation was used to identify genomic regions interacting with lamins in individual cell types. We classified a region as constitutively in the nuclear periphery if the region was associated with lamins in all cell types examined; conversely, a region was categorized as constitutively in the nuclear core if it did not overlap with lamin-associated domains in any of the cell types

analyzed. A subset of these regions also shows preferential positioning at the nuclear core (iLAD^c) and periphery (cLAD^c) in an evolutionarily conserved manner (Supplementary Figure 3). C) cLADs tend to have a significantly higher AMR compared to iLADs. The number of samples in each cohort is as described in Figure 1B. Mann Whitney U test p-value < 1e-05 for all cohorts. Similar results are observed when mutations in iLAD^c and cLAD^c regions are considered (Supplementary Fig. 3). In the boxplots, upper whisker is defined to be 1.5×IQR more than the third quartile or the maximal value of the adjusted mutation rate (depending on which value is greater) and the lower whisker is defined to be 1.5×IQR lower than the first quartile or the minimum value of the adjusted mutation rate (depending on which value is smaller) respectively, where IQR is the difference between the third quartile and the first quartile, i.e. the box length. (D) Mutational signatures differ between the nuclear core and periphery across different cancer types. Somatic mutational signatures were identified based on the non-negative matrix factorization and principal component analysis, using the somaticSignature[40] R package. Details of the mutation signatures S1–S5 are provided in Supplementary Fig. 4.

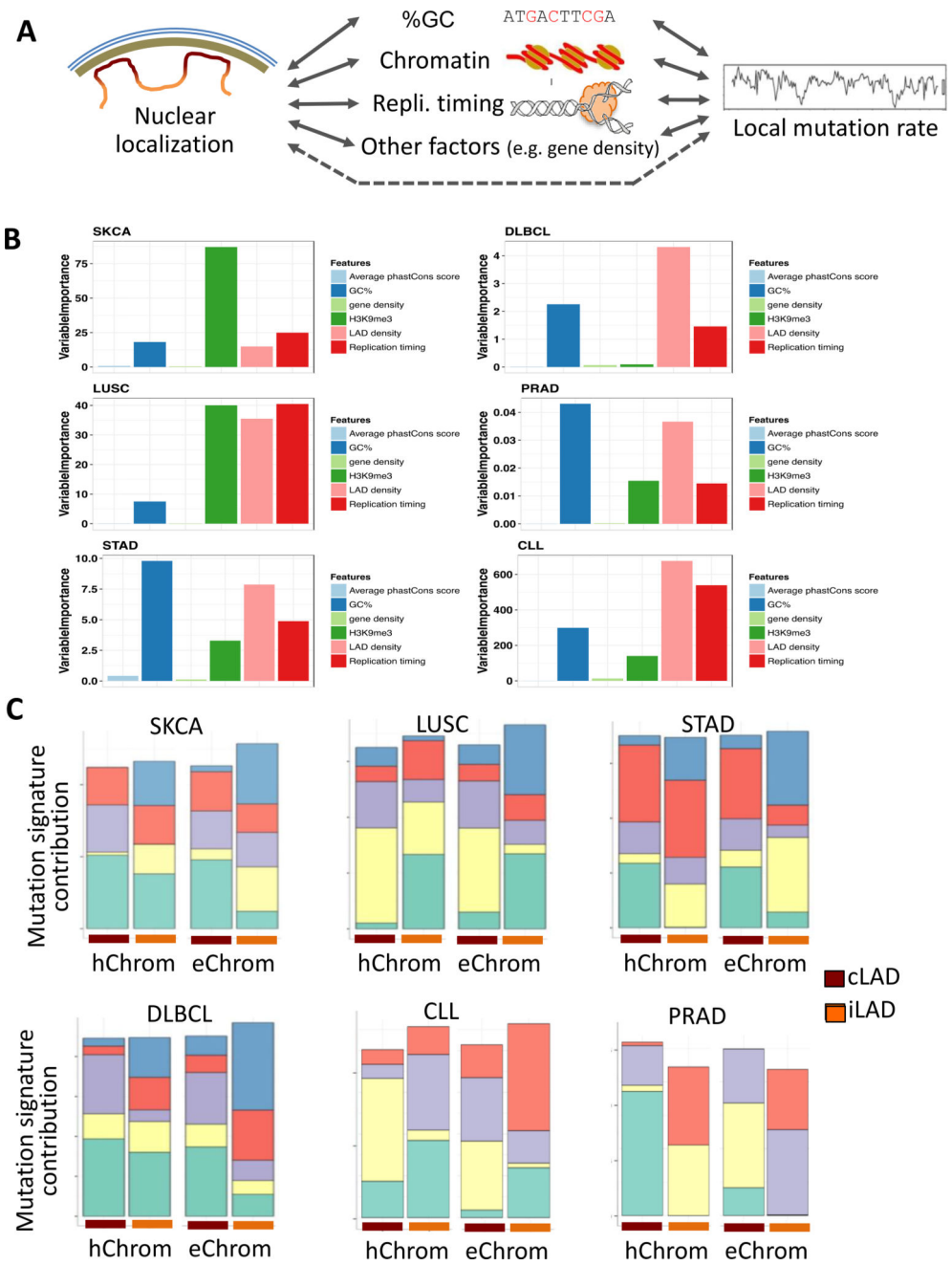

Figure 3. Certain differences in somatic mutation patterns between nuclear core versus periphery are not due to chromatin and other factors

A) Nuclear localization of genomic DNA is, at least partly, associated with chromatin, and also other features such as GC content, replication timing and gene density, which modulate the local mutation rate. Effects of nuclear localization beyond that explained via these known covariates are investigated. B) Marginal variable importance metrics for different genomic features computed from random forest regression were compared for six cancer types. C) Mutational signatures differ between genomic material localized at the nuclear core and periphery even when assessed within similar euchromatic or heterochromatic contexts. Note that color codes of mutation signatures and their relative contributions are

comparable only within respective cancer cohorts. See Supplementary Fig. 5 for additional comparative assessments.

