nature COMMUNICATIONS

ARTICLE

Received 23 Jul 2012 | Accepted 16 Jan 2013 | Published 19 Feb 2013

DOI: 10.1038/ncomms2502

DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes

Lin Liu[1,2], Subhajyoti De[3] & Franziska Michor[1,2]

Single-nucleotide substitutions are a defining characteristic of cancer genomes. Many single-nucleotide substitutions in cancer genomes arise because of errors in DNA replication, which is spatio-temporally stratified. Here we propose that DNA replication patterns help shape the mutational landscapes of normal and cancer genomes. Using data on five fully sequenced cancer types and two personal genomes, we determined that the frequency of intergenic single-nucleotide substitution is significantly higher in late DNA replication timing regions, even after controlling for a number of genomic features. Furthermore, some substitution signatures are more frequent in certain DNA replication timing zones. Finally, integrating data on higher-order nuclear organization, we found that genomic regions in close spatial proximity to late-replicating domains display similar mutation spectra as the late-replicating regions themselves. These data suggest that DNA replication timing together with higher-order genomic organization contribute to the patterns of single-nucleotide substitution in normal and cancer genomes.

[1] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Instituteh, Boston, Massachusetts 02215, USA. [2] Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02215, USA. [3] Department of Medicine, University of Colorado School of Medicine, and Molecular Oncology Program, University of Colorado Cancer Center, Aurora, Colorado 80045, USA. Correspondence and requests for materials should be addressed to F.M. (email: michor@jimmy.harvard.edu).

Human cancer genomes exhibit complex mutational landscapes, often characterized by a large number of single-nucleotide substitutions (SNS) found throughout the genome[1–3]. The patterns of SNS have been shown to depend on the type of cancer, the number of cell divisions leading to the initiation and progression of the tumour and tissue-specific patterns of driver events in cancer[4–7]. Mutation rates also vary according to different genomic features such as GC content, recombination rate, CpG islands and others[8–10].

Recent advances in genomic profiling methods have enabled the characterization of the spatial arrangement of genomic material within inter-phase nuclei[11,12]. The use of such databases has enabled an unprecedented mapping of genomic regions not only relative to each other but also with regard to different higher-order structures within individual cell types[11,13,14]. Furthermore, the temporal order of DNA replication in human cells displays marked variability across genomic regions, in that some areas are replicated early, whereas others are replicated late during S phase[15–17]. To date, such data has been used to investigate evolutionary divergence between species and human nucleotide diversity, showing that late-replicating regions display larger point mutation rates than early-replicating regions[18]. It was also recently elucidated that genomic regions of similar replication timing are clustered spatially in the nucleus, that the two boundaries of somatic copy number alterations (SCNAs) in cancer genomes tend to be found in regions with the same replication timing and that regions replicated early and late display distinct patterns of frequencies of SCNA boundaries, SCNA size and a preference for deletions over insertions[19]. For example, deletions are generally more frequent than amplifications in late as compared with early replication timing zones.

Recently available genome-wide sequencing data have enabled us to investigate the patterns of SNS in different temporal phases and spatial compartments during the DNA replication process. Several studies have illustrated associations between mutation frequencies and other genetic and epigenetic factors[20–22]. Woo *et al.*[20] utilized information on selection and DNA replication timing to study the local variation of mutation frequencies, whereas Schuster-Bockler *et al.*[21] and the The Cancer Genome Atlas (TCGA) lung cancer consortium[21,22] proposed a multivariate analysis approach to investigate epigenetic markers using data from different cell types.

Here we investigated the patterns of SNS across the genome by using replication timing data conserved across several cell lines based on data from Hansen *et al.*[23] and regions not under strong selection pressure[24]. We then comprehensively catalogued individual mutation signatures in these constant late and early replication timing zones. Finally, we utilized information on higher-order chromatin interactions between genomic material to demonstrate the coordinative effects between replication timing and nuclear architecture on the mutational landscape of cancer genomes.

Results

Description of analysed data sets. We integrated SNS data from completely sequenced genomes of five cancer types (melanoma[25,26], prostate cancer[27], small cell lung cancer[28], chronic lymphocytic leukemia[29] and colorectal cancer[30]; the number of samples analysed is shown in Table 1), two completely sequenced personal genomes[31,32], genome-wide DNA replication timing data[23,24] and data on single-nucleotide differences between the human (hg18) and chimpanzee (panTro2) genomes[33]. All data were mapped to the human genome version hg18. Genome-wide replication timing data, obtained using a technique based on massively parallel sequencing (Repli-Seq) across different human cell types[23], were used to classify genomic regions as 'constant early', 'constant mid', 'constant late' and 'variable', according to the extent of consistency of replication timing regions across the different cell types.

Table 1 | Cancer types and numbers of samples used in this study.

Cancer types	Source	Number of samples
Melanoma Study 1	Pleasance *et al.*[25]	1
Melanoma Study 2	Berger *et al.*[26]	25
Prostate cancer	Berger *et al.*[27]	7
Small cell lung cancer	Pleasance *et al.*[28]	1
Chronic lymphocytic leukaemia	Puente *et al.*[29]	4
Colorectal cancer	Bass *et al.*[30]	9

The table displays the number of samples and citation for each cancer type analysed.

Because cancer development encompasses two intertwined processes—the acquisition of mutations and natural selection affecting the frequency of the resultant phenotypes[3], we first excluded regions such as the centromere and telomere, Y chromosome, genes and promoters (±2 kb), repeat elements and ultra-conserved regions[33] from the data. The remaining sequences were expected to evolve nearly neutrally and were termed *filtered intergenic regions* (FIRs). Using FIRs only, we were also able to avoid some challenging issues of variant calling outside of these regions[34]. The frequency of mutations detected in these regions was referred to as *adjusted intergenic mutation frequency* (AIMF). We mapped SNS data for each cancer genome onto the FIRs and calculated the AIMF for both the whole genome and each chromosome individually. Our analysis revealed that the AIMF varies substantially across the four cancer types (Table 2 and Supplementary Table S1, based on analysis of variance adjusted by multiple comparisons); such variation could be explained by biological differences in the cancer types and/or differences in the experimental design, sequencing technologies and variant calls. Nevertheless, similar trends were observed in the two completely sequenced personal genomes[31,32], pointing towards meaningful differences (Table 2). We then also repeated these analyses using genome-wide data instead of FIRs and obtained consistent results (Table 3).

