Topography of mutational signatures in human cancer

Burcak Otlu[1-4], Marcos Díaz-Gay[1-3], Ian Vermes[5], Erik N Bergstrom[1-3], Maria Zhivagui[1-3], Mark Barnes[1-3], and Ludmil B. Alexandrov[1-3*]

Affiliations

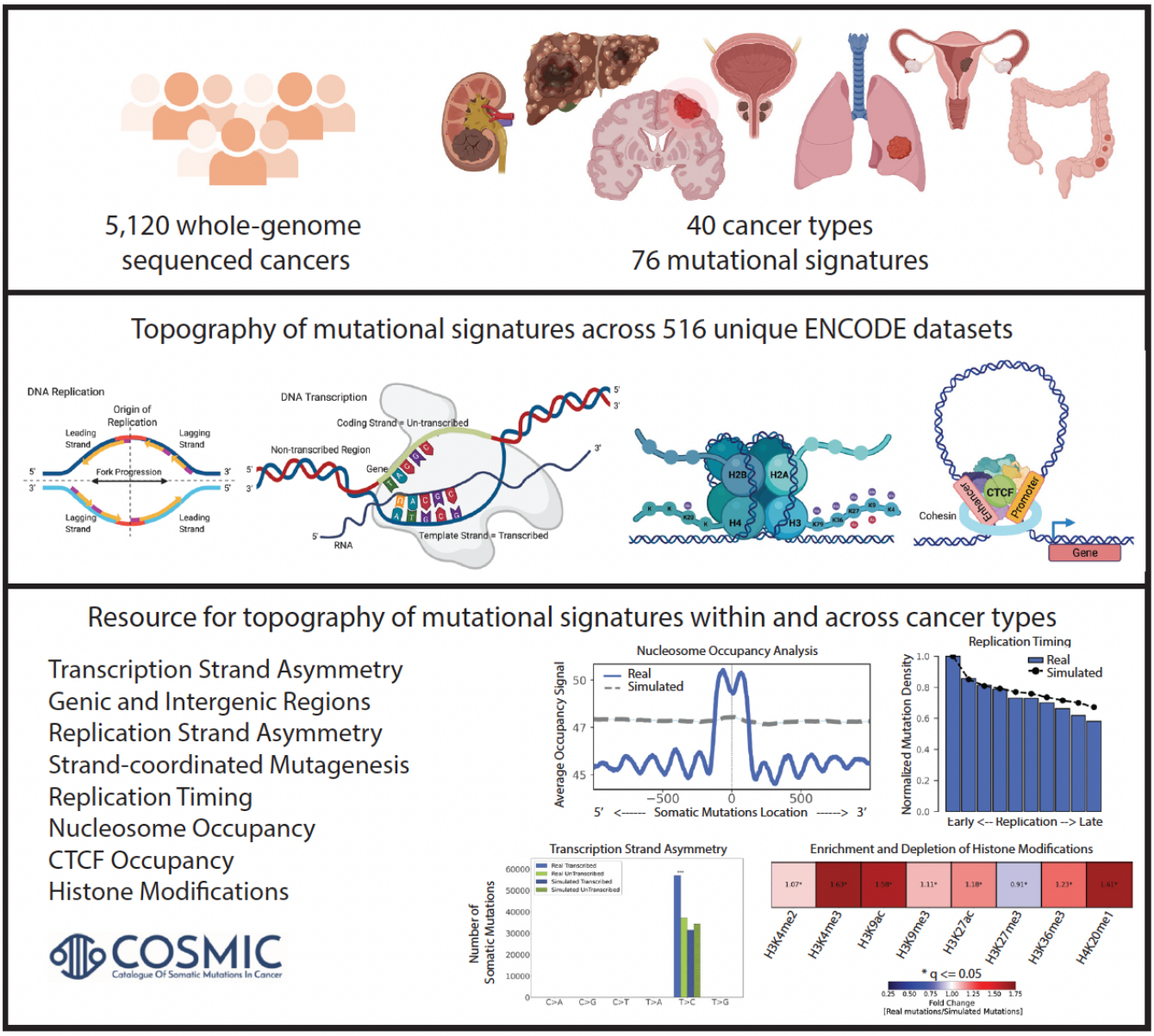
[1]Department of Cellular and Molecular Medicine, UC San Diego, La Jolla, CA, 92093, USA
[2]Department of Bioengineering, UC San Diego, La Jolla, CA, 92093, USA
[3]Moores Cancer Center, UC San Diego, La Jolla, CA, 92037, USA
[4]Graduate School of Informatics, Middle East Technical University, 06800, Ankara, Turkey
[5]COSMIC, Wellcome Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA, UK

[*]Correspondence should be addressed to L2alexandrov@health.ucsd.edu.

