Nucleosome positioning stability is a significant modulator of germline mutation rate variation across the human genome

Cai Li[1,#] and Nicholas M. Luscombe[1,2,3]

[1] The Francis Crick Institute, 1 Midland Road, London, NW1 1AT, UK

[2] Okinawa Institute of Science & Technology Graduate University, Okinawa, 904-0495, Japan

[3] UCL Genetics Institute, University College London, Gower Street, London, WC1E 6BT, UK

Correspondence: cai.li@crick.ac.uk

Abstract

Understanding the patterns and genesis of germline *de novo* mutations is important for studying genome evolution and human diseases. Nucleosome organization is suggested to be a contributing factor to mutation rate variation across the genome. However, the small number of published *de novo* mutations and the low resolution of earlier nucleosome maps limited our understanding of how nucleosome organization affects germline mutation rates in the human genome. Here, we systematically investigated the relationship between nucleosome organization and fine-scale mutation rate variation by analyzing >300,000 *de novo* mutations from whole-genome trio sequencing and high-resolution nucleosome maps in human. We found that *de novo* mutation rates are elevated around strong, translationally stable nucleosomes, a previously under-appreciated aspect. We confirmed this observation having controlled for local sequence context and other potential confounding factors. Analysis of the underlying mutational processes suggests that the increased mutation rates around strong nucleosomes are shaped by a combination of low-fidelity replication, frequent DNA damage and insufficient/error-prone repair in these regions. Interestingly, strong nucleosomes are preferentially located in young SINE/LINE elements, implying frequent nucleosome re-positioning (i.e. shifting of dyad position) and their contribution to hypermutation at new retrotransposons during evolution. These findings provide novel insights into how chromatin organization affects germline mutation rates and have important implications in human genetics and genome evolution.

1 Introduction

Germline *de novo* mutations, which arise during gametogenesis and can be passed to offspring, are the primary source of genetic variation, contributing substantially to biological diversity and evolution. *De novo* mutations are also thought to play significant roles in early-onset genetic disorders such as intellectual disability, autism spectrum disorder, and other developmental diseases (Veltman and Brunner 2012; Acuna-Hidalgo et al. 2016). Thus, investigating the patterns and genesis of *de novo* mutations in the germline is important for understanding genome evolution and human diseases.

It is well known that germline and somatic mutation rates vary across the human genome over several scales of resolution, from nucleotide- to whole chromosome levels (Hodgkinson and Eyre-Walker 2011; Segurel et al. 2014). Previous studies of germline and somatic mutations (mainly cancer genomic variants) revealed several factors linked to mutation rate variation across the human genome, including local sequence context (Michaelson et al. 2012), replication timing (Stamatoyannopoulos et al. 2009), recombination rate (Francioli et al. 2015), DNA accessibility (Sabarinathan et al. 2016) and histone modifications (Michaelson et al. 2012; Schuster-Bockler and Lehner 2012). Despite these findings, however, two recent studies (Terekhanova et al. 2017; Smith et al. 2018) reported that less than 40% of local mutation rate variation at 100Kb to 1Mb resolution can be explained by the genomic features identified so far, implying that unknown but important factors remain to be found. Moreover, the underlying mutational processes causing the observed mutation rate variation remain poorly understood. A few recent studies pointed to the contribution of error-prone replicative processes (Harris and Nielsen 2014; Lujan et al. 2014; Reijns et al. 2015; Seplyarskiy et al. 2017; Seplyarskiy et al. 2018) and differential DNA repair efficiency (Supek and Lehner 2015; Perera et al. 2016; Sabarinathan et al. 2016; Frigola et al. 2017). Despite these advances, it remains a big challenge to understand the mutational mechanisms associated with mutation rate variation, particularly in the germline.

Chromatin organization has been reported to contribute to regional variation in mutation rates (Makova and Hardison 2015) because of structural constraints on the mutability of genomic sequences. In particular, nucleosome positioning was suggested as a significant factor in humans and other eukaryotes (Sasaki et al. 2009; Tolstorukov et al. 2011; Chen et al. 2012; Michaelson et al. 2012; Lujan et al. 2014; Pich et al. 2018). Studies in different lineages (Sasaki et al. 2009; Tolstorukov et al.

2011; Lujan et al. 2014) reported increased substitution rates around the centers of nucleosomal sequences and increased insertion/deletion rates in nucleosome-free DNA. However, there are also disagreements between published studies. For example, Michaelson et al. (2012) suggested that high nucleosome occupancy tends to suppress *de novo* mutations, but Smith et al. (2018) found that a comparative analysis using datasets from different studies resulted in opposing conclusions, likely affected by different variant calling strategies. Due to the limited availability of germline *de novo* mutations for humans, most previous studies focused on coarse-grained mutation rate variation (e.g. ≥1kb windows), or used within-species polymorphisms and inter-species divergence for analysis. The drawback of using polymorphism and divergence data is that they might be affected by natural selection and other evolutionary processes such as GC-biased gene conversion. In addition, because of the limitation of available nucleosome maps, some previous studies treated all annotated nucleosomes equally, ignoring the diverse contexts in which they form. Therefore, the effects of nucleosome organization on germline mutation rate variation, particularly at high resolution remain to be elucidated. Here we take advantage of the rapid increase in the number of germline *de novo* mutation datasets and better understanding of nucleosome organization in the human genome to perform a systematic analysis of this topic.

2 Results

2.1 Datasets used for analysis

We used germline *de novo* mutations identified in multiple large-scale trio sequencing projects; these total >300,000 single-nucleotide variants (SNVs) and >30,000 short insertions/deletions (INDELs) (**Fig. 1a, Supplementary Fig.1a**), after removing those in genomic regions that could confound downstream analysis (see Methods). Most of these *de novo* mutations originated from three studies (Jonsson et al. 2017; Turner et al. 2017a; Yuen et al. 2017), each presenting about 100,000 mutations. In addition to *de novo* mutations from trio sequencing, we also examined extremely rare variants (allele frequency ≤ 0.0001) from the gnomAD database (Lek et al. 2016), which are thought to be subject to limited selection and non-adaptive evolutionary processes, and so can be approximated to *de novo* mutations (Carlson et al. 2018).

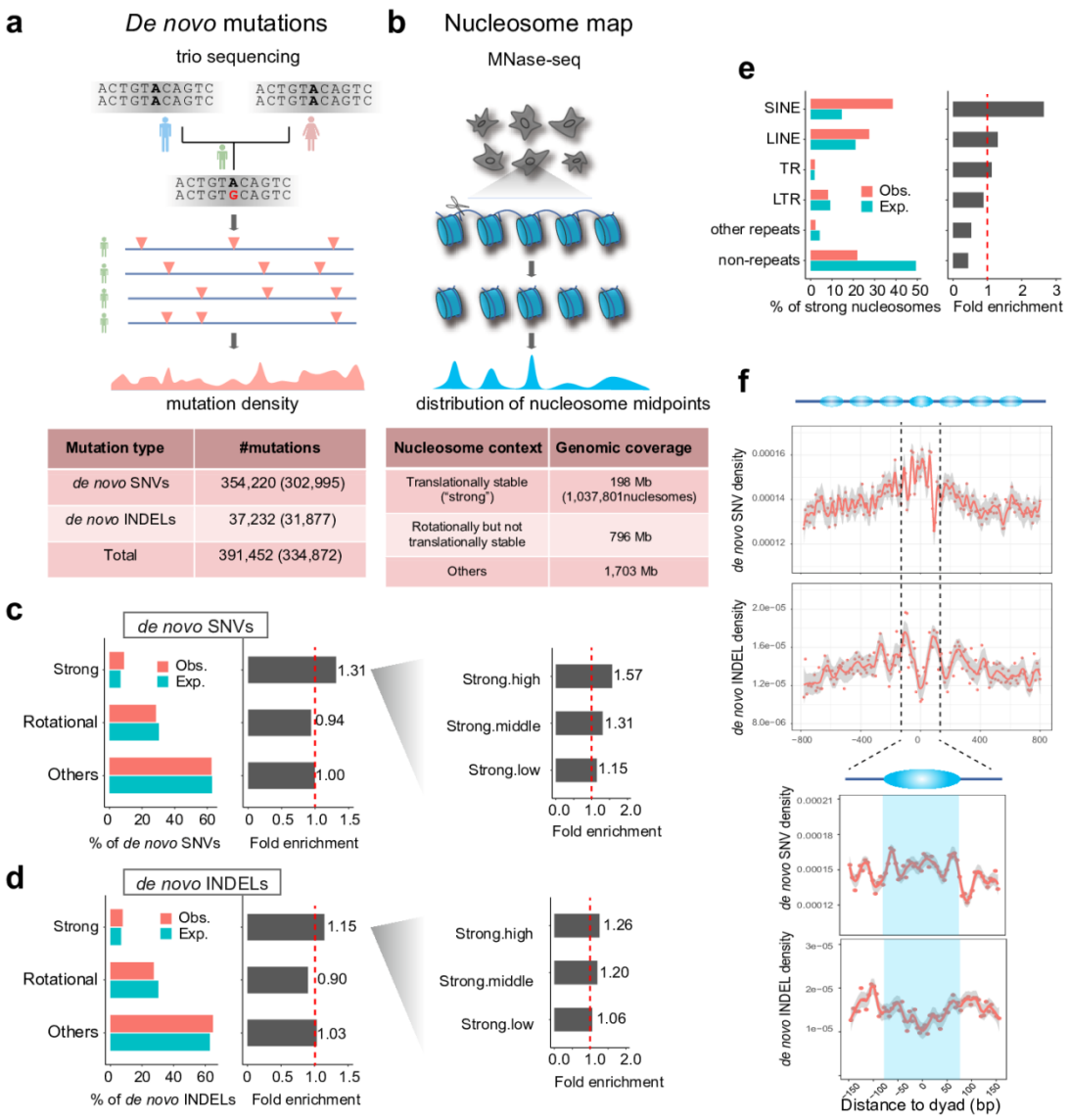


Fig. 1 *De novo* mutations in different nucleosome contexts. (**a**) Germline *de novo* mutation data used for analysis. Numbers of *de novo* mutations after filtering out those in confounding regions are shown in the brackets. (**b**) Illustration of MNase-seq method and the nucleosome datasets used in this analysis. (**c**) Fold enrichment/depletion of *de novo* SNVs in different nucleosome contexts. 'Strong', translationally stable positioning; 'Rotational', rotationally but not translationally stable positioning; 'Others', the remaining genomic regions. On the left is the fold enrichment for three subgroups of strong nucleosomes with different stabilities. (**d**) Fold enrichment/depletion of *de novo* INDELs in different nucleosome contexts. (**e**) Fold enrichment/depletion of strong nucleosomes in different repeat sequence contexts. SINE, short interspersed nuclear element; LINE, long interspersed nuclear element; TR, tandem repeat; LTR, long terminal repeat. (**f**) Above, meta-profile of *de novo* SNV/INDEL densities, centered at dyads of strong nucleosomes. Below, the same meta-profile zoomed in on the middle nucleosome in the array.

123 Nucleosome positioning is described by the translational setting, which defines the
124 location of the nucleosomal midpoint (also called 'dyad'), and the rotational setting,
125 which defines the orientation of the DNA helix on the histone surface (Gaffney et al.
126 2012). Using MNase-seq measurements across seven lymphoblastoid cell lines,
127 Gaffney et al. (2012) previously identified about a million 'strong' nucleosomes that
128 adopt highly stable translational positions across the sequenced cell lines (covering
129 198Mb of the genome). We also considered rotationally stable nucleosomes
130 identified in DNase-seq samples across 43 cell types (Winter et al. 2013), covering
131 892Mb of the genome. There is a ~50Mb overlap between regions bound by strong
132 nucleosomes and rotationally stable nucleosomes.

133 We divided the human genome (sex chromosomes excluded) into three categories
134 based on the type of nucleosome context (**Fig. 1b**): i) regions containing
135 translationally stable ('strong') nucleosomes (198Mb); ii) regions with rotationally but
136 not translationally stable nucleosomes (796Mb); and iii) all other non-N base genomic
137 regions (1,703Mb). Although nucleosome positioning can vary between cell types,
138 particularly in transcriptionally regulated regions, West et al. (2014) suggested that
139 such variation is very limited overall. As the nucleosome maps were not derived from
140 germ cells, we excluded nucleosomes that were found to differ in positioning
141 between cell types (~23Mb; see Methods).

142 **2.2 *De novo* mutations in different nucleosome contexts**

143 We first checked the numbers of mutations in the different genomic regions defined
144 above. Regions associated with strong nucleosomes have ~30% more *de novo*
145 SNVs (**Fig. 1c**) and ~15% more *de novo* INDELs (**Fig. 1d**) relative to expectations,
146 whereas regions containing rotationally stable nucleosomes are slightly depleted of
147 both mutation types. These enrichments are also apparent for the extremely rare
148 variants from gnomAD (**Supplementary Fig. 1b,c**), though effect sizes are smaller
149 compared with *de novo* mutations; the smaller enrichment here could be due to the
150 fact that highly mutable sites are under-represented among extremely rare variants
151 owing to higher allele frequencies in the population (Harpak et al. 2016). When
152 separating the one million strong nucleosomes to three subgroups based on the
153 translational positioning stability scores defined by Gaffney et al., we found that the
154 nucleosomes with higher stability exhibit higher mutation rates (**Fig. 1c,d**),
155 suggesting the nucleosome translational positioning stability (termed positioning
156 stability for short hereafter) is associated with the mutation rate variation across the
157 genome.

158 We further noticed that ~80% of strong nucleosomes overlap with repetitive elements,
159 especially short interspersed nuclear elements (SINEs) and long interspersed
160 nuclear elements (LINEs) (**Fig. 1e**). Genetic variations in repetitive sequences are
161 difficult to detect because of poor mappability, leading to a high proportion of false
162 negatives (Lee and Schatz 2012). Thus the enrichments of *de novo* mutations and
163 rare polymorphisms in strong nucleosomal regions are probably underestimated.