Mutation frequencies depend on replication timing. We first sought to investigate the effects of DNA replication timing onto the patterns of SNS frequency in cancer genomes. We utilized only constant late and constant early replication timing zones[23] to exclude tissue specificity as a confounding factor. The constant mid category represented a much smaller part of the human genome and was thus discarded. We first analysed the melanoma genomes[25]. We observed that the mutation frequency in the FIRs was intimately linked to DNA replication timing: FIRs with constant late replication timing displayed a significantly higher AIMF compared with those with constant early replication timing (Mann–Whitney *U*-test *P*-value $= 2.075 \times 10^{-7}$). This effect was consistent across all 23 chromosomes (chr1-22 and chrX). We did not identify a significant trend when investigating the 23 chromosomes individually (Fig. 1 and Supplementary Figs S1–S13). We then repeated our analysis for the other four cancer types (prostate cancer, small cell lung cancer, chronic lymphocytic leukaemia and colorectal cancer) and two personal genomes (Watson[31] and HuRef[32] genomes, analysed separately)

Table 2 | The genome-wide AIMF.

AIMF	Melanoma Study 1	Melanoma Study 2	Prostate cancer	SCLC	CLL	Colorectal cancer	Watson	HuRef
Constant early	0.100	0.203	0.00874	0.0439	0.00230	0.0309	7.99	10.7
Constant late	0.184	0.356	0.0149	0.135	0.00653	0.0801	9.07	11.5
Core	0.096	0.199	0.00886	0.0405	0.00223	0.0297	7.92	10.6
	0.181	0.352	0.0135	0.129	0.00640	0.0801	8.78	11.3
Periphery	0.124	0.226	0.00806	0.0630	0.00270	0.0372	8.41	11.0
	0.185	0.357	0.0153	0.138	0.00658	0.0802	9.18	11.5
Low CpGI	0.105	0.212	0.00763	0.0605	0.00352	0.0353	8.36	10.4
	0.183	0.362	0.0150	0.140	0.00833	0.0794	9.02	11.4
High CpGI	0.0998	0.202	0.00885	0.0423	0.00305	0.0307	7.96	10.7
	0.189	0.319	0.0143	0.108	0.00605	0.0586	9.36	11.8
Low RR	0.105	0.200	0.00926	0.0447	0.00313	0.0314	7.15	9.35
	0.192	0.377	0.0150	0.141	0.00836	0.0793	8.63	10.8
High RR	0.0972	0.205	0.00839	0.0432	0.00307	0.0310	8.53	11.5
	0.171	0.323	0.0146	0.119	0.00747	0.0717	9.76	12.5
Low GC percent	0.0973	0.206	0.00989	0.0464	0.00348	0.0365	7.25	9.21
	0.179	0.362	0.0149	0.141	0.00830	0.0836	8.91	11.1
High GC percent	0.101	0.202	0.00842	0.0432	0.00298	0.0295	8.20	11.1
	0.198	0.338	0.0149	0.119	0.00721	0.0723	9.51	12.4
Gene poor	0.0933	0.197	0.00833	0.0554	0.00357	0.0368	8.38	10.8
	0.187	0.341	0.0149	0.139	0.00821	0.0780	9.15	11.6
Gene rich	0.102	0.192	0.00883	0.0415	0.00299	0.0300	7.91	10.6
	0.166	0.284	0.0148	0.111	0.00659	0.0628	8.48	10.5
GPos	0.0978	0.206	0.00809	0.0459	0.00320	0.0320	7.95	10.4
	0.185	0.359	0.0151	0.141	0.00819	0.0790	9.05	11.4
GNeg	0.102	0.202	0.00902	0.0429	0.00303	0.0301	8.02	10.8
	0.179	0.343	0.0141	0.119	0.00745	0.0720	9.17	11.8

Abbreviations: CLL, chronic lymphocytic leukaemia; CpGI, CpG island; GPos, Giemsa positive; Gneg, Giemsa negative; RR, recombination rate; SCLC, small cell lung cancer.
The AIMF ($\times 10^{-4}$) is shown for five cancer types (1 sample of melanoma of study 1[25], 25 samples of melanoma of study 2[26], 7 samples of prostate cancer[27], 1 sample of small cell lung cancer[28], 4 samples of chronic lymphocytic leukaemia[29], 9 samples of colorectal cancer[30]) and two completely sequenced personal genomes (Watson[31]) and HuRef genomes[32]) for 'constant early' (upper row) and 'constant late' replicating regions (lower row); and after stratifying by six different genomic features (nuclear lamina-associated domains[11], CpG islands, recombination rate, GC percentage, gene density and chromatin status). CpG island data generated by Wu *et al.*[53] were obtained from the UCSC genome browser. GC percentage was calculated for each 1 Mb window using gc5base[54]. Chromatin status was derived from Giemsa-staining-based g-banding patterns[50]. We also used RefSeq genes in the control analyses. The recombination rate based on either the decode[55], Marshfield[56] or Genethon genetic maps[57] was downloaded from UCSC genome browser.

and obtained similar results (Fig. 1). Using a permutation test based on randomly permuting the number of mutations in the adjusted intergenic regions (Supplementary Figs S14 and S15), we recalculated the permuted AIMFs and compared them with the observed patterns, obtaining a permutation *P*-value <0.001 for all cancer types. To investigate the confounding effects of different genomic features, we then adjusted for a variety of potential confounders such as gene density, GC percentage, recombination rate, CpG islands, chromatin states[9] and nuclear lamina-associated regions[11] (Supplementary Figs S1–S13 and Table 2). The observed patterns of SNS with regard to replication timing were consistent in different groups categorized by these genomic features for all analysed genomes. This observation suggests that our findings are unlikely biased by these genomic features and the internal biological variation among cancer types.