SUMMARY

The somatic mutations found in a cancer genome are imprinted by different mutational processes. Each process exhibits a characteristic mutational signature, which can be affected by the genome architecture. However, the interplay between mutational signatures and topographical genomic features has not been extensively explored. Here, we integrate mutations from 5,120 whole-genome sequenced tumours from 40 cancer types with 516 topographical features from ENCODE to evaluate the effect of nucleosome occupancy, histone modifications, CTCF binding, replication timing, and transcription/replication strand asymmetries on the cancer-specific accumulation of mutations from distinct mutagenic processes. Most mutational signatures are affected by topographical features with signatures of related aetiologies being similarly affected. Certain signatures exhibit periodic behaviours or cancer-type specific enrichments/depletions near topographical features, revealing further information about the processes that imprinted them. Our findings, disseminated via COSMIC, provide a comprehensive online resource for exploring the interactions between mutational signatures and topographical features across human cancer.

GRAPHICAL ABSTRACT



46 **HIGHLIGHTS**

47 • Comprehensive topography analysis of mutational signatures encompassing 82,890,857
48 somatic mutations in 5,120 whole-genome sequenced tumours integrated with 516 tissue-
49 matched topographical features from the ENCODE project.

50

51 • The accumulation of somatic mutations from most mutational signatures is affected by
52 nucleosome occupancy, histone modifications, CTCF binding sites, transcribed regions, or
53 replication strand/timing.

54

55 • Mutational signatures with related aetiologies are consistently characterized by similar
56 genome topographies across tissue types.

57

58 • Topography analysis allows both separating signatures from different aetiologies and
59 understanding the genomic specificity of clustered somatic mutations.

60

61 • A comprehensive online resource, disseminate through the COSMIC signatures database,
62 that allows researchers to explore the interactions between somatic mutational processes and
63 genome architecture within and across cancer types.

64 **Keywords:** somatic mutations; topography; mutational processes; mutational patterns;
65 mutational signatures; human cancer

66

67 **Manuscript length:** 6,934 words (including figure legends, but excluding STAR Methods,
68 supplemental item legends, and references section).

INTRODUCTION

Cancer genomes are peppered with somatic mutations imprinted by the activities of different endogenous and exogenous mutational processes[1,2]. Due to their intrinsic biophysical and biochemical properties, each mutational process engraves a characteristic pattern of somatic mutations, known as a mutational signature[3]. Our previous analyses encompassing more than 5,000 whole-genome and 20,000 whole-exome sequenced human cancers have revealed the existence of at least 78 single base substitution (SBS), 11 doublet-base substitution (DBS), and 18 indel (ID) mutational signatures[4-7]. Through statistical associations and further experimental characterizations, aetiology has been proposed for approximately half of the identified signatures[4,8-15]. Prior studies have also explored the interactions between somatic mutations imprinted by different mutational processes and the topographical features of the human genome for certain cancer types and a small subset of topographical features. However, previously, there has been no comprehensive evaluation that examined the effect of genome architecture and topographical features on the accumulation of somatic mutations from different mutational signatures across human cancer.