164 Interesting mutation patterns also emerge at higher resolutions. Meta-profiles of
165 mutation densities relative to the nucleosomal dyads reveal increased SNV density
166 and reduced INDEL density compared with the flanking linker regions (**Fig. 1f**),
167 resembling previous observations based on polymorphism data and a much smaller
168 number of nucleosomes (Tolstorukov et al. 2011). Since many strong nucleosomes
169 are associated with Alu elements and LINE-1 (L1) elements (Alu and L1 are the most
170 abundant SINE and LINE families in the human genome, respectively), we separated
171 the strong nucleosomes into three groups: i) Alu-associated, ii) L1-associated and iii)
172 others. Alu-associated nucleosomes display increased SNV rates around the dyads
173 (**Supplementary Fig. 1d**), whereas non-Alu nucleosomes show increased SNV rates
174 ~60bp from the dyads, close to the nucleosome edges. In contrast, the patterns of
175 INDEL densities are relatively similar among the different groups (**Supplementary**
176 **Fig. 1d**).

177 **2.3 Statistical tests confirm the significant association between nucleosome**
178 **positioning stability and mutation rate variation**

179 Local sequence contexts can affect mutation rates; for instance, CpG sites are
180 known to be highly mutable (**Supplementary Fig. 1d**). Other factors such as DNA
181 methylation, histone modification, chromatin accessibility, replication timing and
182 recombination rate have also been reported to affect mutation rates. Therefore, to
183 control for potential confounding factors, we used a logistic regression framework to
184 systematically assess the contribution of nucleosome positioning to mutation rate
185 variation.

186 We defined three variables to quantify nucleosomal properties relative to a specific
187 nucleotide position. The first two relate to the translational positioning of nearby
188 nucleosomes: i) d_{mean}, the mean distance between the focal site and the midpoints of
189 mapped MNase-seq fragments (we considered MNase-seq midpoints within ±95bp of
190 the focal site, corresponding to the average nucleosome repeat length reported by
191 Gaffney et al. 2012) and ii) d_{var}, the variance of these distances (**Fig. 2a**). A smaller
192 d_{mean} indicates that the site is closer to the nucleosome dyad and smaller d_{var}

193 represents more stable translational positioning around that site. Because the
194 relationship between d_{mean} and SNV rates is not monotonic nor linear (see **Fig. 1f,**
195 **Supplementary Fig. 1d**), we considered d_{mean} be a categorical variable by binning
196 values into five intervals (see Methods). The third variable is nucleosome occupancy,
197 calculated as the per-base MNase-seq fragment coverage after correcting for
198 sequencing and mapping biases (see Methods).

199 We also considered the local nucleotide sequence (5bp upstream and downstream of
200 the focal site), RNA expression in spermatogonial stem cells (SSCs), DNA
201 methylation (SSCs), chromatin accessibility (ATAC-seq in SSCs), eight histone
202 modifications (ENCODE ChIP-seq in ESCs), replication timing (Repli-seq in
203 GM12878) and recombination rates (HapMap project). We found that nucleosome
204 positioning stability (d_{var}) has a relatively weak but statistically significant correlation
205 with many of these factors (**Supplementary Fig. 2**), suggesting non-independence of
206 factors; therefore, it is important to control for them in assessing the contribution of
207 positioning stability to mutation rate variation.

208 We compared the full logistic regression model encompassing all variables with
209 reduced models missing individual variables using likelihood-ratio tests (**Fig. 2b,c**;
210 see Methods for more detail). As CpG sites are highly mutable and contribute to a
211 large fraction of mutations, we tested CpG and non-CpG sites separately (**Fig. 2b,c**).
212 We found that positioning stability (d_{var}) is significantly associated with mutation rate
213 variation in both CpG and non-CpG contexts while controlling for other factors.
214 Generally, the effects of positioning stability on mutation rates for A/T sites are
215 greater than for C/G sites, and the effects on SNV mutation rates more significant
216 than for INDELs. The lower significance for INDELs could partly be due to the fact
217 that there are fewer such *de novo* mutations available for modeling. More specifically,
218 among SNV mutation rates, we found highly significant contributions of positioning
219 stability for all mutation types in CpG and non-CpG contexts (**Supplementary Fig. 3**),
220 except for CpG>TpG mutations (adjusted $p = 0.10$). Notably, the regression
221 coefficients for d_{var} are always negative (**Supplementary Table 1**), indicating that
222 positioning stability is inversely associated with mutation rates, corroborating the
223 patterns observed in **Fig. 1**.

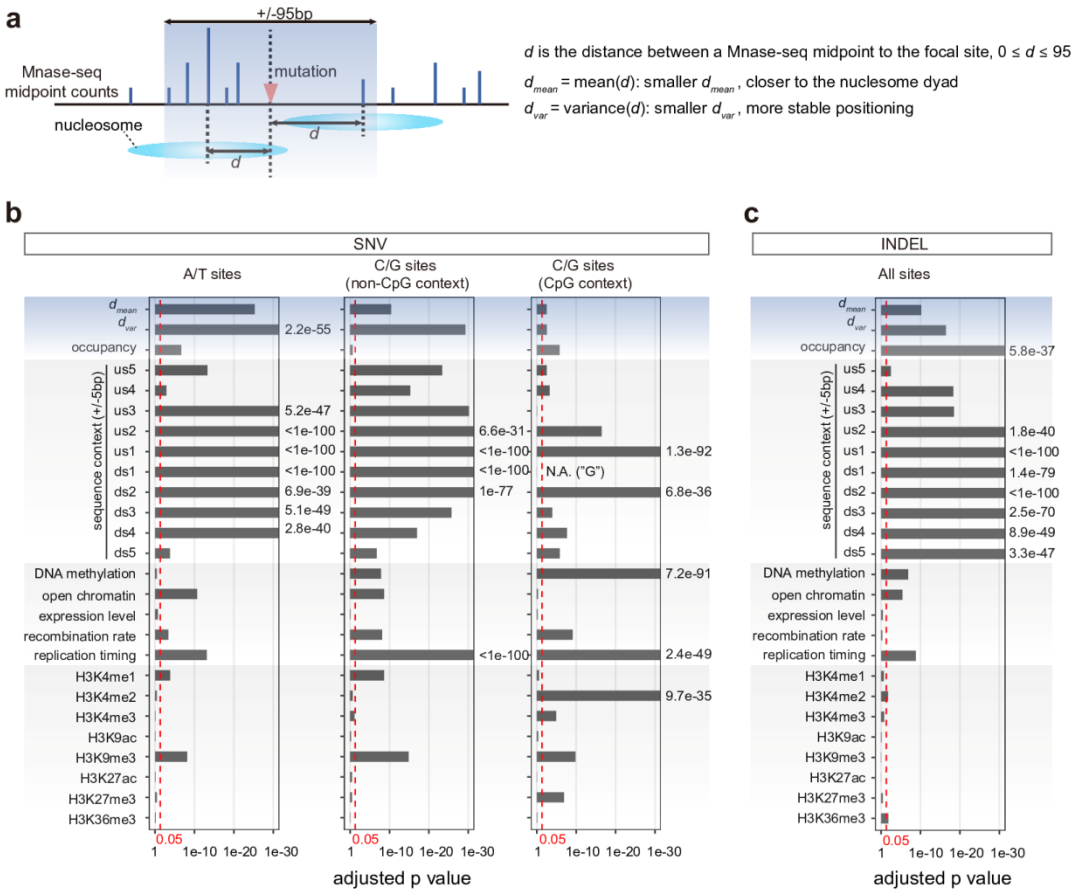


Fig. 2 Statistical tests for evaluating contribution of nucleosome organization to mutation rate variation. (**a**) Definition of two nucleosome positioning-related variables (d_{mean} and d_{var}) for a given site in the genome. (**b-c**) Statistical significance of each variable by performing likelihood ratio tests for assessing their contribution to mutation rate variation: **b** for *de novo* SNVs and **c** for *de novo* INDELs. For *de novo* SNVs, we performed tests for A/T sites, C/G sites in non-CpG context and C/G sites in CpG context separately. The red vertical lines represent the significance cut-off (0.05) for the adjusted p values (Bonferroni-Holm correction). 'us', upstream; 'ds', downstream.

The mean distance to the dyads, d_{mean}, also has statistically significant effects on SNV/INDEL mutation rates (**Fig. 2b,c**). The relationship with SNVs (but not INDELs) is non-linear, as evidenced by the loss of significance upon treating d_{mean} as a continuous variable (**Supplementary Fig. 4**). Nucleosome occupancy is also statistically significant; in contract to the positioning variables however, here the effect is much larger for INDELs than SNVs (**Fig. 2b,c;** INDELs, adjusted p = 5.8e-37; SNVs, adjusted p = 0.21, 1.6e-6, 2.2e-7).

In agreement with previous observations on rare variants (Carlson et al. 2018), local sequence context appears to be the biggest contributing factor for local mutation rate

242 variation (**Fig. 2b,c**). Contributions generally decrease with increasing distance to the
243 surveyed site. Although the larger sequence context could still have influence on
244 mutation rates, including more distant bases can lead to some issues such as
245 overfitting in regression models. We appreciate that nucleosome positioning stability
246 is at least partly determined by the occupied DNA sequence and thus its effects on
247 mutation rates to some degree can be attributed to the associated sequence (this
248 also applies to other reported factors such as replication timing). However, higher-
249 order interactions among the long stretches of nucleotides which guide nucleosome
250 positioning are difficult to model properly. We achieved similar statistical significance
251 for positioning stability after including non-additive two-way interaction effects for ±5
252 nucleotides and the 7-mer mutability estimates from Carlson et al. (**Supplementary**
253 **Fig. 5**).

254 Finally in addition to sequence context, we recapitulated the effects of previously
255 reported factors (**Fig. 2b,c, Supplementary Fig. 3**). DNA methylation and H3K9me3
256 are two common epigenetic signatures that statistically significantly contribute to
257 mutation rate variation (Schuster-Bockler and Lehner 2012). H3K4me1, H3K4me2,
258 H3K4me3 and H3K27me3 are also statistically significant for specific mutation types.
259 Replication timing has highly statistically significant effects on both SNV and INDEL
260 mutation types. Recombination rate and open chromatin (measured by ATAC-seq)
261 also have significant effects on many mutation types. However, transcription levels
262 appear to have no effect on local mutation rates according to our result.

263 Taken together, our modeling analysis recapitulated previously reported factors
264 affecting mutation rates and confirmed the importance of nucleosome positioning
265 stability as another contributing factor.

266 2.4 Mutational processes associated with elevated mutability around strong 267 nucleosomes

268 2.4.1 Mutational signature analysis

269 Next we sought to investigate the underlying mutational processes that explain the
270 elevated mutation rates associated with strong nucleosome positioning. For this, we
271 first used the MutationalPatterns software (Blokzijl et al. 2018) to calculate the
272 contribution of different COSMIC mutational signatures to germline *de novo* SNVs.
273 Three major signatures (Signatures 1, 5 and 16) explain 88% of all *de novo* SNVs
274 (**Fig. 3a**); however, Signatures 3 (3.4%), 6 (14%), 7 (2.8%) and 12 (9.2%) are
275 enriched only among SNVs associated with strong nucleosomes (**Fig. 3a**). There are

further differences between nucleosome-associated SNVs overlapping Alus, L1s and 'other' regions (**Fig. 3a);** for example, Signatures 6 and 12 make major contributions to Alu-associated SNVs (27% and 17% respectively) but less marked to 'Others' group (3.7% and 3.9%) and none to L1 group. Signature 3 is found in all three groups, but shows relatively higher contribution in the L1 group. The mutational signatures of L1-associated SNVs are similar to that of genome-wide background (**Fig. 3a**), which might be because (i) L1 elements make up a considerable fraction (>20%) of the whole genome and (ii) strong nucleosomes are under-detected in very young L1 elements owing to difficulties in mapping short MNase-seq reads (Alus are easier to map because they are smaller).

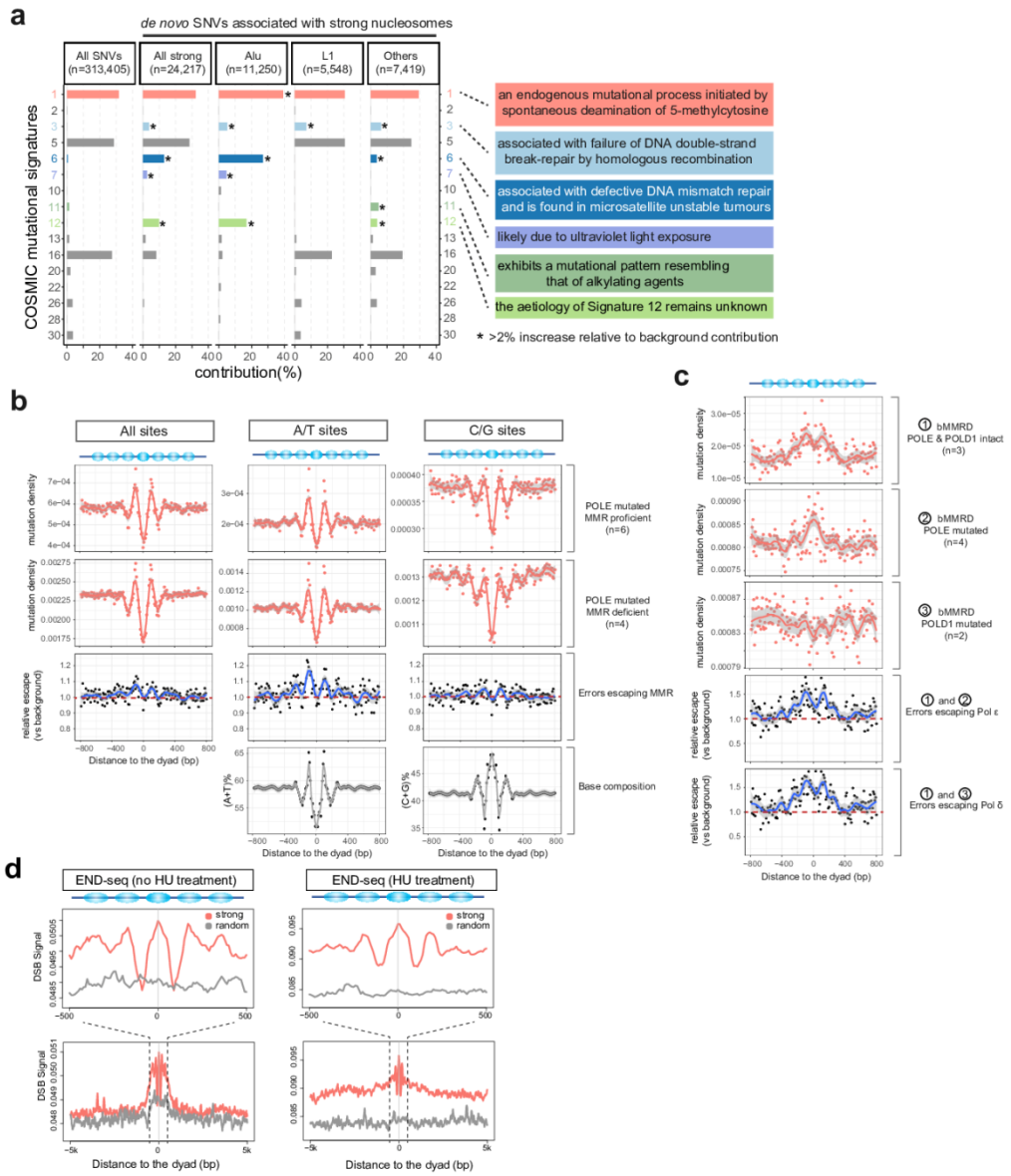


Fig. 3 Analysis of mutational processes related to strong nucleosomes. (**a**) Comparison of contribution of COSMIC mutational signatures in different sets of *de novo*