We then repeated our analyses using genome-wide mutation frequencies in constant late and constant early replication timing regions (Table 3). In general, we obtained robust results. Surprisingly, in prostate cancer and small cell lung cancer, the genome-wide mutation frequencies were higher than the AIMF (Tables 2 and 3); these findings might arise because of an excess of mutations in repeat elements in these two cancer types, which could be due to mapping issues, different criteria used for variant calls or diverse biological mechanisms of tumorigenesis. After adjusting for several genomic features, we again obtained results consistent with previous studies showing that genomic regions, which (i) have a high gene density, (ii) reside in euchromatin regions or (iii) have a high CpG content, display lower mutation rates. When analysing adjusted intergenic regions instead of the whole genome, however, some of these associations were not observed: for instance, we observed a relatively higher AIMF in melanoma samples as well as the Watson and HuRef genomes in regions with higher CpG density compared with lower CpG density. One possible reason for this observation is that SNS in FIRs might not be strongly affected by the active elements around the regions. Alternatively, this trend might also be because of sequencing or mapping issues in repeat elements. We also calculated the SNS frequencies in genes only: the SNS frequency in genes was much lower than the AIMF (χ^2-test, *P*-value <0.0001), and constant late replication timing regions had larger SNS frequencies in genes (Supplementary Table S2). To account for the potential inconsistencies of replication timing across cell lines, we used six alternative replication timing data sets[24,35,36] from the *Replication Domain* database to confirm our findings (Supplementary Fig. S16).

Recent evidence suggests that DNA replication timing may be coordinated across megabase-scale domains in metazoan genomes, and that early and late replication initiation occurs in spatio-temporally separate nuclear compartments[13,14,19]. Thus, it is possible that DNA replication timing domains within a larger genomic region (for example, 1 Mb) might affect the SNS frequency. For instance, overall, constant late genomic material could reside in regions that are either predominantly replicated late or not, and vice versa. To address this issue, we segmented the human genome into 1 Mb non-overlapping windows and dichotomized these windows into those with a large versus small proportion of late-replicating domains based on the prevalence of late-replicating base pairs within them. Using different cutoffs to categorize these 1 Mb windows, we found that in the stratum with a large proportion of late RT material, the SNS frequencies are higher than in the stratum with a small proportion of late RT

Table 3 | The genome-wide mutation frequency (MF).

Genome-wide MF	Melanoma Study 1	Melanoma Study 2	Prostate cancer	SCLC	CLL	Colorectal cancer	Watson	HuRef
Constant early	0.068	0.197	0.0119	0.047	0.00187	0.0300	6.63	9.84
Constant late	0.176	0.352	0.0172	0.139	0.00571	0.0746	8.27	12.1
Core	0.064	0.194	0.0119	0.045	0.00185	0.0295	6.56	9.72
	0.173	0.343	0.0167	0.124	0.00616	0.0752	8.05	11.8
Periphery	0.104	0.226	0.0121	0.065	0.00208	0.0351	7.25	10.9
	0.177	0.356	0.0173	0.143	0.00555	0.0744	8.34	12.3
Low CpGI	0.079	0.216	0.0112	0.054	0.00228	0.0352	7.03	10.1
	0.178	0.362	0.0172	0.145	0.00596	0.0772	8.23	12.0
High CpGI	0.068	0.204	0.0120	0.047	0.00183	0.0296	6.59	9.80
	0.164	0.318	0.0169	0.105	0.00429	0.0603	8.45	12.1
Low RR	0.067	0.216	0.0122	0.047	0.00171	0.0303	5.91	8.65
	0.184	0.358	0.0173	0.149	0.00630	0.0775	7.86	11.5
High RR	0.069	0.195	0.0117	0.048	0.00199	0.0298	7.17	10.7
	0.161	0.323	0.0165	0.123	0.00489	0.0685	8.92	13.0
Low GC percent	0.063	0.208	0.0112	0.051	0.00167	0.0313	5.77	8.20
	0.178	0.360	0.0175	0.147	0.00601	0.0778	8.16	11.9
High GC percent	0.069	0.196	0.0120	0.047	0.00189	0.0299	6.71	10.0
	0.165	0.316	0.0154	0.114	0.00428	0.0592	8.78	12.8
Gene poor	0.073	0.204	0.0115	0.051	0.00220	0.0328	7.27	9.74
	0.179	0.357	0.0172	0.142	0.00593	0.0766	8.37	11.1
Gene rich	0.068	0.196	0.0120	0.047	0.00182	0.0296	6.52	10.4
	0.154	0.322	0.0168	0.111	0.00427	0.0617	7.59	12.3
GPos	0.066	0.201	0.0117	0.049	0.00201	0.0308	6.52	9.64
	0.177	0.354	0.0171	0.143	0.00596	0.0757	8.25	12.2
GNeg	0.070	0.195	0.0120	0.046	0.00181	0.0296	6.28	9.94
	0.164	0.337	0.0163	0.126	0.00504	0.0676	8.21	12.0

Abbreviations: CLL, chronic lymphocytic leukaemia; CpGI, CpG island; GPos, Giemsa positive; Gneg, Giemsa negative; RR, recombination rate; SCLC, small cell lung cancer.
The MF ($\times 10^{-4}$) is shown for five cancer types (1 sample of melanoma of study 1[25], 25 samples of melanoma of study 2[26], 7 samples of prostate cancer[27], 1 sample of small cell lung cancer[28], 4 samples of chronic lymphocytic leukaemia[29], 9 samples of colorectal cancer[30]) and two completely sequenced personal genomes (Watson[31] and HuRef genomes[32]) for 'constant early' (upper row) and 'constant late' replicating regions (lower row); and after stratifying by six different genomic features (nuclear lamina-associated domains[11], CpG islands, recombination rate, GC percentage, gene density, and chromatin status). CpG island data generated by Wu *et al.*[53] were obtained from the UCSC genome browser. GC percentage was calculated for each 1 Mb window using gc5base[54]. Chromatin status was derived from Giemsa-staining-based g-banding patterns[50]. We also used RefSeq genes in the control analyses. The recombination rate based on either the decode[55], Marshfield[56], or Genethon genetic maps[57] was downloaded from UCSC genome browser.

material (χ^2-test, P-value <0.001 in all cases), but the differences of mutation frequencies between specific early and late replication timing regions hold in both strata (Supplementary Fig. S17). This observation was also consistent across the five cancer types. Therefore, the prevalence of late replication timing zones on a larger scale is unlikely to affect our observations. Interestingly, although it has been reported that the transition regions between late and early replication timing zones are less stable than other parts of the genome[37], we did not observe significant differences in terms of mutation rates between regions at the centre versus at the boundary of individual replication timing zones based on the constant late and early replication timing data (Supplementary Fig. S18).