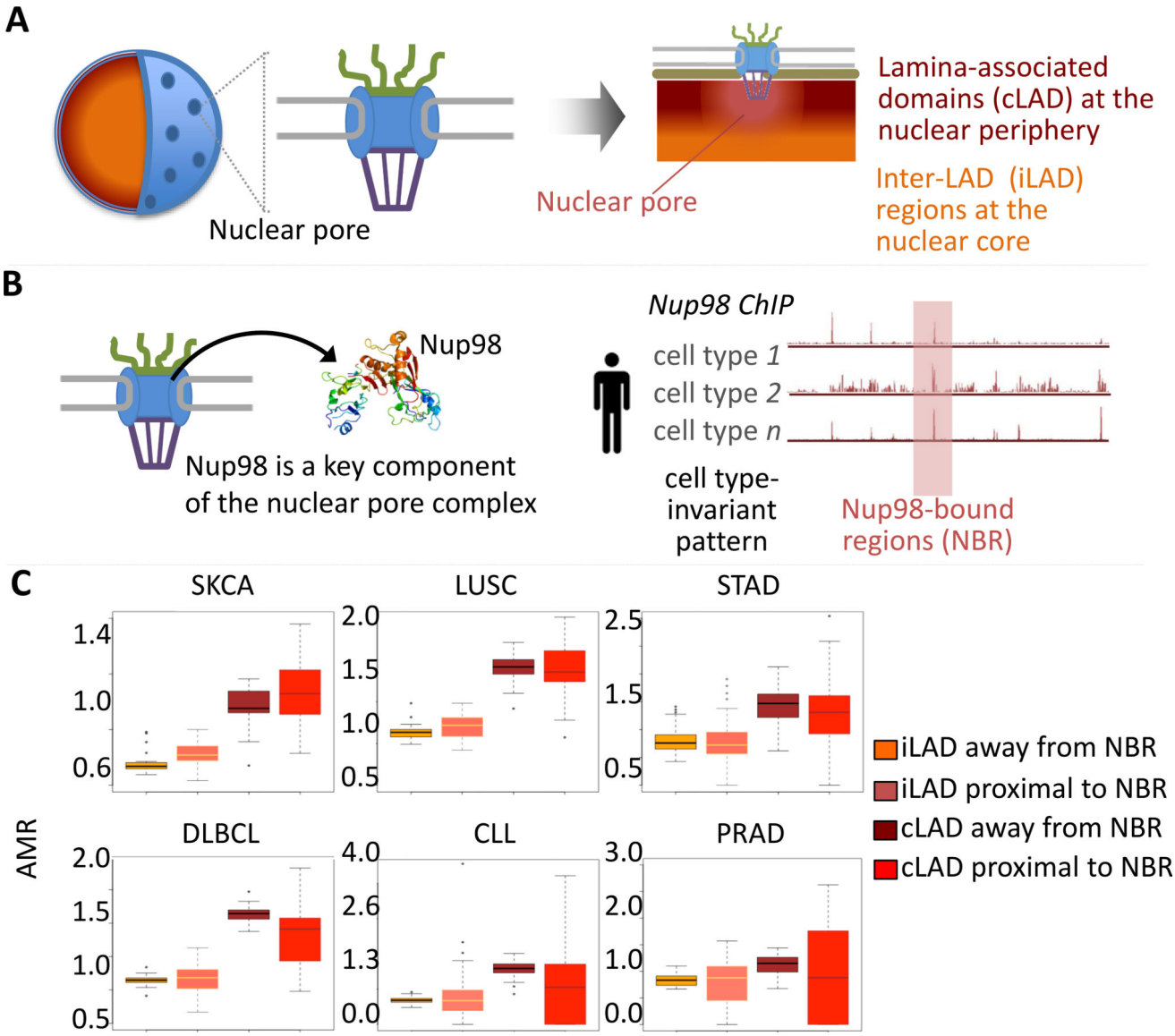

Figure 4. Nuclear pore-proximal genomic regions have characteristic somatic mutation patterns

A) Schematic representation of nuclear pores that are large, multi-protein complexes on the nuclear envelope, regulating nuclear transport of biomolecules including some mutagens and DNA repair factors. B) Nup98 is a key component of the nuclear pore complex, and based on Nup98 ChIP data, Nup98-bound regions were identified, using an approach similar to that in Fig 1b. We classified genomic regions as nuclear pore-proximal if they were within 50kb of Nup58 ChIP peaks in all cell types examined. Conversely, genomic regions that were at least 50kb 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 away from nuclear pore regions were compared. FDR-adjusted Mann Whitney U test p-value < 5e-02 in the STAD, lymphoma, and CLL cohorts. In the boxplots, the upper whisker is defined as 1.5×IQR more than the third quartile or the maximal value of the adjusted mutation rate (depending on which value is greater) and the lower whisker is defined as 1.5×IQR lower than the first quartile or the minimum value of the adjusted mutation rate (depending on which value is

smaller) respectively, where IQR is the difference between the third quartile and the first quartile, i.e. the box length.