289 SNV data. Only signatures with non-zero values are shown. Asterisks indicate the
290 mutational signatures showing ≥2% increase relative to the background contribution (i.e.
291 'All SNVs'). On the right is brief information about aetiologies (from the COSMIC website)
292 of highlighted signatures. (**b**) Mutation profiles around strong nucleosomes for the
293 PCAWG cancer genomes harboring driver mutations in *POLE* and genes involved in
294 mismatch repair (MMR). Numbers of cancer genomes used are indicated in the brackets.
295 The relative escape ratios of MMR were estimated by comparing '*POLE*-mutated MMR
296 proficient' and '*POLE*-mutated MMR deficient' samples. The profiles are centered at the
297 annotated strong nucleosomes. (**c**) Mutation profiles around strong nucleosomes for
298 bMMRD cancer genomes and the estimated relative escape ratios of Pol ε or Pol δ. (**d**)
299 Signal of DSBs around strong nucleosomes based on the END-seq data. Only the strong
300 nucleosomes of high 75-mer mappability within ±500bp were considered. HU,
301 hydroxyurea.

302 Having assigned mutational signatures, we can infer possible underlying DNA
303 damage and repair mechanisms that explain the elevated local mutation rates (**Fig.**
304 **3a**). Signature 1, a common mutational process due to spontaneous deamination of
305 5-methylcytosine, displays increased contribution to the Alu-associated SNVs; this is
306 expected because young Alu elements harbor many CpG sites and are usually highly
307 methylated. For other suggested mutational processes in **Fig. 3a**, we took advantage
308 of multiple genomic datasets for further investigation (see below).

309 **2.4.2 Mismatch repair**

310 Signature 6, which contributes ~14% to strong nucleosome-associated SNVs, is
311 linked to mismatch repair (MMR). This is a major repair pathway active during DNA
312 replication: it mainly repairs mismatches and short INDELs introduced during DNA
313 synthesis that escaped polymerase proofreading. We assessed the impact of the
314 MMR pathway by analyzing somatic mutations from two sets of cancer genomes
315 (Campbell et al. 2017). The first possessed driver mutations in catalytic subunit gene
316 *POLE* of DNA polymerase ε (Pol ε, the major replicase for the leading strand) and
317 one or more of the core MMR genes (*MLH1*, *MSH2*, *MSH6*, *PMS1* and *PMS2*). The
318 second set comprised cancers with mutated *POLE* but intact MMR pathway.
319 Mutations in *POLE* in these cancers can lead to reduced base selectivity and/or
320 deficient proofreading during replication, producing unusually large numbers of
321 mutations (so called 'ultra-hypermutation') which facilitated our analysis. Because
322 detecting somatic mutations in tumors is more difficult than for ordinary individual re-
323 sequencing data and mutations in poor-mappability regions could be easily missed,
324 we focused on strong nucleosomes with high mappability (see Methods).

325 For samples with defective *POLE* and MMR, mutation rates around the strong
326 nucleosome boundaries and adjacent linker DNA are much higher than surrounding
327 regions (**Fig. 3b**). This suggests replication errors occur frequently around strong
328 nucleosomes. We detected a similar pattern for samples with mutated *POLE* but
329 proficient MMR. This implies that many errors introduced during error-prone
330 replication of mutated *POLE* also escaped the repair by MMR at strong nucleosomes,
331 which were then translated into somatic mutations.

332 By comparing the mutation rates in the two sets of *POLE*-mutated cancers, we were
333 able to estimate the MMR escape ratio, which quantifies the relative amount of
334 replication errors that remains unrepaired (see Methods for more detail). Strong-
335 nucleosome regions (especially boundaries and adjacent linkers) have ~10% higher
336 MMR escape ratio than the genome-wide background, suggesting that replication
337 errors at these regions are more likely converted to mutations in daughter cells.
338 Differences in the escape ratio in these regions cannot be ascribed to local base
339 composition, as both C/G sites and A/T sites exhibit similar patterns (**Fig. 3b**), though
340 A/T sites appear to have a higher ratio. Moreover, the apparent ~200-bp periodic
341 patterns in escape ratios and mutation densities (**Fig. 3b**) are suggestive of
342 associations with nucleosome positioning rather than sequence alone.

343 **2.4.3 DNA polymerase fidelity**

344 We also analyzed data from children with inherited biallelic mismatch repair
345 deficiency (bMMRD) (Shlien et al. 2015), which included ultra-hypermutated
346 genomes with mutations in *POLE* or the catalytic subunit gene *POLD1* of polymerase
347 δ (Pol δ, the major replicase for the lagging strand). The bMMRD data can be used
348 for investigating polymerase fidelity around strong nucleosomes. Here, we estimated
349 Pol ε and Pol δ escape ratios using the same method as that for MMR escape ratios
350 (**Fig. 3c**, see Methods). We found that strong nucleosomes have higher escape
351 ratios for both polymerases relative to the genomic background (**Fig. 3c**), implying
352 that replication polymerases have lower fidelity in these regions. These escape ratios
353 are even higher than that for MMR (**Fig. 3b,c**), and A/T sites also have larger escape
354 ratios than C/G sites (**Supplementary Fig. 6a**). In addition, the periodic pattern in the
355 relative escape profiles (**Fig. 3c, Supplementary Fig. 6a**) suggests that nucleosome
356 positioning contributes to the heterogeneity in replicase fidelity across the genome.

357 Signature 12 is another prominent mutational signature that contributes 9.2% to *de*
358 *novo* SNVs associated with strong nucleosomes, but whose aetiology is currently
359 unknown. By performing mutational signature analysis, we found that Signature 12

360 contributes 21.15%~21.99% to mutations in *POLD1*-mutated bMMRD samples
361 (**Supplementary Fig. 6b,c**), but much less for other bMMRD samples (0~2.88% for
362 *POLE*-mutated, and 3.32%~10.43% for *POLE*/*POLD1*-intact). This suggests that the
363 previously uncharacterized signature is likely associated with Pol δ, and that many *de*
364 *novo* mutations around strong nucleosomes arise from errors escaping Pol δ proof-
365 reading and post-replicative MMR during lagging strand replication. On the other
366 hand, Signature 10, the only signature known to be associated with Pol ε deficiency,
367 exhibits little contribution to strong nucleosome-associated SNVs (**Fig. 3a**). This
368 suggests that although both Pol ε and Pol δ have low fidelity around strong
369 nucleosomes (**Fig. 3c**), the majority of the replication errors that are eventually
370 converted to *de novo* mutations are derived from Pol δ.

371 Previous analysis in budding yeast suggested that Okazaki fragment-associated
372 mutagenesis during lagging strand replication contribute significantly to mutation rate
373 variation across the genome (Reijns et al. 2015). They found that in budding yeast
374 Okazaki junctions tend to have elevated mutation rates and are usually close to
375 nucleosome dyads. We tested this point in human by re-analyzing the OK-seq data
376 (Petryk et al. 2016), which sequenced the Okazaki fragments in human cells. We
377 found that, unlike that in yeast, Okazaki junctions in human appear to be relatively
378 more frequently located around the linker regions (**Supplementary Fig. 7**), rather
379 than the dyads, implying the mutagenic effects of Okazaki junctions in yeast are not
380 the same in human. The difference may be partly due to the reason that budding
381 yeast lacks the typical H1 histone found in human and other eukaryotes. However,
382 the very short reads (single-ended 50bp) of OK-seq data only allowed us to examine
383 the nucleosomes with high mappability with 50-mers (~10% of strong nucleosomes),
384 the distribution of Okazaki junctions in unexamined regions is unclear.

385 **2.4.4 Double-strand breaks**

386 Double-strand break (DSB) repair represented by Signature 12 is another potential
387 mechanism involved in strong nucleosome-associated mutations (**Fig. 3a**). A recent
388 study (Tubbs et al. 2018) investigating genome-wide DSBs using END-seq
389 suggested that poly(dA:dT) tracts are recurrent sites of replication-associated DSBs.
390 Our analysis of this data revealed a higher frequency of DSBs around strong
391 nucleosomes compared with genomic background (**Fig. 3d**). The trend holds for
392 experiments with and without hydroxyurea treatment (HU, a replicative stress-
393 inducing agent), suggesting that strong nucleosomes contribute to endogenous
394 replication stress and higher frequency of DSBs. It is notable that young Alu and L1

395 elements harbor prominent poly(dA:dT) tracts, which are enriched at the boundary
396 and linker regions of strong nucleosomes (**Supplementary Fig. 8a**). The patterns of
397 high DSB frequency still hold true when looking at strong nucleosomes associated
398 with different repeats (**Supplementary Fig. 8b-c**). Yet, because the END-seq data
399 was sequenced with single-ended 75bp reads and majority of young Alu and L1
400 elements can not be assessed with such short reads, we did not perform further
401 analysis regarding this. Since DSB repair can be error-prone (Rodgers and McVey
402 2016), even using high-fidelity homologous recombination, frequent DSB-formation
403 and subsequent error-prone repair likely contribute to the elevated mutation rates
404 around strong nucleosomes.

405 **2.5 Strong nucleosome positioning is mostly associated with young repeat**
406 **elements and undergoes frequent turnover**

407 ~70% of strong nucleosomes are located in SINE and LINE retrotransposons (**Fig.**
408 **1e**). Upon examination of the subfamilies (**Fig. 4a,b**), we uncovered specially strong
409 enrichments for evolutionarily young L1s (e.g. L1PA2 to L1PA11) and Alus (e.g. AluY
410 to AluSx). Since younger repeats have poorer mappability, these observations
411 probably underestimate the true enrichment. This may also explain why several of
412 the youngest L1 subfamilies (L1PA2 to L1PA5) have lower enrichments relative to
413 the slightly older subfamilies (**Fig. 4a**). It has long been reported that nucleosomes
414 preferentially bind specific segments of Alu elements, supported by both *in vitro* and
415 *in vivo* evidence (Englander et al. 1993; Englander and Howard 1995). Later studies
416 confirmed preferential nucleosome positions within Alus (Salih et al. 2008; Tanaka et
417 al. 2010). These studies also suggested that A-rich sequences and CpG dinucleotides
418 in Alus likely play important roles in the unique positioning patterns. We
419 recapitulated their findings using MNase-seq data of Gaffney et al. (**Fig. 4c**) with two
420 hotspots of strong nucleosomes in young Alus, which fade away in older elements.
421 Young Alu elements exhibit elevated *de novo* mutation rates (**Fig. 4c**), which could
422 be partly contributed by the associated strong nucleosome positioning as evaluated
423 in previous sections.

424 The weakening of positioning stability in older Alu elements is accompanied by
425 reduced *de novo* mutation rates for both SNVs and INDELs, although the pattern in
426 INDELs is less clear compared with SNVs (**Fig. 4c**). We hypothesize that the
427 sequence changes in young Alu elements arising from *de novo* mutations might
428 affect the stability of nucleosome positioning. The histone octamer is thought to
429 preferentially bind DNA sequences presenting lower deformation energy costs

(Tolstorukov et al. 2008). Using the nuScore software (Tolstorukov et al. 2008) to estimate deformation energies based on the DNA sequence and nucleosome core particle structure, we indeed found that Alus exhibit lower deformation energies than surrounding regions. Furthermore, the deformation energies increased in older subfamilies, suggesting that the evolutionary changes in Alu sequences reduced their nucleosome-binding affinity. We further analyzed the 3' end sequences of L1 elements harboring strong nucleosomes and observed similar patterns (**Supplementary Fig. 9**).