Different temporal phases of DNA replication have been reported to associate with the existence of DNA secondary structures[38], common fragile sites[39] and sometimes *cis*-regulatory elements[40]. To examine whether these factors could confound the different mutation frequencies in early and late replication timing zones, stratification analyses were performed based on these factors (Supplementary Fig. S19). The preference of SNS in constant late over constant early DNA replication timing was not masked by these factors, demonstrating remarkable robustness of our observation, in addition to other control analyses. Besides, we focused on intergenic mutations, whose function is difficult to be inferred computationally or verified experimentally[41]. However, some portion of the intergenic regions can potentially be transcribed[42]; for instance, noncoding RNAs, especially large intergenic noncoding RNAs (lincRNAs), may also contribute to the local variation of the distribution of mutation frequencies[41]. A recent study has catalogued all known lincRNAs with the most thorough annotation to date[43]. Because those adjusted intergenic mutations included in our study are far away from protein-coding genes (median distance to the closest transcription start site: 400 kb), it is possible that these mutations have a role in acting on those lincRNAs. We observed that the SNS did not display any global preference towards residing within FIR regions overlapping with lincRNAs (Supplementary Table S3). However, because we cannot rule out that mutations varying lincRNAs are more frequent in cancer genomes and the effects of variation in lincRNAs may be subtle compared with variation in protein-coding genes, more work is required to delineate these effects.

Mutation signatures depend on replication timing. When investigating the different types of SNS in cancer FIRs, we observed that the patterns depended on whether FIRs were located in constant early versus constant late replication timing zones. We considered six types of SNS signatures for each nucleotide in the genome: A→C: T→G, A→G: T→C, A→T: T→A, C→A: G→T, C→T: G→A and C→G: G→C. The proportions of these six types of substitutions were calculated for the constant late and constant early replication timing FIRs (Fig. 2). The overall patterns were significantly different between constant early and constant late replication timing (χ^2-test, P-values <0.01 in all cases, Fig. 2). Similar differences of substitution patterns between early and late replication timing zones were obtained after controlling for the effects of gene density, GC percentage, chromatin state, CpG islands, recombination rate and nuclear lamina-associated regions (Supplementary Figs S20–S31). Interestingly, we also obtained a similar trend using the single-nucleotide polymorphism data from the two completely sequenced personal genomes (Fig. 2 and Supplementary Figs

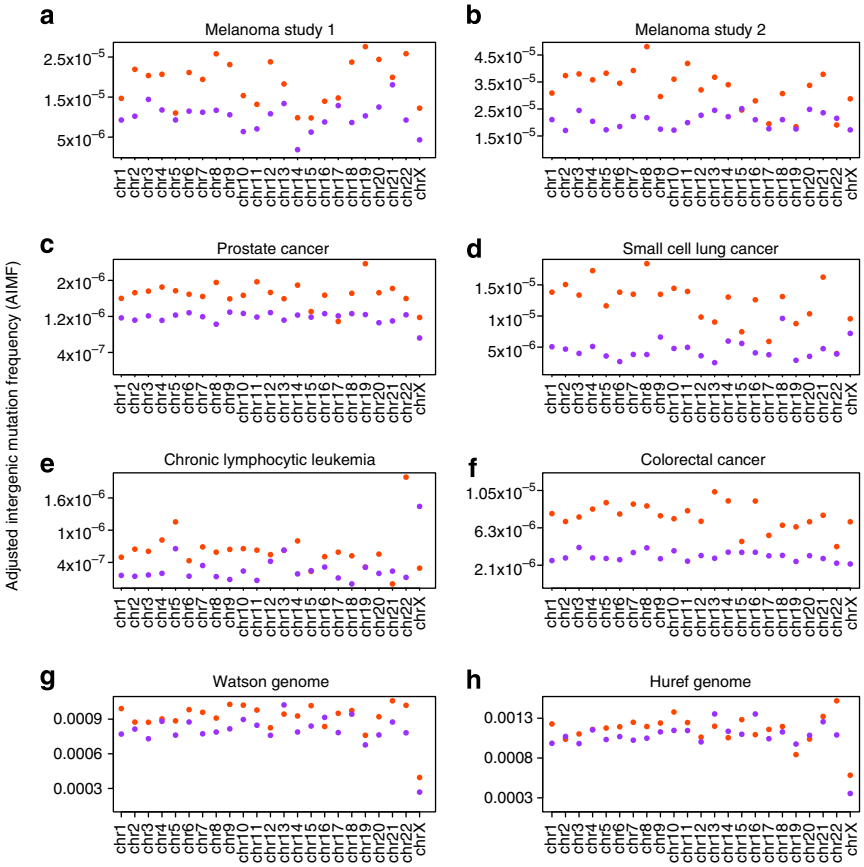

Figure 1 | Effects of DNA replication timing on mutation rates. The figure shows the AIMF for regions residing within constant early (purple) and constant late DNA replication timing zones (orange) for completely sequenced genomes of five cancer types and two personal genomes: (**a**) melanoma of study 1 (1 sample)[25], (**b**) melanoma of study 2 (25 samples in total)[26], (**c**) prostate cancer (7 samples in total)[27], (**d**) small cell lung cancer (1 sample)[28], (**e**) chronic lymphocytic leukaemia (4 samples in total)[29], (**f**) colorectal cancer (9 samples in total)[30], (**g**) Watson[31] and (**h**) HuRef genomes[32]. The AIMF represents the number of SNS observed per base pair in the FIRs, which overlap with constant early and constant late DNA replication timing zones, respectively. The horizontal axes display the results for chr1–chr22 and chrX.

S20–S31). The mutation signatures within genes and promoters were also investigated (Supplementary Fig. S32) to allow a comparison between genic and intergenic regions. We found similar patterns in genes and FIRs in terms of mutation signatures (Supplementary Fig. S32).