Early studies have shown that late replicating regions and condensed chromatin regions accumulate more mutations when compared to early replicating regions, actively transcribed regions, and open chromatin regions[16-19]. Subsequent analyses of hundreds of cancer genomes have revealed that differential DNA repair can explain variations in mutation rates across some cancer genomes[20] as well as that chromatin features originating from the cell of origin, which gave rise to the tumour, can affect mutation rate and the distribution of somatic mutations[17]. Recently, Morganella *et al.* examined the effect of the genomic and the epigenomic architecture

92 on the activity of 12 SBS signatures in breast cancer[21]. These analyses demonstrated that
93 mutations generated by different mutational processes exhibit distinct strand asymmetries and
94 that mutational signatures are differently affected by replication timing and nucleosome
95 occupancy[21]. Pan-cancer exploration of strand asymmetries was also conducted for different
96 mutation types across multiple cancer types[22] as well as for different mutational signatures[23]. In
97 particular, pan-cancer analyses of more than 3,000 cancers have revealed the strand asymmetries
98 and replication timings of the 30 SBS mutational signatures from the Catalogue of Somatic
99 Mutations in Cancer version 2 signatures database (COSMICv2)[23]. Similarly, more than 3,000
100 cancer genomes were used to elucidate the effect of nucleosome occupancy for the 30
101 substitution signatures from COSMICv2[24]. More recently, a study has also shown the interplay
102 between the three-dimensional genome organization and the activity of certain mutational
103 signatures[25,26].
104
105 Here, we report the most comprehensive evaluation of the effect of nucleosome occupancy,
106 histone modifications, CCCTC-binding factor (CTCF) binding sites, replication timing,
107 transcription strand asymmetry, and replication strand asymmetry on the cancer-specific
108 accumulation of somatic mutations from distinct mutational signatures. Our analysis leverages
109 the complete set of known COSMICv3.3 signatures (78 SBS, 11 DBS, and 18 ID) and it
110 examines 5,120 whole-genome sequenced cancers while simultaneously utilizing 516 unique
111 tissue-matched topographical features from the ENCODE project (**Table S1**)[27]. In all analyses,
112 the observed patterns of somatic mutations are compared to background simulation models of
113 mutational signatures that mimic both the trinucleotide pattern of these signatures as well as their
114 mutational burden within each chromosome in each examined sample (**Methods**). Our results

115 confirm many of the observations previously reported for strand asymmetry, replication timing,
116 and nucleosome periodicity for the original COSMICv2 signatures. Further, the richer and larger
117 COSMICv3.3 dataset allowed us to elucidate novel biological findings for some of these 30 SBS
118 signatures revealing previously unobserved pan-cancer and cancer-specific dependencies.
119 Additionally, this report provides the first-ever map of the genome topography of indel, doublet-
120 base, and another 24 substitution signatures in human cancer. Moreover, our study is the first to
121 examine the tissue-specific effect of CTCF binding and 11 different histone modifications on the
122 accumulation of somatic mutations from different mutational signatures. As part of the results,
123 we provide a global view of the topography of mutational signatures across 5,120 whole-genome
124 sequenced tumours from 40 types of human cancer and we include cancer type specific
125 examples. As part of the discussion, we zoom into two distinct case studies: *(i)* the topography of
126 different types of clustered somatic mutations; and *(ii)* using the topography of mutational
127 signatures to separate mutational signatures with similar patterns. Lastly, the reported results are
128 released as part of the COSMICv3.3 signatures database, https://cancer.sanger.ac.uk/signatures,
129 providing an unprecedented online resource for examining the topography of mutational
130 signatures within and across human cancer types.
131

132 **RESULTS**

133 **Transcription Strand Asymmetries**

134 Transcription strand asymmetries have been generally attributed to transcription-coupled
135 nucleotide excision repair (TC-NER) since bulky adducts (*e.g.*, ones due to tobacco carcinogens)
136 in actively transcribed regions of the genome will be preferentially repaired by TC-NER[28].
137 Additionally, transcription-coupled damage may also lead to transcription strand asymmetry due
138 to one of the strands being preferentially damaged during transcription[22].