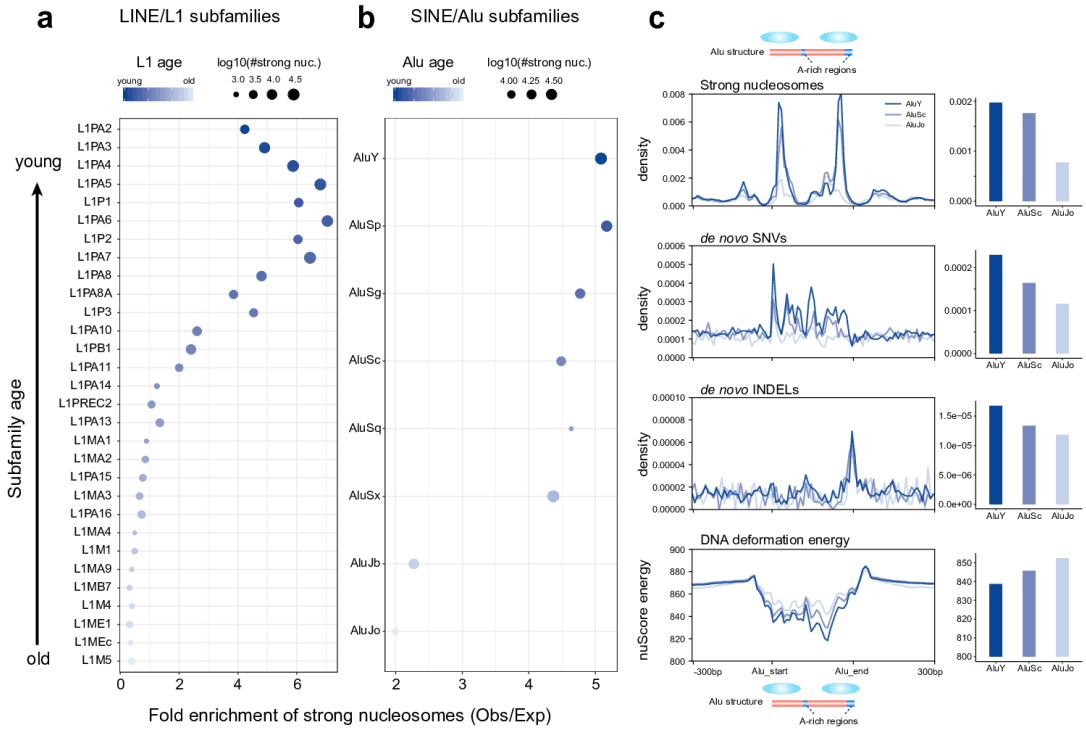


Fig. 4 Strong nucleosomes are mostly associated with young LINE/SINE elements. (**a**) Fold enrichment/depletion of strong nucleosomes in LINE/L1 subfamilies of different ages (top 30 abundant subfamilies are shown), with dot sizes and colors depicting numbers of associated strong nucleosomes and subfamily ages respectively. (**b**) Fold enrichment/depletion of strong nucleosomes in SINE/Alu subfamilies of different ages. (**c**) Densities of strong nucleosome dyads, *de novo* SNVs and *de novo* INDELs along the Alu sequence and surrounding regions for Alu subfamilies of different ages, with the average values for all Alu regions of different subfamilies in the barplots on the right. Two major loci for strong nucleosomes are shown at the top. The deformation energy estimated by nuScore is shown at the bottom panel. Alu elements of >=250bp were used for plotting the meta-profiles and all elements were scaled to a 300bp region in the plots. The structure of the Alu consensus sequence is shown at the top and bottom.

Several studies investigating genetic changes around nucleosomes suggested that natural selection appears to preserve nucleosome positioning during evolution (Prendergast and Semple 2011; Tolstorukov et al. 2011; Drillon et al. 2016), but they had different opinions on how selection acts on the sequence. On the other hand, another model (Warnecke et al. 2013) suggested that the observed sequence divergence patterns around nucleosomes can be explained by frequent nucleosome re-positioning after mutations, rather than by natural selection. These studies were mainly based on human polymorphisms or inter-species divergence and so were potentially confounded by mutation rate heterogeneity associated with nucleosomes and other non-adaptive processes such as GC-biased gene conversion. The use of germline *de novo* mutations helps resolve this debate. As we showed before, there is considerable *de novo* mutation rate variation around strong nucleosomes (**Fig. 1f**, **Supplementary Fig. 1**), which can not be ignored in any selection analysis. Furthermore, analysis on strong nucleosomes associated with different SINE/LINE subfamilies clearly showed that strong positioning is preferentially present in young elements and the positioning stability decays substantially over time (**Fig. 4**), which strongly supports the re-positioning model. Since LINE/SINE elements make up a large majority of strong nucleosomes in the genome, selection for preserving positioning is unlikely widespread as previously suggested, though it may happen at some particular regions.

3 Discussion

While nucleosome organization is long known to be involved in DNA damage/repair processes (Smerdon 1991), its effects on germline mutation rate variation across the genome (particularly in higher eukaryotes) are poorly understood, largely due to the limited available *de novo* mutations and poor-resolution of nucleosome mapping. Our analysis combining several large-scale *de novo* mutation and nucleosome datasets in human have provided several important insights into this topic.

One main finding in our work is that strong nucleosome translational positioning is associated with elevated germline *de novo* mutation rates, which is also supported by the analysis with extremely rare variants in polymorphism data. The direct use of *de novo* mutations in the analysis avoided many confounding evolutionary factors in some previous studies. Importantly, our statistical tests controlling for nucleosome occupancy and other related factors confirmed the significant contribution of nucleosome positioning stability to the mutation rate variation. While nucleosome organization was reported to be associated with mutation rate variation in human,

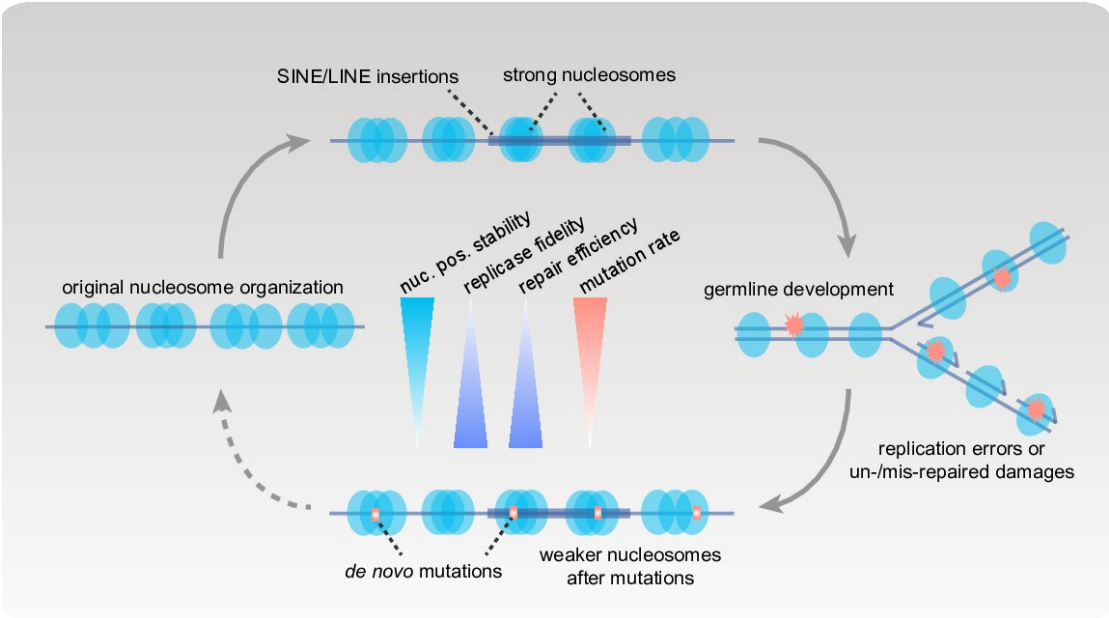
486 previous research usually focused on nucleosome occupancy and positioning
487 stability was rarely considered. Therefore we discovered a novel factor that
488 significantly modulate the germline mutation rate variation, which we think is an
489 important aspect for future related research.

490 Investigating mutational processes in the germline and somatic tissues remains
491 challenging. Nevertheless, by analyzing published genomic datasets, we obtained
492 several informative results regarding potential mutational processes associated with
493 strong nucleosomes. Analysis based on the COSMIC mutational signatures
494 suggested that MMR and DSB are likely involved in the elevated mutation rates
495 around strong nucleosomes. Our further analysis with multiple independent datasets
496 indeed supported MMR and DSB as contributing mutational mechanisms. In addition,
497 analysis with the bMMRD cancers suggests that lower replicase fidelity of both Pol ε
498 and Pol δ around strong nucleosomes also contributes to the elevated mutation rates.
499 The underlying molecular mechanisms about how strong positioning affects replicase
500 fidelity and associated repair (MMR or DSB repair) are still not clear. Based on the
501 evidence from our analysis and previous studies (Li et al. 2009; Reijns et al. 2015;
502 Tubbs et al. 2018), we speculate that strong nucleosomes may act as strong barriers
503 which impair the performance of the replication and repair machineries in those
504 regions. Strong nucleosome positioning could also affect other unexamined DNA
505 damage/repair processes related to germline development, but many published
506 genomic datasets about DNA damage/repair were generated with very short
507 sequencing reads (e.g. <100bp), which hinder detailed analysis in highly repetitive
508 regions. Improved sequencing strategies such as long-read sequencing should be
509 beneficial for future related studies.

510 Interestingly, we found that strong nucleosomes are preferentially located in young
511 LINE and SINE elements, two common types of retrotransposons in the human and
512 other mammalian species. Due to their potentially deleterious effects, newly inserted
513 retrotransposons are highly repressed by multiple regulatory mechanisms, such as
514 DNA methylation and H3K9me3 (Slotkin and Martienssen 2007). Strong nucleosome
515 positioning could be another layer of the repressive system. Strong positioning in
516 young SINEs/LINEs may mask access to the transcription machinery. Furthermore,
517 the contribution of strong nucleosomes to hypermutation in young SINEs/LINEs can
518 reduce the retotransposition capacity more rapidly. Therefore, strong positioning and
519 hypermutation in SINEs/LINEs might have facilitated the their spread across the
520 genome and help produce large amounts of genetic variations (substrates for natural
521 selection) for the host.

522 The decreasing numbers of strong nucleosomes in older LINE/SINE elements imply
523 widespread nucleosome re-positioning during evolution. Since nucleosome
524 positioning is affected by the surrounding sequences, the re-positioning probably
525 results from accumulated mutations. Our data disagree with the previous hypothesis
526 of widespread selection for maintaining nucleosome positioning in the human
527 genome (Prendergast and Semple 2011). Another reason for favoring the re-
528 positioning model is that most genomic regions do not employ strong positioning,
529 possibly due to its relatively high mutagenic potential.

530 Finally, we summarized our major findings in a proposed model in **Fig. 5**, which
531 demonstrates the relationship among nucleosome positioning, mutation rate variation,
532 retrotransposons and evolution. Given the importance of germline *de novo* mutations
533 in evolution and human diseases and the universal roles of nucleosomes in
534 eukaryotic genome organization and regulation, our work should have profound
535 implications in related research areas.



536

537 **Fig. 5 Proposed model for how nucleosome positioning stability modulates**
538 **mutation rate variation and influences evolution.** While the genome is generally
539 covered by weak nucleosomes, newly inserted SINE/LINE retrotransposons are
540 enriched with strong nucleosomes. The regions bound by strong nucleosomes are
541 subject to high mutation rates during germline development, which in turn reduce the
542 nucleosome binding affinity of the sequence and the transposition capacity of associated
543 retrotransposons.

544

545

546 **Methods**

547 **Mutation datasets**

548 *De novo* mutations identified in multiple large-scale trio sequencing project were
549 downloaded from denovo-db v1.6.1 (Turner et al. 2017b). Seven studies with >1000
550 *de novo* mutations (Genome of the Netherlands 2014; Turner et al. 2016; Yuen et al.
551 2016; Jonsson et al. 2017; Turner et al. 2017a; Yuen et al. 2017; Werling et al. 2018)
552 were considered in our analysis (**Supplementary Fig. 1a**). Extremely rare variants
553 (derived allele frequency ≤ 0.0001) were obtained from Genome Aggregation
554 Database (gnomAD, release 2.0.2) (Lek et al. 2016).

555 **Nucleosome datasets**

556 We used 1,037,801 strong nucleosomes (i.e. translationally stable nucleosomes)
557 identified based on MNase-seq data of sequenced seven lymphoblastoid cell lines
558 from Gaffney et al. (Gaffney et al. 2012). The original hg18-based coordinates of
559 annotated nucleosomes were converted to hg19 using the 'liftOver' tool from UCSC
560 genome browser. The rotationally stable nucleosomes identified based on 49 DNase-
561 seq samples (43 distinct cell types) were from Winter et al. (Winter et al. 2013). We
562 classified the human genome into three groups based on the nucleosome contexts
563 (**Fig. 1b**): 1) regions covered by translationally stable ('strong') nucleosomes; 2)
564 regions covered by rotationally but not stable translationally nucleosomes; and 3) the
565 remaining genome regions. Chromosomes X and Y were excluded from analysis as
566 some other datasets used in our work lacked data for these chromosomes. As the
567 nucleosome maps we used were not derived from germ cells, for downstream
568 analysis we excluded the genomic regions in which nucleosome positioning were
569 found to differ between human embryonic stem cells and differentiated fibroblasts
570 (West et al. 2014). Based on the positioning stability scores defined in Gaffney et al.,
571 we divided the one million strong nucleosomes into three categories of equal sizes
572 with different levels of stability – 'high', 'middle' and 'low', which were used for
573 analysis shown in **Fig. 1** and **Supplementary Fig. 1.**

574 **Accounting for mappability**

575 Sequencing read mappability can significantly affect variant calling results and other
576 aligned read-depth based measurements (e.g. nucleosome occupancy). The
577 sequencing reads for detecting *de novo* mutations used in our analysis were mainly
578 150bp paired-end reads, with fragment sizes ranging from 300-700bp
579 (**Supplementary Fig. 1**). We used the Genome Mappability Analyzer (GMA) (Lee

580 and Schatz 2012) to generate the mappability scores for simulated PE150 reads with
581 fragment sizes set to be 400bp. Only the regions with GMA mappability scores
582 of >=90 (~2.6Gb) were considered for most analyses, unless specified otherwise. We
583 did not use the mappability tracks from ENCODE for the de novo mutation data,
584 because those tracks were only for single-ended reads. For some analyses,
585 additional filtering were applied if other associated datasets suffered from more
586 severe mappability issues. For measuring nucleosome occupancy, we used the
587 method described in the Gaffney et al. to simulate paired-end 25bp reads matching
588 the base compositions of MNase-seq data in the human genome, and then
589 calculated per-base coverage depth by the simulated fragments. The 10bp-bin ratios
590 between the MNase-seq read coverage and the simulated read coverage were used
591 for measuring the occupancy.