Comparing the data across cancer types, we observed some common patterns: some signatures were more prevalent in the constant late regions, whereas others were preferentially located in constant early regions. For instance, A→T: T→A transversions occurred most often in the constant late replication timing regions in all five cancer types. Out of the five cancer types and the two personal genomes studied, the differences of the proportion of A→T: T→A in early and late replication timing regions were significant in prostate cancer samples, melanoma samples from study 2, and Watson and HuRef genomes (adjusted *P*-values <0.01 after multiple testing correction). Overall, the higher proportion of A→T: T→A in late replication timing zones was observed in 38 out of all the 47 samples analysed in our study (Supplementary Figs S33–S40). In contrast, the frequencies of mutations and the relative proportions of the six types of substitution signatures differed among the five cancer genomes and two personal genomes; for example, the most frequent type of substitution in melanoma was the C→T transition[25]. In general, the consistency in the relative proportions of substitution signatures in constant early versus constant late replication timing regions might indicate common mutagenic mechanisms in different temporal phases of DNA replication.

Mutation frequencies and higher-order nuclear organization. The spatio-temporal segregation of DNA replication timing leads to the formation of DNA replication factories in which DNA synthesis takes place on multiple strands simultaneously[13,14]. We therefore aimed to test the hypothesis that those regions brought in close spatial proximity by the proposed fractal organization of the genome[12] display similar mutation frequencies. To address this question, we divided the whole genome into 100 Kb non-overlapping windows and obtained Hi-C-based long-range interaction data from the GM06990 and K562 cell lines from Lieberman-Aiden *et al.*[12] to measure the spatial proximity between two individual windows. We excluded any two loci that were closer than 20 Kb from each other on linear DNA. We then stratified all pairs of windows according to the number of Hi-C reads between them and investigated those windows close to but outside of the constant late DNA replication timing zones. Those regions that overlapped with FIRs were referred to as 'transition-to-late' regions; these are the regions that do not reside in constant late replication timing zones but are linked to constant late regions with at least one Hi-C read. Compared with the AIMF in constant late and constant early DNA replication timing zones, we found that the AIMF in the 'transition-to-late' regions

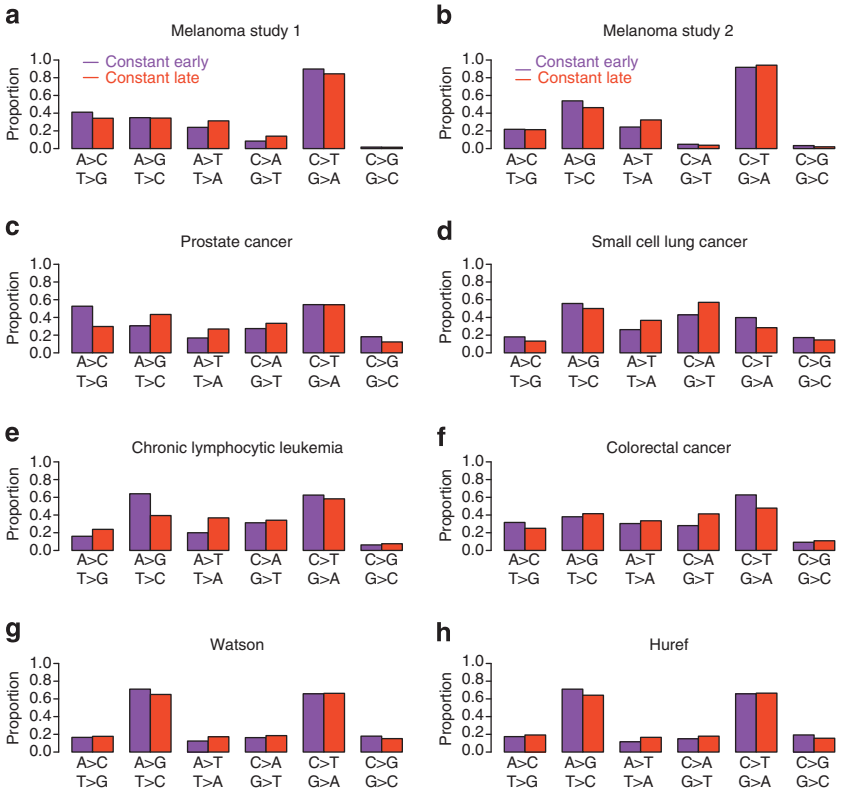

Figure 2 | Relationship between DNA replication timing and substitution patterns. The figure shows the proportions of different types of SNS in the constant early (purple) and constant late (orange) DNA replication timing zones for completely sequenced genomes of five cancer types and two personal genomes: (**a**) melanoma of study 1[25], (**b**) melanoma of study 2[26], (**c**) prostate cancer[27], (**d**) small cell lung cancer[28], (**e**) chronic lymphocytic leukaemia[29], (**f**) colorectal cancer[30], (**g**) Watson[31] and (**h**) HuRef genomes[32]. The proportions were calculated based on the hg18 reference allele so that Prob(A→C: T→G) + Prob(A→G: T→C) + Prob(A→T: T→A) = 100%, and Prob(C→A: G→T) + Prob(C→T: G→A) + Prob(C→G: G→C) = 100% for each of the constant late and constant early categories. Note that A→T: T→A is a signature commonly higher in late replication timing in all cancer types. Using the χ^2-test and correcting for multiple hypothesis testing by false discovery rate, **b**, **c**, **g** and **h** are significantly different with adjusted *P*-values <0.01.

were much closer to, yet still smaller than that in constant late DNA replication timing zones. Interestingly, the AIMF was positively associated with the interaction counts (linear regression *P*-value <0.01 for each cancer type, Fig. 3). Furthermore, in most cases, the AIMF in these regions was higher than the genome-wide AIMF (Fig. 3). These observations were consistent across the Hi-C data from the GM06990 and K562 cell lines and the Hi-C data for the GM06990 cell line generated using different restriction enzymes (*Hin*dIII and *Nco*I) (Supplementary Figs S41–S43).

We also examined whether the different proportions of DNA replication timing (including constant early, constant mid, constant late and variable) in the transition zones confounded our results. To address this issue, we performed the following analysis: the FIRs were divided into four groups—(i) constant late regions linked with Hi-C reads to constant late regions, (ii) constant late regions linked with Hi-C reads to constant early regions, (iii) constant early regions linked with Hi-C reads to constant late regions and (iv) constant early regions linked with Hi-C reads to constant early regions. We found that group (i) had the highest mutation frequency, whereas group (iv) had the lowest. Moreover, the mutation frequency of group (ii) was closer to, but still lower than that of group (i), and a similar trend was observed between groups (iii) and (iv) (Fig. 4). Interestingly, all pairwise comparisons were significantly different (Mann–Whitney *U*-test, false discovery rate-adjusted *P*-value <0.03 in all cases). Taken together, we found that those regions close to late DNA replication timing zones had similar, though lower, mutation frequencies, suggesting a potential role for higher-order chromatin organization on the mutagenic mechanisms during DNA replication.