139

140 Mutational signatures with similar aetiologies generally exhibited consistent patterns of
141 transcription strand asymmetries across cancer types. Specifically, most signatures attributed to
142 exogenous mutational processes showed transcription strand bias with mutations usually
143 enriched on the transcribed strand (**Figure 1*A&E***). This included signatures SBS4/DBS2 (both
144 previously attributed to mutagens in tobacco smoking), SBS16 (alcohol consumption), SBS24
145 (aflatoxin), SBS29 (tobacco chewing), SBS25/SBS31/SBS35/DBS5 (prior chemotherapy), and
146 SBS32 (prior treatment with azathioprine). Nevertheless, for some exogenous signatures, strand
147 asymmetries could differ between cancer types. For example, while transcriptional asymmetries
148 for C>A and T>A mutations from SBS4 were observed across most cancer types, asymmetries
149 for C>G mutations were only observed in lung adenocarcinoma and cancers of the head and neck
150 (**Figure 1*C***). Interestingly, C>T mutations attributed to SBS4 had strand asymmetry only in lung
151 adenocarcinoma. In contrast, mutational signatures due to direct damage from ultraviolet light
152 (*viz.*, SBS7a/b/c/d and DBS1) were the only known exogenous mutational processes to exhibit
153 transcription strand asymmetry with strong enrichment of mutations on the untranscribed strand,
154 consistent with damage from ultraviolet light on cytosine (**Figure 1*A&E***).

155 Transcription strand asymmetry with consistent enrichment of mutations on the transcribed
156 strand was also observed for clock-like signature SBS5 and for multiple mutational signatures
157 with unknown aetiology, including: SBS12, SBS19, and ID14 (**Figure 1*A&E***). Strand bias with
158 preferences for the untranscribed strand was observed for signatures ID11 and SBS33 (both with
159 unknown aetiology). Lastly, other mutational signatures exhibited transcription strand
160 asymmetry in only a small subset of cancer types (**Figure 1*A&E***).
161
162 **Mutational Signatures in Genic and Intergenic Regions**
163 Except for SBS16 and ID11, all other signatures were enriched in intergenic regions across most
164 cancer types with the enrichment ranging from 1.30-fold (*e.g.*, SBS24) to more than 2-fold (*e.g.*,
165 SBS17a/b; **Figure S1*A-C***). The observed depletion of mutations in genic regions was not due to
166 transcription strand asymmetries as correcting the asymmetries, by assigning the number of
167 mutations on both transcribed and untranscribed strands to their highest value, resulted in only a
168 minor alterations of the fold change increases (**Figure S1*D-E***). Overall, these results suggest that
169 transcription strand asymmetry, usually attributed to the activity of TC-NER, does not account
170 for the high enrichment of somatic mutations in intergenic regions.
171
172 SBS16 and ID11 showed enrichment of mutation in genic regions in liver and oesophageal
173 cancers, while ID11 was also enriched in genic regions in cancers of the head and neck. SBS16
174 has been previously associated with exposure to alcohol[29-31] and attributed to the activity of
175 transcription-coupled damage[22]. Prior studies have also associated ID11 to alcohol consumption
176 in oesophageal cancers[7]. Re-examining ID11 in the current cohort of whole-genome sequenced
177 liver cancers, by comparing the mutations attributed to ID11 in 32 heavy drinkers to the

178 mutations attributed to ID11 in 94 light drinkers, reveals a 2-fold enrichment in heavy drinkers
179 (p-value: 1.31 x 10^{-3}; Mann-Whitney U test). This and the prior associations in oesophageal
180 cancers[7] strongly suggest a similar exogenous mutational processes, related to alcohol
181 consumption, accounting for the enrichment of mutation in genic regions for both signatures
182 SBS16 and ID11.
183
184 **Replication Strand Asymmetries**
185 Replication strand bias was consistently observed in most signatures attributed to aberrant or
186 defective endogenous mutational processes with strand bias either on the leading or on the
187 lagging strand (**Figure 1*B&F***). Strong replication strand asymmetries with enrichment of
188 mutations on the leading strand was observed for signatures previously attributed to the defective
189 activity of polymerases, including: *(i)* SBS10a/SBS10b/DBS3 found in samples with
190 exonuclease domain mutations in DNA polymerase epsilon (*POLE*); *(ii)* SBS9, attributed to
191 infidelity of polymerase eta (POLH); and *(iii)* SBS10c due to defective polymerase delta
192 (*POLD1*). Interestingly, SBS28 (unknown aetiology) exhibited a strong replication strand bias
193 when found at high levels in POLE deficient samples.
194
195 Mutational signatures associated with defective DNA mismatch repair exhibited statistically
196 significant replication strand bias either predominately on the leading strand (*viz.*, SBS6) or on
197 the lagging strand (*viz.*, SBS14, SBS15, SBS20, SBS21, SBS26, SBS44, ID1). There were some
198 minor inconsistencies of replication strand bias across cancer types. For example, SBS44 did not
199 have replication strand asymmetry for C>T, T>A, and T>C mutations in oesophageal squamous
200 cell carcinoma (**Figure 1*D***). Somatic mutations due to signatures SBS2 and SBS13, both