592 **Enrichment analysis for *de novo* mutations in different nucleosome contexts**

593 The tool Genomic association tester (GAT) (Heger et al. 2013) was used to estimate
594 the expected numbers of mutations in different contexts (sampling >=1000 times),
595 which were then compared to the observed numbers. Low-mappability regions were
596 excluded from analysis. A similar analysis was also done for the extremely rare
597 variants of gnomAD. Analysis of meta-profiles along strong nucleosomes was done
598 using deepTools (Ramirez et al. 2014).

599 **Statistical modelling of the contribution of different factors to mutation rate**
600 **variation**

601 As described in the main text, for a given site, we defined two variables regarding the
602 translational positioning of nearby nucleosomes (**Fig. 2a**):

$$d_{mean} = \frac{\sum_{i=1}^{n} d_i}{n}, 0 \leq d \leq 95,$$
$$d_{var} = \frac{\sum_{i=1}^{n} (d_i - d_{mean})^2}{n}$$

603 where *d* is the distance between a MNase-seq midpoint to the focal site, $0 \leq d \leq 95$.
604 We considered MNase-seq midpoints within ±95bp of the focal site, because
605 genome-wide nucleosome repeat length was estimated to be 191.4bp (Gaffney et al.
606 2012). Genomic sites without any MNase-seq midpoint within ±95bp were excluded
607 from analysis. The measurements for nucleosome occupancy were 10bp-bin ratios
608 between the MNase-seq read coverage and the simulated read coverage. We did not
609 use the positioning score *S(i)* defined in Gaffney et al. to measure positioning stability

610 in our modelling analysis, because *S(i)* was designed for identifying the stable dyads
611 and so for non-dyad positions it does not represent the positioning stability properly.

612 RNA expression, DNA methylation and chromatin accessibility (ATAC-seq) data from
613 human spermatogonial stem cells was from Guo et al. (Guo et al. 2017). For the
614 RNA-seq and ATAC-seq data from Guo et al., because the genome-wide read signal
615 tracks were not available, we downloaded, processed and mapped the raw reads to
616 generate the genome-wide tracks. Since suitable data for histone modifications in
617 human germ cells was not available, we used the ChIP-seq data of human
618 embryonic stem cells from ENCODE (ENCODE Consortium 2012). Replication timing
619 data (Repli-seq of GM12878) was from ENCODE. The data of recombination rates
620 was from the HapMap project (International HapMap Consortium et al. 2007).

621 A binary logistic regression framework was used to assess the contribution of
622 different factors to mutation rate variation across the genome systematically. The
623 logistic regression model is described as below:

$$\mu = \Pr(y = 1) = \frac{\exp\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right)}{1 + \exp\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right)} = \frac{\exp(\boldsymbol{X}\beta)}{1 + \exp(\boldsymbol{X}\beta)}$$

$$\mathrm{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \boldsymbol{X}\beta$$

624 where $\mu = \Pr(y = 1)$ denotes the probability that a genomic site is mutated (for
625 testing individual SNV mutation types (e.g. A>T), μ is the probability that a site is
626 mutated to a specific nucleotide), $\boldsymbol{X}$ represents the observations for the considered
627 variables (categorical or continuous, e.g. d_{mean}, d_{var}, adjacent nucleotides, etc.), and β
628 is the vector of parameters to be estimated.

629 We used the Bayesian logistic regression model implemented in the 'bayesglm'
630 (Gelman et al. 2008) of the R package 'arm', which was reported to perform well in
631 handling the complete separation issue in logistic regression models (Gelman et al.
632 2008). The complete separation issue is common when one class is rare relative to
633 the other and (or) there are many regressors in a model. As we had only ~300,000
634 *de novo* mutations, the probability for a given site to be mutated in our data is
635 ~1/10,000, which is a rare event.

636 Within the logistic regression framework, we compared the full model with all
637 considered variables to a reduced model without one specific variable by performing
638 likelihood-ratio tests in R ('anova' function) to evaluate the significance for each
639 variable. The resulting p values of a set of likelihood-ratio tests were adjusted for
640 multiple testing with Bonferroni-Holm correction.

641 To perform the regression analysis, we generated the data of all variables for the *de*
642 *novo* mutation sites and subsampled a fraction of the non-mutated sites as the
643 control sites. We did not use all the non-mutated sites in the genome as it would lead
644 to a large imbalance in the sizes of two classes ('mutated' and 'non-mutated') and
645 much larger computational burden. For *de novo* SNVs, we randomly generated
646 2,561,953 non-mutated sites (about 1/1000 of the accessible genome, about 10
647 times as many as *de novo* SNVs) and 256,337 non-mutated sites (about 1/10,000 of
648 the accessible genome, about 10 times as many as *de novo* INDELs) for INDELs.
649 For *de novo* INDELs, we used the INDELs of ≤5bp for regression analysis, because
650 long INDELs were rare and may have a high false-positive rate). For RNA expression,
651 DNA methylation, chromatin accessibility, replication timing, recombination rate and
652 histone modifications data, we used the average value of the ±10bp of a focal site for
653 the specific feature based on the genome-wide signal tracks. We also assessed
654 different window sizes (±5bp and ±20bp), which led to similar results.

655 For SNVs, we performed logistic regression tests for mutation types at A/T sites and
656 C/G sites separately and distinguished C/G sites in CpG and non-CpG contexts. We
657 also tested for nine individual SNV mutation types (three for A/T sites, three for C/G
658 sites at CpG contexts, and three for non-CpG contexts, **Supplementary Fig. 3**). The
659 regression coefficients for the full model of each test are given in **Supplementary**
660 **Table 1.**

661 Since the variable d_{mean} has a non-monotonic relationship with mutation rates, we
662 binned the values into five categories: [0,18], [19, 36], [37, 54], [55, 73] and [74, 95]
663 (first four bins implying nucleosome-bound regions, and the last bin implying close to
664 the linker). We also did tests when considering d_{mean} as a continuous variable
665 (**Supplementary Fig. 4)**.

666 In the regression models mentioned above, we did not consider the non-additive
667 effects of adjacent nucleotides (±5 bp). When we tried adding non-additive effects for
668 ±5 nucleotides (considering only two-way interactions; taking a much longer running
669 time), we got similar results regarding the association of positioning stability (d_{var}) and
670 mutation rates (**Supplementary Fig. 5)**. We also tried using the 7-mer mutability

671 estimates from Carlson et al. (Carlson et al. 2018), which incorporated non-additive
672 effects among ±3 nucleotides, as predictors in the regression models and which gave
673 a similar significance for d_{var} (**Supplementary Fig. 5)**.

674 **Analysis of mutational processes**

675 The contribution of COSMIC mutational signatures (Alexandrov et al. 2013) to
676 different sets of mutations (*de novo* SNVs and somatic mutations from bMMRD
677 samples) was predicted using the 'fit_to_signatures' function in the R package
678 'MutationalPatterns' (Blokzijl et al. 2018).

679 *POLE* mutated genomes from PCAWG project (Campbell et al. 2017) were used to
680 evaluate the differential MMR efficiency between strong and non-strong nucleosome
681 regions, We compared the mutation densities in cancer genomes with *POLE* mutated
682 and a deficient MMR to those with *POLE* mutated and a proficient MMR. The MMR
683 pathway was considered deficient if a driver mutation (annotated by the PCAWG
684 consortium) was found in one of five MMR core genes - *MLH1*, *MSH2*, *MSH6*, *PMS1*
685 and *PMS2*.

686 For a given bin (10bp-size) in the meta-profile, we calculated the relative MMR
687 escape ratio relative to genomic background around strong nucleosomes as
688 described in the following formula,

$$R_i^{escape} = \frac{\frac{m_i^{POLE^*,\, MMR^{WT}}}{m_i^{POLE^*,\, MMR^*}}}{\frac{\overline{m}^{POLE^*,\, MMR^{WT}}}{\overline{m}^{POLE^*,\, MMR^*}}}$$

689 where m_i is the mutation density for the *i*th bin, and $\overline{m}$ is the genome-wide average
690 mutation density of a specific sample group, estimated by simulating random
691 windows in the genome. A similar logic was used when evaluating relative escape
692 ratios of Pol ε (mutated POLE) and Pol δ (mutated POLD1) using the somatic
693 mutations generated by the bMMRD project (Shlien et al. 2015).

694 When analyzing PCAWG and bMMRD data, to account for potential mappability
695 issues, we focused on the highly mappable regions based on the CrgMapability
696 scores from ENCODE. We used CrgMapability scores here, which are more stringent
697 than GMA ones, because detecting somatic mutations in tumors is more difficult than
698 for ordinary individual re-sequencing data. We considered the strong nucleosomes
699 which have a 100mer CrgMapability score of 1 (meaning any 100-bp read from these
700 regions can be mapped uniquely in the genome) within ±800bp of the dyads. We

701 then simulated a same number of 1600bp-sized regions from the genome that satisfy
702 the mappability requirement to calculate the background mutation density. Note that
703 in theory the mappabiity issue in the relative escape ratios should be very small
704 because the two sets of samples have the same mappability for a given bin and the
705 ratio calculation normalizes the effects of different mappability among regions.

706 The raw reads of OK-seq data (Petryk et al. 2016) were downloaded from NCBI and
707 mapped to the human genome. We kept only the uniquely mapped reads for inferring
708 Okazaki junctions. The very 5' end sites of aligned reads (separating reads mapped
709 to Watson and Crick strands) were considered putative Okazaki junction signals.

710 To investigate DSBs around strong nucleosmes, we downloaded the genome-wide
711 tracks of human END-seq data (GSM3227951 and GSM3227952) (Tubbs et al.
712 2018). Because the reads of END-seq data were single-ended 75bp, we considered
713 the strong nucleosomes which have a 75mer CrgMapability score of 1 within ±500bp
714 of the strong nucleosome dyads for analysis.

715 **Enrichment analysis for strong nucleosomes in different repeat contexts**

716 GAT (Heger et al. 2013) was used to estimate the expected numbers of strong
717 nucleosomes in different contexts (sampling >=1000 times), which were compared to
718 the observed numbers. The annotations of repeat elements (Feb 2009, Repeat
719 Library 20140131) were downloaded from RepeatMasker (Tempel 2012). We also
720 did GAT analysis for LINE-1(L1) and Alu subfamilies of different ages. The age
721 information of repeat families was from Giordano et al. (Giordano et al. 2007). For
722 generating the MNase-seq midpoints along the repeat consensus sequences, we
723 made use of the alignment information in the RepeatMasker result files
724 ('hg19.fa.align.gz') and mapped the hg19-based coordinates to the coordinates in the
725 consensus sequences. Nucleosome deformation energy scores of all sites in the
726 human genome were estimated using nuScore (Tolstorukov et al. 2008). For the L1
727 analysis shown in **Supplementary Fig. 9**, we only considered the 3' end regions of
728 L1 subfamilies, because 5' end regions of L1 elements are usually truncated in the
729 genome and their subfamily identities are difficult to be determined.

730

731

732 **References**

733 Acuna-Hidalgo R, Veltman JA, Hoischen A. 2016. New insights into the generation
734 and role of de novo mutations in health and disease. *Genome biology* **17**(1):
735 241.
736 Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell
737 GR, Bolli N, Borg A, Borresen-Dale AL et al. 2013. Signatures of mutational
738 processes in human cancer. *Nature* **500**(7463): 415-421.
739 Blokzijl F, Janssen R, van Boxtel R, Cuppen E. 2018. MutationalPatterns:
740 comprehensive genome-wide analysis of mutational processes. *Genome*
741 *medicine* **10**(1): 33.
742 Campbell PJ, Getz G, Stuart JM, Korbel JO, Stein LD. 2017. Pan-cancer analysis of
743 whole genomes. *BioRxiv*: 162784.
744 Carlson J, Locke AE, Flickinger M, Zawistowski M, Levy S, Myers RM, Boehnke M,
745 Kang HM, Scott LJ, Li JZ et al. 2018. Extremely rare variants reveal patterns
746 of germline mutation rate heterogeneity in humans. *Nature communications*
747 **9**(1): 3753.
748 Chen X, Chen Z, Chen H, Su Z, Yang J, Lin F, Shi S, He X. 2012. Nucleosomes
749 suppress spontaneous mutations base-specifically in eukaryotes. *Science*
750 **335**(6073): 1235-1238.
751 Drillon G, Audit B, Argoul F, Arneodo A. 2016. Evidence of selection for an accessible
752 nucleosomal array in human. *BMC genomics* **17**: 526.
753 ENCODE Consortium. 2012. An integrated encyclopedia of DNA elements in the
754 human genome. *Nature* **489**(7414): 57-74.
755 Englander EW, Howard BH. 1995. Nucleosome positioning by human Alu elements in
756 chromatin. *The Journal of biological chemistry* **270**(17): 10091-10096.
757 Englander EW, Wolffe AP, Howard BH. 1993. Nucleosome interactions with a human
758 Alu element. Transcriptional repression and effects of template methylation.
759 *The Journal of biological chemistry* **268**(26): 19565-19573.
760 Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, Genome of the
761 Netherlands C, van Duijn CM, Swertz M, Wijmenga C et al. 2015. Genome-
762 wide patterns and properties of de novo mutations in humans. *Nature*
763 *genetics* **47**(7): 822-826.
764 Frigola J, Sabarinathan R, Mularoni L, Muinos F, Gonzalez-Perez A, Lopez-Bigas N.
765 2017. Reduced mutation rate in exons due to differential mismatch repair.
766 *Nature genetics* **49**(12): 1684-1692.
767 Gaffney DJ, McVicker G, Pai AA, Fondufe-Mittendorf YN, Lewellen N, Michelini K,
768 Widom J, Gilad Y, Pritchard JK. 2012. Controls of nucleosome positioning in
769 the human genome. *PLoS genetics* **8**(11): e1003036.
770 Gelman A, Jakulin A, Pittau MG, Su Y-S. 2008. A weakly informative default prior
771 distribution for logistic and other regression models. *The Annals of Applied*
772 *Statistics* **2**(4): 1360-1383.
773 Genome of the Netherlands C. 2014. Whole-genome sequence variation, population
774 structure and demographic history of the Dutch population. *Nature genetics*
775 **46**(8): 818-825.
776 Giordano J, Ge Y, Gelfand Y, Abrusan G, Benson G, Warburton PE. 2007.
777 Evolutionary history of mammalian transposons determined by genome-wide
778 defragmentation. *PLoS computational biology* **3**(7): e137.
779 Guo J, Grow EJ, Yi C, Mlcochova H, Maher GJ, Lindskog C, Murphy PJ, Wike CL,
780 Carrell DT, Goriely A et al. 2017. Chromatin and Single-Cell RNA-Seq Profiling
781 Reveal Dynamic Signaling and Metabolic Transitions during Human
782 Spermatogonial Stem Cell Development. *Cell stem cell* **21**(4): 533-546 e536.