Evolutionary and cancer mutations share genomic locations. We then sought to compare the regions prone to accumulating adjusted intergenic SNS in cancer genomes versus mutations arising on evolutionary time scales. To this end, we obtained data on differences between the human hg18 and chimpanzee panTro2 genomes from the UCSC genome browser[33], using a similar approach as in Stamatoyannopoulos *et al.*,[18] and compared the number of such changes with the number of SNS in each cancer type in 1 Mb non-overlapping windows[44]. The five cancer types had very different regions that overlapped with those regions harbouring human–chimpanzee SNS (Supplementary Fig. S44). After collapsing the windows with SNS in each of the five cancer types together, we identified 1,039 such windows with at least one SNS in any of the five cancer types in early DNA replication timing zones. We then fixed the number of windows with cancer mutations and selected the same number of windows with the highest number of human–chimpanzee SNS. Out of these 1,039 windows, 775 were also present among the human–chimpanzee SNS windows. We then performed similar analyses in late DNA replication timing zones and found that out of 1,240 windows, 1,208 overlapped in

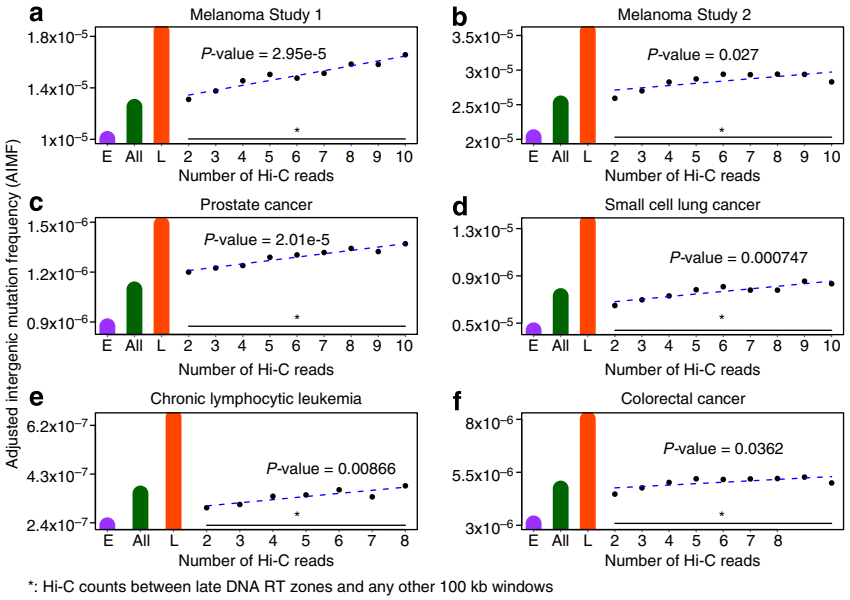

Figure 3 | Higher-order nuclear architecture is associated with mutation frequencies. The figure shows the AIMF in the 'transition-to-late' regions defined by different numbers of Hi-C interaction counts from the GM06990 cell line between regions inside and outside the constant late DNA replication timing zones for (**a**) the melanoma sample of study 1[25], (**b**) melanoma samples of study 2[26], (**c**) prostate cancer samples[27], (**d**) the small cell lung cancer sample[28], (**e**) chronic lymphocytic leukaemia samples[29] and (**f**) colorectal cancer samples[30]. Statistical significance was evaluated using simple linear regression, and *P*-values were obtained. All *P*-values were <0.01. The green bar shows the genome-wide AIMF, the orange bar shows the AIMF in constant late DNA replication timing FIR and the purple bar shows the AIMF in constant early DNA replication timing FIR. The blue dashed line, that is, the fitted linear model, shows the positive association between the AIMF and the Hi-C counts that was used to stratify the regions. Because of the small mutation number in the chronic lymphocytic leukaemia genome, we only used 2–8 Hi-C counts in **d**. The x-axes display the groups of regions stratified by the number of Hi-C interactions with constant late replication timing regions.

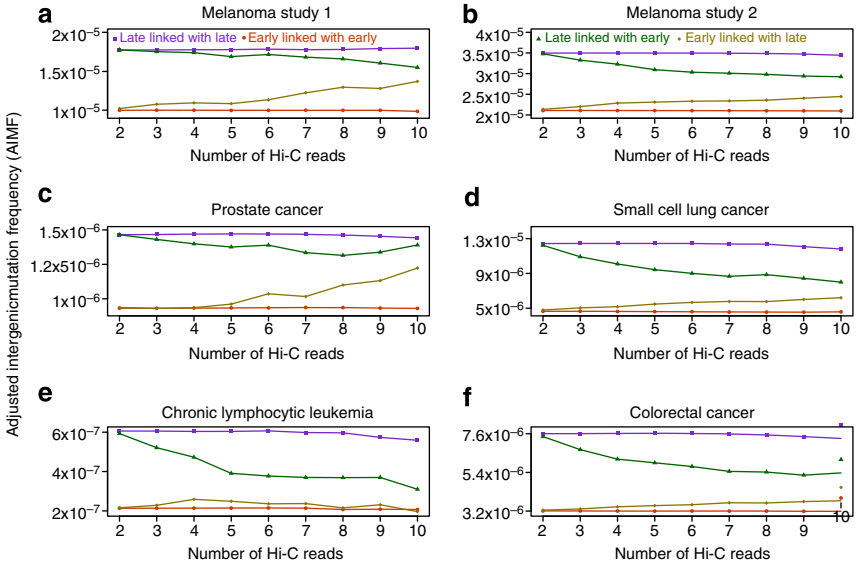

Figure 4 | Effects of transition regions on mutation frequencies. The figure shows the AIMF for (**a**) the melanoma sample of study 1[25], (**b**) melanoma samples of study 2[26], (**c**) prostate cancer samples[27], (**d**) the small cell lung cancer sample[28], (**e**) chronic lymphocytic leukaemia samples[29] and (**f**) colorectal cancer samples[30] in four groups of adjusted intergenic regions: constant late replication timing regions linked with constant late replication timing regions by Hi-C interactions (purple); constant late replication timing regions linked with constant early replication timing regions by Hi-C interactions (green); constant early replication timing regions linked with constant late replication timing regions by Hi-C interactions (gold); and constant early replication timing regions linked with constant early replication timing regions by Hi-C interactions (red). The x-axes display the groups of paired regions stratified by the number of Hi-C reads (2–10). All pairwise comparisons were significantly different from each other (Mann–Whitney *U*-test, false discovery rate-adjusted *P*-values <0.03 in all cases).