783 Harpak A, Bhaskar A, Pritchard JK. 2016. Mutation Rate Variation is a Primary
784 Determinant of the Distribution of Allele Frequencies in Humans. *PLoS*
785 *genetics* **12**(12): e1006489.
786 Harris K, Nielsen R. 2014. Error-prone polymerase activity causes multinucleotide
787 mutations in humans. *Genome research* **24**(9): 1445-1454.
788 Heger A, Webber C, Goodson M, Ponting CP, Lunter G. 2013. GAT: a simulation
789 framework for testing the association of genomic intervals. *Bioinformatics*
790 **29**(16): 2046-2048.
791 Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across
792 mammalian genomes. *Nature reviews Genetics* **12**(11): 756-766.
793 International HapMap Consortium Frazer KA Ballinger DG Cox DR Hinds DA Stuve LL
794 Gibbs RA Belmont JW Boudreau A Hardenbol P et al. 2007. A second
795 generation human haplotype map of over 3.1 million SNPs. *Nature*
796 **449**(7164): 851-861.
797 Jonsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson MT,
798 Hjorleifsson KE, Eggertsson HP, Gudjonsson SA et al. 2017. Parental influence
799 on human germline de novo mutations in 1,548 trios from Iceland. *Nature*
800 **549**(7673): 519-522.
801 Lee H, Schatz MC. 2012. Genomic dark matter: the reliability of short read mapping
802 illustrated by the genome mappability score. *Bioinformatics* **28**(16): 2097-
803 2105.
804 Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria
805 AH, Ware JS, Hill AJ, Cummings BB et al. 2016. Analysis of protein-coding
806 genetic variation in 60,706 humans. *Nature* **536**(7616): 285-291.
807 Li F, Tian L, Gu L, Li GM. 2009. Evidence that nucleosomes inhibit mismatch repair in
808 eukaryotic cells. *The Journal of biological chemistry* **284**(48): 33056-33061.
809 Lujan SA, Clausen AR, Clark AB, MacAlpine HK, MacAlpine DM, Malc EP, Mieczkowski
810 PA, Burkholder AB, Fargo DC, Gordenin DA et al. 2014. Heterogeneous
811 polymerase fidelity and mismatch repair bias genome variation and
812 composition. *Genome research* **24**(11): 1751-1764.
813 Makova KD, Hardison RC. 2015. The effects of chromatin organization on variation in
814 mutation rates in the genome. *Nature reviews Genetics* **16**(4): 213-223.
815 Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D,
816 Bhandari A et al. 2012. Whole-genome sequencing in autism identifies hot
817 spots for de novo germline mutation. *Cell* **151**(7): 1431-1442.
818 Perera D, Poulos RC, Shah A, Beck D, Pimanda JE, Wong JW. 2016. Differential DNA
819 repair underlies mutation hotspots at active promoters in cancer genomes.
820 *Nature* **532**(7598): 259-263.
821 Petryk N, Kahli M, d'Aubenton-Carafa Y, Jaszczyszyn Y, Shen Y, Silvain M, Thermes C,
822 Chen CL, Hyrien O. 2016. Replication landscape of the human genome.
823 *Nature communications* **7**: 10208.
824 Pich O, Muinos F, Sabarinathan R, Reyes-Salazar I, Gonzalez-Perez A, Lopez-Bigas N.
825 2018. Somatic and Germline Mutation Periodicity Follow the Orientation of the
826 DNA Minor Groove around Nucleosomes. *Cell* **175**(4): 1074-1087 e1018.
827 Prendergast JG, Semple CA. 2011. Widespread signatures of recent selection linked
828 to nucleosome positioning in the human lineage. *Genome research* **21**(11):
829 1777-1787.
830 Ramirez F, Dundar F, Diehl S, Gruning BA, Manke T. 2014. deepTools: a flexible
831 platform for exploring deep-sequencing data. *Nucleic acids research* **42**(Web
832 Server issue): W187-191.

833 Reijns MAM, Kemp H, Ding J, de Proce SM, Jackson AP, Taylor MS. 2015. Lagging-
834 strand replication shapes the mutational landscape of the genome. *Nature*
835 **518**(7540): 502-506.
836 Rodgers K, McVey M. 2016. Error-Prone Repair of DNA Double-Strand Breaks.
837 *Journal of cellular physiology* **231**(1): 15-24.
838 Sabarinathan R, Mularoni L, Deu-Pons J, Gonzalez-Perez A, Lopez-Bigas N. 2016.
839 Nucleotide excision repair is impaired by binding of transcription factors to
840 DNA. *Nature* **532**(7598): 264-267.
841 Salih F, Salih B, Kogan S, Trifonov EN. 2008. Epigenetic nucleosomes: Alu sequences
842 and CG as nucleosome positioning element. *Journal of biomolecular structure*
843 *& dynamics* **26**(1): 9-16.
844 Sasaki S, Mello CC, Shimada A, Nakatani Y, Hashimoto S, Ogawa M, Matsushima K,
845 Gu SG, Kasahara M, Ahsan B et al. 2009. Chromatin-associated periodicity in
846 genetic variation downstream of transcriptional start sites. *Science*
847 **323**(5912): 401-404.
848 Schuster-Bockler B, Lehner B. 2012. Chromatin organization is a major influence on
849 regional mutation rates in human cancer cells. *Nature* **488**(7412): 504-507.
850 Segurel L, Wyman MJ, Przeworski M. 2014. Determinants of mutation rate variation
851 in the human germline. *Annual review of genomics and human genetics* **15**:
852 47-70.
853 Seplyarskiy VB, Akkuratov EE, Akkuratova N, Andrianova MA, Nikolaev SI, Bazykin
854 GA, Adameyko I, Sunyaev SR. 2018. Error-prone bypass of DNA lesions
855 during lagging-strand replication is a common source of germline and cancer
856 mutations. *Nature genetics*.
857 Seplyarskiy VB, Andrianova MA, Bazykin GA. 2017. APOBEC3A/B-induced
858 mutagenesis is responsible for 20% of heritable mutations in the TpCpW
859 context. *Genome research* **27**(2): 175-184.
860 Shlien A, Campbell BB, de Borja R, Alexandrov LB, Merico D, Wedge D, Van Loo P,
861 Tarpey PS, Coupland P, Behjati S et al. 2015. Combined hereditary and
862 somatic mutations of replication error repair genes result in rapid onset of
863 ultra-hypermutated cancers. *Nature genetics* **47**(3): 257-262.
864 Slotkin RK, Martienssen R. 2007. Transposable elements and the epigenetic
865 regulation of the genome. *Nature reviews Genetics* **8**(4): 272-285.
866 Smerdon MJ. 1991. DNA repair and the role of chromatin structure. *Current opinion*
867 *in cell biology* **3**(3): 422-428.
868 Smith TCA, Arndt PF, Eyre-Walker A. 2018. Large scale variation in the rate of germ-
869 line de novo mutation, base composition, divergence and diversity in humans.
870 *PLoS genetics* **14**(3): e1007254.
871 Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev
872 SR. 2009. Human mutation rate associated with DNA replication timing.
873 *Nature genetics* **41**(4): 393-395.
874 Supek F, Lehner B. 2015. Differential DNA mismatch repair underlies mutation rate
875 variation across the human genome. *Nature* **521**(7550): 81-84.
876 Tanaka Y, Yamashita R, Suzuki Y, Nakai K. 2010. Effects of Alu elements on global
877 nucleosome positioning in the human genome. *BMC genomics* **11**: 309.
878 Tempel S. 2012. Using and understanding RepeatMasker. *Methods in molecular*
879 *biology* **859**: 29-51.
880 Terekhanova NV, Seplyarskiy VB, Soldatov RA, Bazykin GA. 2017. Evolution of Local
881 Mutation Rate and Its Determinants. *Molecular biology and evolution* **34**(5):
882 1100-1109.

883 Tolstorukov MY, Choudhary V, Olson WK, Zhurkin VB, Park PJ. 2008. nuScore: a
884 web-interface for nucleosome positioning predictions. *Bioinformatics* **24**(12):
885 1456-1458.
886 Tolstorukov MY, Volfovsky N, Stephens RM, Park PJ. 2011. Impact of chromatin
887 structure on sequence variability in the human genome. *Nature structural &*
888 *molecular biology* **18**(4): 510-515.
889 Tubbs A, Sridharan S, van Wietmarschen N, Maman Y, Callen E, Stanlie A, Wu W,
890 Wu X, Day A, Wong N et al. 2018. Dual Roles of Poly(dA:dT) Tracts in
891 Replication Initiation and Fork Collapse. *Cell* **174**(5): 1127-1142 e1119.
892 Turner TN, Coe BP, Dickel DE, Hoekzema K, Nelson BJ, Zody MC, Kronenberg ZN,
893 Hormozdiari F, Raja A, Pennacchio LA et al. 2017a. Genomic Patterns of De
894 Novo Mutation in Simplex Autism. *Cell* **171**(3): 710-722 e712.
895 Turner TN, Hormozdiari F, Duyzend MH, McClymont SA, Hook PW, Iossifov I, Raja A,
896 Baker C, Hoekzema K, Stessman HA et al. 2016. Genome Sequencing of
897 Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory
898 DNA. *American journal of human genetics* **98**(1): 58-74.
899 Turner TN, Yi Q, Krumm N, Huddleston J, Hoekzema K, HA FS, Doebley AL, Bernier
900 RA, Nickerson DA, Eichler EE. 2017b. denovo-db: a compendium of human de
901 novo variants. *Nucleic acids research* **45**(D1): D804-D811.
902 Veltman JA, Brunner HG. 2012. De novo mutations in human genetic disease. *Nature*
903 *reviews Genetics* **13**(8): 565-575.
904 Warnecke T, Becker EA, Facciotti MT, Nislow C, Lehner B. 2013. Conserved
905 substitution patterns around nucleosome footprints in eukaryotes and
906 Archaea derive from frequent nucleosome repositioning through evolution.
907 *PLoS computational biology* **9**(11): e1003373.
908 Werling DM, Brand H, An JY, Stone MR, Zhu L, Glessner JT, Collins RL, Dong S, Layer
909 RM, Markenscoff-Papadimitriou E et al. 2018. An analytical framework for
910 whole-genome sequence association studies and its implications for autism
911 spectrum disorder. *Nature genetics* **50**(5): 727-736.
912 West JA, Cook A, Alver BH, Stadtfeld M, Deaton AM, Hochedlinger K, Park PJ,
913 Tolstorukov MY, Kingston RE. 2014. Nucleosomal occupancy changes locally
914 over key regulatory regions during cell differentiation and reprogramming.
915 *Nature communications* **5**: 4719.
916 Winter DR, Song L, Mukherjee S, Furey TS, Crawford GE. 2013. DNase-seq predicts
917 regions of rotational nucleosome stability across diverse human cell types.
918 *Genome research* **23**(7): 1118-1129.
919 Yuen R, Merico D, Bookman M, J LH, Thiruvahindrapuram B, Patel RV, Whitney J,
920 Deflaux N, Bingham J, Wang Z et al. 2017. Whole genome sequencing
921 resource identifies 18 new candidate genes for autism spectrum disorder.
922 *Nature neuroscience* **20**(4): 602-611.
923 Yuen RK, Merico D, Cao H, Pellecchia G, Alipanahi B, Thiruvahindrapuram B, Tong X,
924 Sun Y, Cao D, Zhang T et al. 2016. Genome-wide characteristics of de novo
925 mutations in autism. *NPJ genomic medicine* **1**: 160271-1602710.

926

927

928 **Data availability**

929 All the analyses in this study were based on published datasets.

930 **Acknowledgments**

931 We are grateful to Tobias Warnecke, John Diffley, Anob Chakrabarti and Sara
932 Rohban for insightful comments. We thank Peter Van Loo, Jonas
933 Demeulemeester and Maxime Tarabichi for assistance in accessing the
934 PCAWG genomic data. We appreciate obtaining access to the *de novo*
935 mutation data on SFARI Base. This work is supported by the Francis Crick
936 Institute which receives its core funding from Cancer Research UK
937 (FC001110), the UK Medical Research Council (FC001110), and the
938 Wellcome Trust (FC001110) (N.M.L.). N.M.L. is also supported by a Wellcome
939 Trust Investigator Award and core funding from the Okinawa Institute of
940 Science & Technology. C.L. is funded by an EMBO long-term postdoctoral
941 fellowship (ALTF 1499-2016).

942 **Author contributions**

943 C.L. conceived the project, performed the analyses and drafted the
944 manuscript; N.M.L. supervised the project and co-wrote the manuscript.