cancer and human–chimpanzee SNS (Supplementary Fig. S45). Although the overlap between regions with cancer SNS and the regions with the top human–chimpanzee SNS varied across different cancer types, after pooling them together, the overlap became larger. Therefore, we concluded that at the scale of 1 Mb, most regions harbouring human–chimpanzee SNS were also

regions harbouring SNS in any one of the five cancer types. This finding suggests some common mechanisms between human–chimpanzee evolutionary transversion and cancer mutagenesis, with no obvious differences in early versus late DNA replication timing zones.

Discussion

In this paper, we have demonstrated that mutational landscapes of cancer genomes differ between early and late DNA replication timing zones, with higher mutation frequencies in late replication timing regions. We identified different patterns of mutation signatures across these zones; for example, A→T: T→A mutation signatures commonly appeared in most cancer samples investigated. This finding implies that some mutagenic and repair mechanisms might depend on the DNA replication timing of genomic material. The differences in mutation frequencies and signatures between early and late replication timing also hold after controlling for several genomic features such as GC percentage, CpG density, recombination rate, chromatin accessibility, gene density and lamina-associated domains. Also, the transition-to-late regions defined based on Hi-C interactions, although not located in constant late replication timing regions, has higher mutation frequencies than the overall AIMF. Taken together, we conclude that (i) DNA replication timing is a robust genomic feature affecting SNS frequencies in both cancer and personal genomes, after controlling for many variables such as GC percentage, gene density, recombination rate, higher-order DNA replication timing domains, CTCF-binding sites, secondary structures and lincRNAs; (ii) SNS display specific patterns in early versus late DNA replication timing regions; and (iii) higher-order nuclear organization, together with DNA replication timing, affects the mutation frequencies. Furthermore, we found that in general, genome-wide mutation frequencies were lower than AIMFs. The exceptions in prostate cancer and small cell lung cancer could be because of an excess of mutations in repeat elements observed in our analysis, because the majority of the regions excluded from the genome to determine the AIMF were genes, promoters and repeat elements. The overall higher genome-wide mutation frequency in late replication timing regions also holds after controlling for several genomic features.

The higher SNS frequencies in late DNA replication timing zones in cancer genomes could partly arise from the accumulation of single-stranded DNAs, given similar observations in our analyses and others[18] and given that a certain fraction of regions harbouring mutations overlapped between cancer and personal genomes (Supplementary Fig. S45). DNA repair processes can often repair the errors arising during replication[45], and it has been suggested that both DNA replication timing and the efficiency of DNA repair are related to higher-order chromatin structure[45,46]. Our findings suggest that some portions of the genome have similar mutation frequencies as their counterparts residing closely within the three-dimensional structure of the nucleus. Chromatin organization and replication timing are intertwined, and could be a driving force of carcinogenesis by disrupting specific processes such as replication initiation and replication fork progression[46]. However, because most mutations analysed reside in noncoding parts of the genome, these patterns might only have indirect applicability to an understanding of the origins of cancer. Our study represents a novel approach to study the replication process-related SNS in cancer genomes together with the higher-order nuclear organization of the genome. This approach can lead to a better understanding of the mutational landscape of cancer genomes from the perspective of replication, epigenetics and chromatin structure.

Methods

Data sets and analyses. Cancer types and sample numbers analysed are listed in Table 1. All analyses were performed using human genome version hg18 as the reference genome. To obtain the FIRs, we employed a similar approach as was used by two other studies[18,47]. We removed all Refseq genes and promoters (up to 2 kb upstream of a gene), ultra-conserved elements with a conservation score >300 and also intronic sequences, which are related to transcription-coupled DNA repair. We also excluded repeat elements, centromeres and telomeres to minimize variant calling complexity in these regions[48], as well as the Y chromosome. All of these data were downloaded from the UCSC genome browser from the NCBI36/hg18 human genome[49]. The remaining genomic regions were termed FIRs. The total length of FIRs was approximately 780 Mb. We then overlaid the DNA replication timing data obtained from Hansen *et al.*[23] onto the FIRs and found that 79.23 Mb and 169.50 Mb of the FIRs resided within replicating regions that were consistently early or late, respectively, across multiple cell types. The human GC percentage, CpG island and recombination rate data were also obtained from UCSC genome browser. Because highly compact heterochromatin stains for Giemsa, whereas euchromatin is often unstained, we were able to characterize euchromatin and heterochromatin states globally across different cell types using Giemsa staining data[50]. Data on nuclear lamina-associated domains from Guelen *et al.*[11] were obtained from the NCBI Gene Expression Omnibus database, accession code GSE8854. Genomic regions harbouring nuclear lamina-associated domains are referred to as the nuclear periphery, whereas the remaining regions are referred to as nuclear core. When analysing the effects of lamina-associated domains on the mutation patterns, we used a bootstrap sampling approach (Supplementary Fig. S13) to take into account the variability of nuclear topology across different cell types. The Hi-C data for GM06990 and K562 cell lines were obtained from Lieberman-Aiden *et al.*[12] through the Gene Expression Omnibus database. Moreover, data on genome-wide common fragile sites were obtained from Durkin and Glover.[39] The G-quadruplex and CTCF-binding site locations were obtained from Quadruplex.org[51] and CTCFBSDB[52], respectively. The lincRNA catalogue can be obtained from http://www.broadinstitute.org/genome_bio/human_lincrnas[43]. All statistical calculations were performed using open source R software. When necessary, 'liftover' software was used to map data from other human genome versions to hg18.