945 **Competing financial interests**

946 The authors declare no competing financial interests.

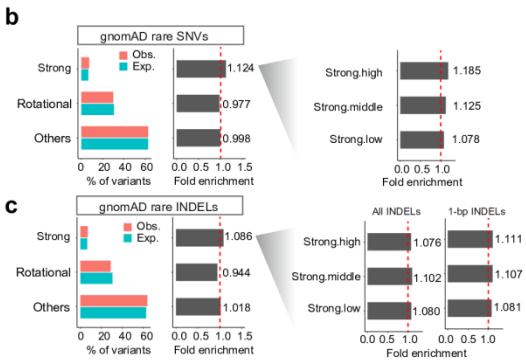
947

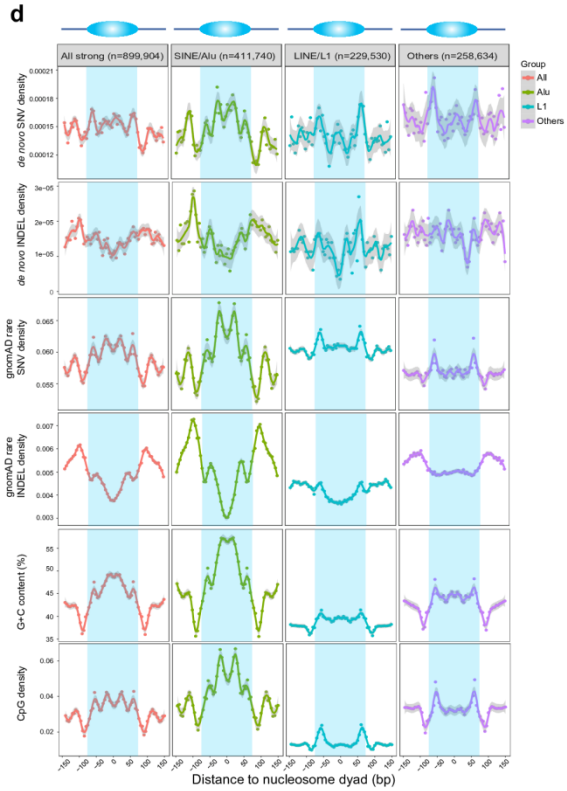
Supplementary Tables and Figures

Supplementary Table 1 Coefficients of variables and other parameters/results from the full regression models for different mutation types (in a separate Excel file). Note that for each of the categorial variables, the first category was used by the regression model as reference category (other categories were compared with the reference category) and thus there is no coefficient for that category. The statistics (4th column) and p-values (5th column) in the table were from Wald tests defaultly produced by 'bayesglm' (shown for reference), which are different from the LRT-based p-values and were not used in our discussion.

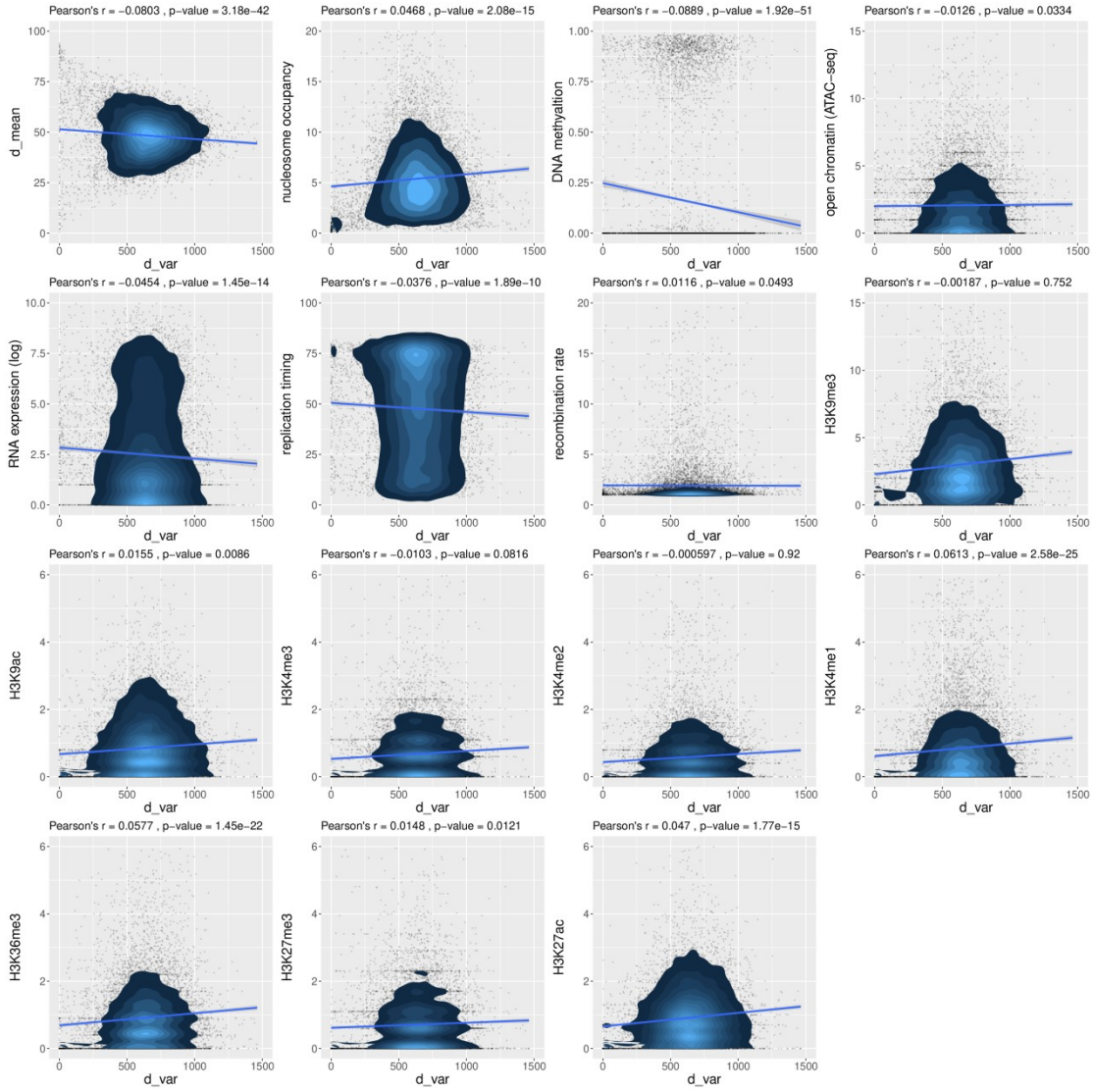
a

Study	#de novo SNVs	#de novo INDELs	Read info
Turner2017	116,101	8,402	PE150 (SFARI website); a median library insert size of 417.8 ± 111.5 bp.
Yuen2017	107,468	18,511	Mainly PE150 (fragmented the DNA into 350 bp average lengths).
Jonsson2017	99,678	7,615	Mainly PE150 (mean target size of 300–400 bp).
GONL2014	11,006	Not available	PE91, 500 bp insert size
Yuen2016	10,040	750	Pair-end DNA libraries with an insert size of 500 bp.
Turner2016	5,570	188	PE150 (SFARI website); a median library insert size of 417.8 ± 111.5 bp.
Werling2018	4,357	1,766	PE150 with a median insert size of 423 bp
Total	354,220	37,232	

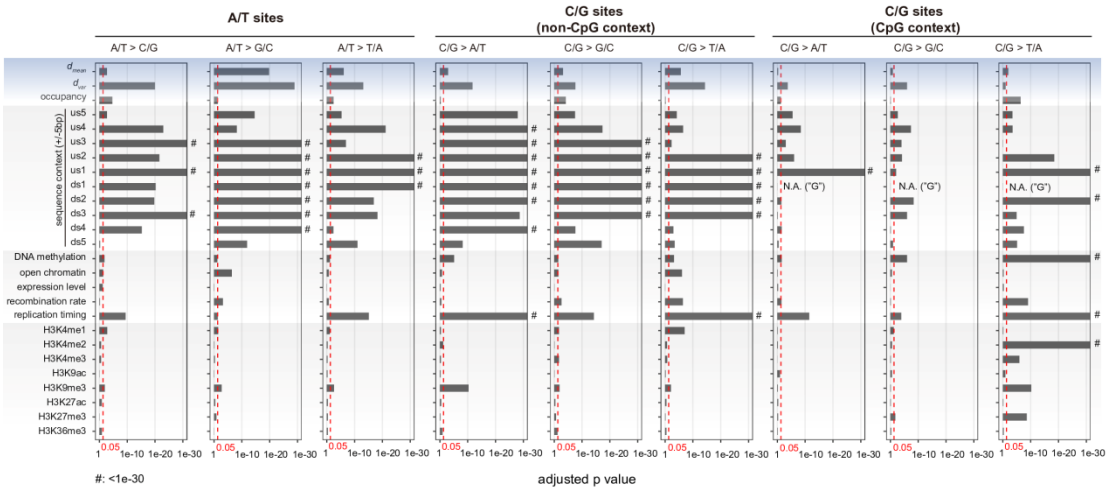




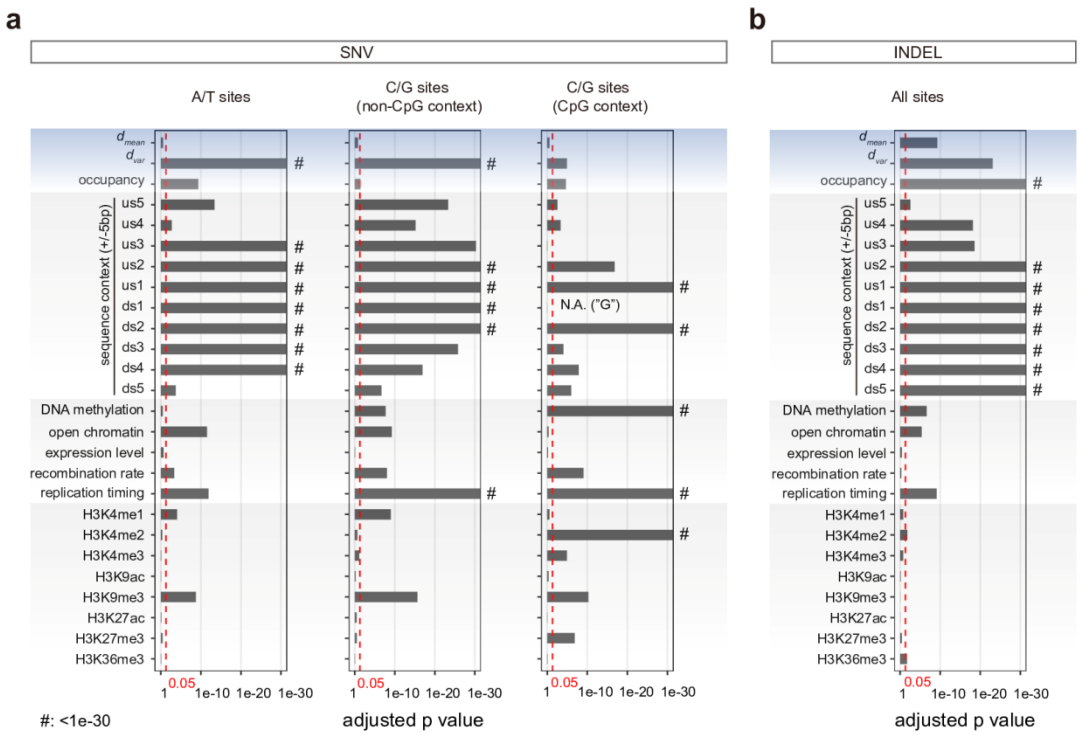
Supplementary Figure 1 Mutations in different nucleosome contexts. (**a**) Information of the *de novo* mutation datasets from seven studies used in analysis. (**b**) Fold enrichment/depletion of gnomAD extremely rare SNVs in different nucleosome contexts. 'Strong', translationally stable positioning; 'Rotational', rotationally but not translationally stable positioning; 'Others', the remaining genomic regions. On the left is the fold enrichment for three subgroups of strong nucleosomes with different stabilities. (**c**) Fold enrichment/depletion of gnomAD INDELs in different nucleosome contexts. When using all INDELs the 'strong.high' group does not have a higher mutation rate than other two groups, but if using the 1-bp INDELs 'strong.high' does have the highest mutation rate among the three groups. We speculated that there may be more false negatives of longer INDELs in the 'strong.high' group.(**d**) Meta-profiles of SNV/INDEL densities (*de novo* or extremely rare variants) around all strong nucleosomes, or in different repeat-associated groups. At the bottom are the G+C content and CpG content profiles.



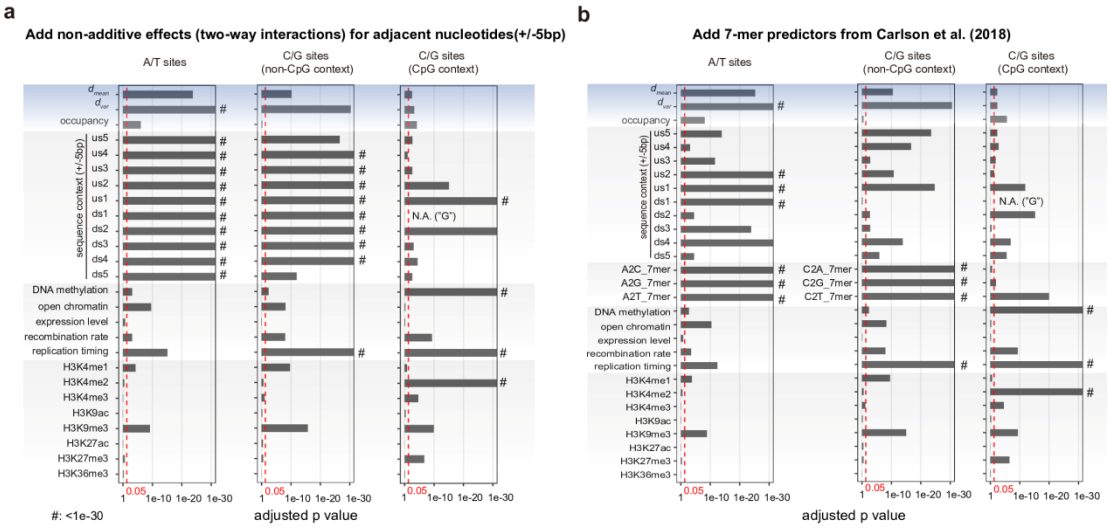
Supplementary Figure 2 Correlation analysis between nucleosome positioning stability (d_{var}) and other factors. On the top of each panel are the Pearson's correlation coefficients and the corresponding p-values.