References

1. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011).
2. Bignell, G. R. *et al.* Signatures of mutation and selection in the cancer genome. *Nature* **463,** 893–898 (2010).
3. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,** 719–724 (2009).
4. Podlaha, O., Riester, M., De, S. & Michor, F. Evolution of the cancer genome. *Trends Genet.* **28,** 155–163 (2012).
5. Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instabilities in human cancers. *Nature* **396,** 643–649 (1998).
6. Heng, H. H. *et al.* Evolutionary mechanisms and diversity in cancer. *Adv. Cancer Res.* **112,** 217–253 (2011).
7. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).
8. Surralles, J., Ramirez, M. J., Marcos, R., Natarajan, A. T. & Mullenders, L. H. Clusters of transcription-coupled repair in the human genome. *Proc. Natl Acad. Sci. USA* **99,** 10571–10574 (2002).
9. Holmquist, G. P. Chromosome bands, their chromatin flavors, and their functional features. *Am. J. Hum. Genet.* **51,** 17–37 (1992).
10. Hodgkinson, A., Chen, Y. & Eyre-Walker, A. The large-scale distribution of somatic mutations in cancer genomes. *Hum. Mutat.* **33,** 136–143 (2012).
11. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453,** 948–951 (2008).
12. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).
13. Yaffe, E. *et al.* Comparative analysis of DNA replication timing reveals conserved large-scale chromosomal architecture. *PLoS Genet.* **6,** e1001011 (2010).
14. Ryba, T. *et al.* Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.* **20,** 761–770 (2010).
15. Jeon, Y. *et al.* Temporal profile of replication of human chromosomes. *Proc. Natl Acad. Sci. USA* **102,** 6419–6424 (2005).
16. Woodfine, K. *et al.* Replication timing of the human genome. *Hum. Mol. Genet.* **13,** 191–202 (2004).
17. Gilbert, D.M. Replication timing and transcriptional control: beyond cause and effect. *Curr. Opin. Cell Biol.* **14,** 377–383 (2002).
18. Stamatoyannopoulos, J. A. *et al.* Human mutation rate associated with DNA replication timing. *Nat. Genet.* **41,** 393–395 (2009).
19. De, S. & Michor, F. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. *Nat. Biotechnol.* **29,** 1103–1108 (2011).

20. 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).
21. Schuster-Bockler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488,** 504–507 (2012).
22. Hammerman, P. S. *et al.* Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489,** 519–525 (2012).
23. 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).
24. Weddington, N. *et al.* ReplicationDomain: a visualization tool and comparative database for genome-wide replication timing data. *BMC Bioinformatics* **9,** 530 (2008).
25. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463,** 191–196 (2010).
26. Berger, M. F. *et al.* Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature* **485,** 502–506 (2012).
27. Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470,** 214–220 (2011).
28. Pleasance, E. D *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463,** 184–190 (2010).
29. Puente, X. S. *et al.* Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475,** 101–105 (2011).
30. Bass, A. J. *et al.* Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nat. Genet.* **43,** 964–968 (2011).
31. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452,** 872–876 (2008).
32. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5,** e254 (2007).
33. Fujita, P. A. *et al.* The UCSC genome browser database: update 2011. *Nucleic Acids Res.* **39,** D876–882 (2011).
34. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).
35. Ryba, T. *et al.* Abnormal developmental control of replication-timing domains in pediatric acute lymphoblastic leukemia. *Genome Res.* **22,** 1833–1844 (2012).
36. Pope, B. D. *et al.* DNA replication timing is maintained genome-wide in primary human myoblasts independent of D4Z4 contraction in FSH muscular dystrophy. *PLoS One* **6,** e27413 (2011).
37. Watanabe, Y. *et al.* Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: disease-related genes in timing-switch regions. *Hum. Mol. Genet.* **11,** 13–21 (2002).
38. McMurray, C. T. DNA secondary structure: a common and causative factor for expansion in human disease. *Proc. Natl Acad. Sci. USA* **96,** 1823–1825 (1999).
39. Durkin, S. G. & Glover, T. W. Chromosome fragile sites. *Annu. Rev. Genet.* **41,** 169–192 (2007).
40. Gondor, A. & Ohlsson, R. Replication timing and epigenetic reprogramming of gene expression: a two-way relationship? *Nat. Rev. Genet.* **10,** 269–276 (2009).
41. Tsai, M. C., Spitale, R. C. & Chang, H. Y. Long intergenic noncoding RNAs: new links in cancer progression. *Cancer Res.* **71,** 3–7 (2011).
42. Manolio, T. A., Brooks, L. D. & Collins, F. S. A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* **118,** 1590–1605 (2008).
43. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25,** 1915–1927 (2011).
44. Karolchik, D. *et al.* The UCSC genome browser database: 2008 update. *Nucleic Acids Res.* **36,** D773–779 (2008).
45. 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).
46. Alabert, C. & Groth, A. Chromatin replication and epigenome maintenance. *Nat. Rev. Mol. Cell Biol.* **13,** 153–167 (2012).
47. Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K. D. & Wray, G. A. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat. Genet.* **39,** 1140–1144 (2007).
48. Mardis, E.R. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9,** 387–402 (2008).
49. Miller, W. *et al.* 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res.* **17,** 1797–1808 (2007).
50. Furey, T. S. & Haussler, D. Integration of the cytogenetic map with the draft human genome sequence. *Hum. Mol. Genet.* **12,** 1037–1044 (2003).
51. Huppert, J. L. & Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.* **33,** 2908–2916 (2005).
52. Bao, L., Zhou, M. & Cui, Y. CTCFBSDB: a CTCF-binding site database for characterization of vertebrate genomic insulators. *Nucleic Acids Res.* **36,** D83–87 (2008).
53. Wu, H., Caffo, B., Jaffee, H. A., Irizarry, R. A. & Feinberg, A. P. Redefining CpG islands using hidden Markov models. *Biostatistics* **11,** 499–514 (2010).
54. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).
55. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat. Genet.* **31,** 241–247 (2002).
56. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380,** 152–154 (1996).
57. Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63,** 861–869 (1998).

Acknowledgements

We gratefully acknowledge support from the Dana-Farber Cancer Institute Physical Sciences-Oncology Centre (NCI U54CA143798), as well as feedback and advice from the Michor lab (michorlab.dfci.harvard.edu).

Author contributions

L.L., S.D. and F.M. conceived the experiments and wrote the paper. L.L. performed the analyses.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Liu L *et al.* DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nat. Commun.* 4:1502 doi: 10.1038/ncomms2502 (2013).