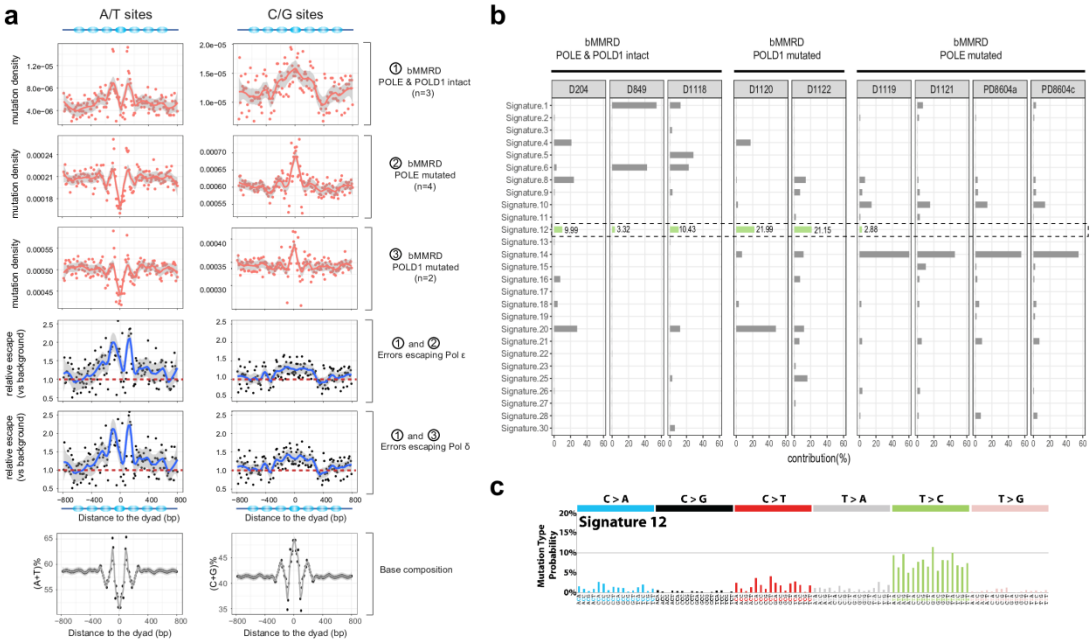
Supplementary Figure 3 Results of statistical tests for nine individual SNV mutation types. C/G sites in non-CpG contexts and C/G sites in CpG contexts were tested separately. The red vertical lines represent the significance cut-off (0.05) for the adjusted p values (Bonferroni-Holm correction). ‘us’, upstream; ‘ds’, downstream. ‘#’ means adjusted p < 1e-30.



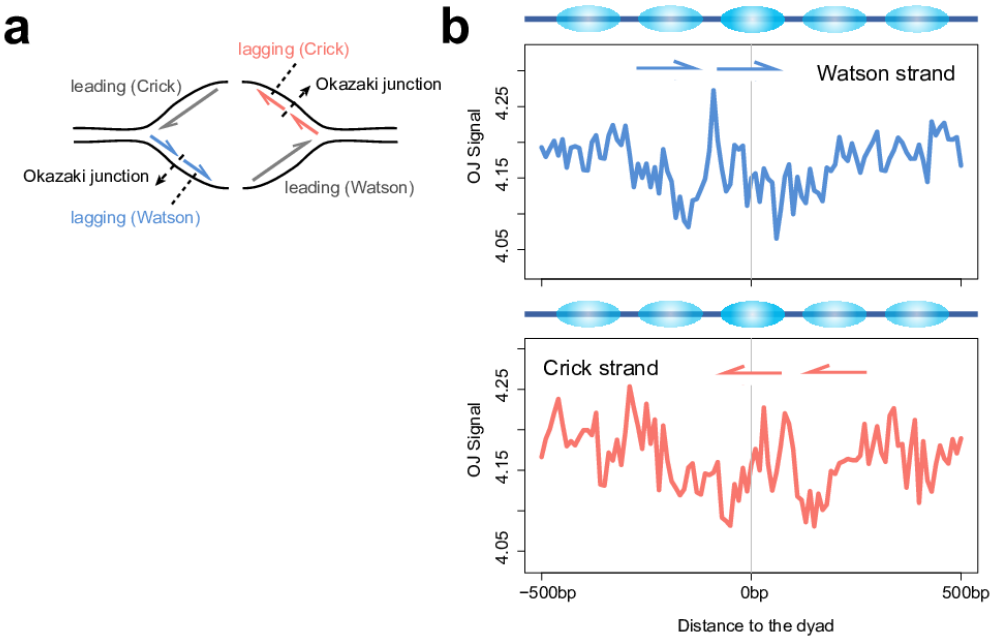
Supplementary Figure 4 Results of statistical tests when considering d_{mean} as a continuous variable. (**a-b**) Statistical significance of each variable when performing likelihood ratio tests for assessing their contribution to mutation rate variation. Panel **a** for *de novo* SNVs and **b** for *de nvo* INDELs. For *de novo* SNVs, we did tests for A/T sites, C/G sites in non-CpG context and C/G sites in CpG context separately. The red vertical lines represent the significance cut-off (0.05) for the adjusted p values (Bonferroni-Holm correction). 'us', upstream; 'ds', downstream. '#' means adjusted p < 1e-30.



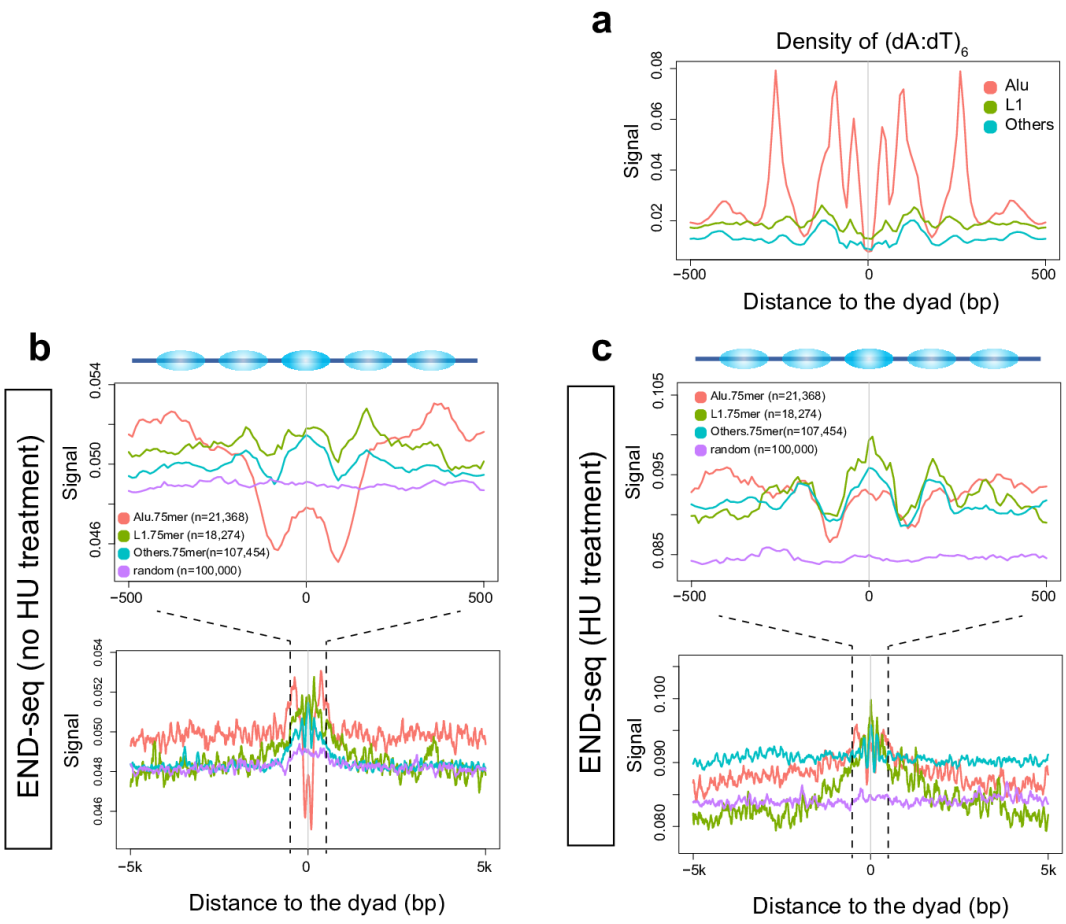
Supplementary Figure 5 **Results of statistical tests when considering non-additive effects and 7-mer mutability estimates from Carlson et al.** (**a**) Adding the two-way interactions for ±5 nucleotides in the regression models. (**b**) Adding the 7-mer mutability estimates from Carlson et al. as predictors in the regression models. We tested SNVs at A/T sites, C/G sites in non-CpG context and C/G sites in CpG context separately. The red vertical lines represent the significance cut-off (0.05) for the adjusted p values (Bonferroni-Holm correction). 'us', upstream; 'ds', downstream. '#' means adjusted p < 1e-30.



Supplementary Figure 6 Analysis of related mutational processes using bMMRD data. (**a**) Mutation profiles around strong nucleosomes for bMMRD cancer genomes and the estimated relative escape ratios of Pol ε or Pol δ, for mutations at A/T sites and C/G sites respectively. Numbers of cancer genomes used are indicated in the brackets. (**b**) Comparison of the contribution of COSMIC mutational signatures predicted by MutationalPatterns in different bMMRD genomes. Highlighted is Signature 12, which shows a particularly high contribution in POLD1-muated bMMRD samples. (**c**) the tri-nucleotide mutational profile of Signature 12, obtained from COSMIC website.

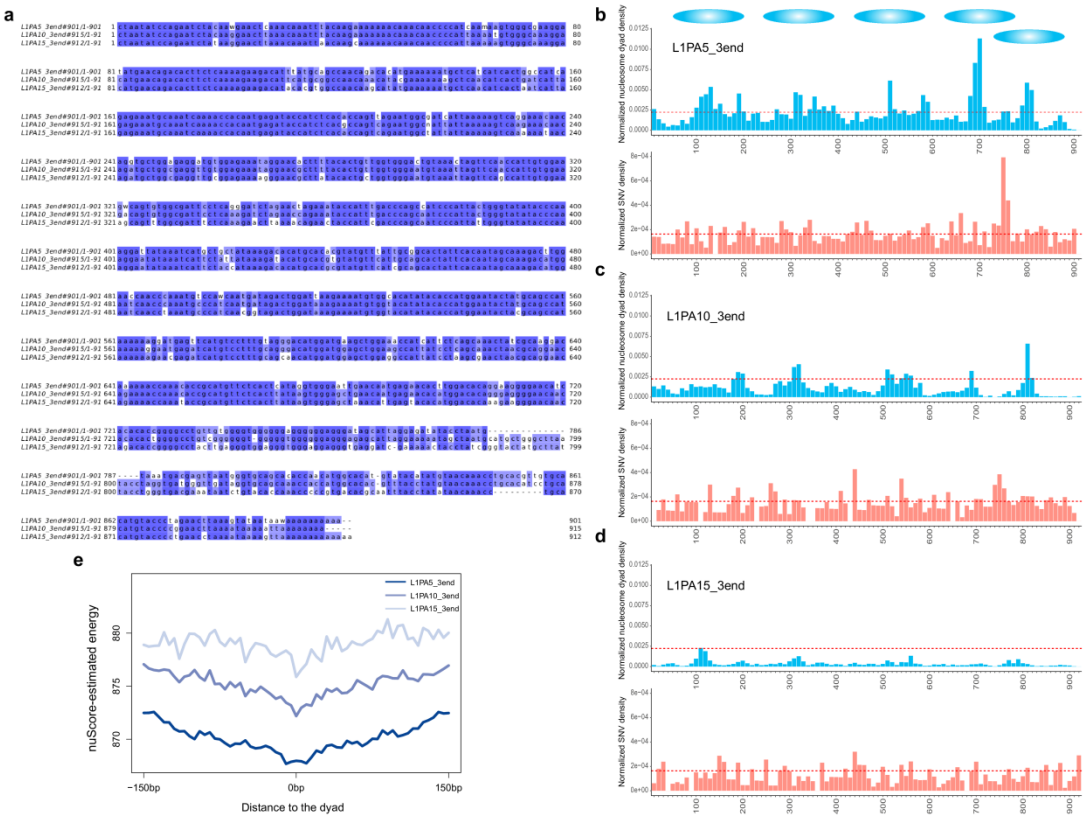


Supplementary Figure 7 Analysis with OK-seq data. (**a**) Schematic illustrating replication strands and Okazaki junctions (OJs). (**b**) Meta-profile of the density of Okazaki junctions inferred from alignments of OK-seq reads around strong nucleosomes (high-mappability). OJ signals for Watson strand and Crick strand were plotted separately. Replication directions of Okazaki fragments are shown by arrows.



1037

1038 **Supplementary Figure 8 Analysis related to the DSBs around strong**
1039 **nucleosomes.** (**a**) Density of poly(dA:dT) tracts (based on occurrence of $(dA:dT)_6$
1040 motifs) around strong nucleosomes. (**b-c**) Signal of DSBs based on the END-seq
1041 data around strong nucleosomes associated with different repeat elements. Only the
1042 strong nucleosomes of high 75-mer mappability within ±500bp were considered.
1043 Numbers of usable strong nucleosomes for each group are given in the brackets. HU
1044 (hydroxyurea) is a replicative stress-inducing agent.



Supplementary Figure 9 Densities of strong nucleosome dyads and *de novo* SNVs along the consensus sequences of 3' ends of different L1 subfamilies. (**a**) Alignment of the consensus sequences of the 3' end sequences of three L1 subfamilies (from young to old: L1PA5, L1PA10 and L1PA15), indicating the high similarity between sequences. (**b-d**) Barplots for normalized densities of strong nucleosome dyads and *de novo* SNVs along the consensus sequences of three L1 subfamilies, using 10-bp bins. Several loci that are enriched for dyads of strong nucleosomes are shown on the top with ellipses. The red dash lines represent the average densities for the L1PA5 subfamily. As shown in the **b-d**, the densities of strong nucleosome dyads and *de novo* SNVs decreased over evolutionary time. (**e**) Meta-profile for the estimated deformation energy around the strong nucleosomes of three L1 subfamilies, centered at the dyads.