201 attributed to the aberrant behaviour of the APOBEC3 family of deaminases[32], were found
202 enriched on the lagging strand in all cancer types. This result is consistent with the observation
203 that single-stranded DNA formed during DNA replication on the lagging strand is a major
204 substrate for the APOBEC3 family of deaminases[33,34]. Lastly, several other mutational
205 signatures, most with unknown aetiology, exhibited replication strand bias within a small set of
206 cancer types (**Figure 1*B&F***).
207
208 **Strand-coordinated Mutagenesis**
209 Prior analyses have shown that certain types of mutations on the same reference allele were
210 observed on the same strand more frequently than expected by chance[21,34,35]. This strand-
211 coordinated clustered mutations usually arise due to damage on single-stranded DNA, and they
212 are often indicative of the formation of hypermutable loci in the genome[33,34].
213
214 SBS7a, attributed to ultraviolet (UV) light, attained the highest strand-coordinated mutagenesis
215 with lengths of subsequent mutations up to 40 consecutive mutations (**Figure 1*G***). In contrast,
216 other mutational signatures attributed to ultraviolet light, mainly, SBS7b/c/d, either did not
217 exhibit or exhibited much lower strand-coordinated mutagenesis. APOBEC3-attributed SBS2
218 and SBS13 showed strand-coordinated mutagenesis with as many as 21 consecutive strand-
219 coordinated mutations. Additionally, strand-coordinated mutations were observed for SBS17b
220 (unknown aetiology), SBS10a/b (*POLE* deficiency), SBS4 (tobacco smoking), SBS26 (defective
221 mismatch repair), and SBS28 (unknown aetiology).
222
223

224 **The Effect of DNA Replication Timing**

225 Consistent with prior reports[18,36-38], the aggregated set of somatic mutations was shown to be
226 enriched in late replicating regions for most cancer types (**Figure 2*A***). Specifically, from the
227 examined 40 cancer types, SBSs were found more common in regions of the genome that
228 undergo late replication in 39/40 cancer types and were not associated with replication only in
229 uveal melanoma (**Figure 2*A***). Similarly, DBSs and IDs were enriched in late replicating regions
230 in 18/18 and 30/32 cancer types, respectively. Note that due to their lower mutational burdens,
231 we could confidently evaluate DBSs and IDs only in a subset of cancer types. In agreement with
232 the aggregates analysis, most mutational signatures imprinted somatic mutations with an
233 increased normalized mutational density from early to late replicating regions (**Figure S2**). For
234 example, SBS3 (defective homologous recombination) was enriched in late replicating regions in
235 all 14 cancer types where the signature can be confidently evaluated. Other examples include
236 signatures DBS2 and ID1, which were also consistently enriched in all examined cancer types
237 (**Figure 2*B***).

238

239 Nevertheless, at least seven mutational signatures were found predominately enriched in early
240 replicating regions, including: ID17 likely due to *TOP2A* mutations; SBS11 due to
241 temozolomide therapy; SBS16, ID11 both associated with alcohol consumption; SBS6, SBS15
242 both attributed to mismatch repair deficiency; and SBS84 due to the aberrant activities of
243 activation-induced (AID) cytidine deaminases (**Figure 2*C***; **Figure S2**). Moreover, multiple
244 mutational signatures were generally unaffected by replication timing, including: SBS7b (UV-
245 light); SBS20, SBS21, SBS44 (attributed to failure of mismatch repair); SBS30 (deficient base
246 excision repair); and SBS39, ID12 (unknown aetiology; **Figure 2*D***; **Figure S2**). The lack of

247 association with replication timing for some of these mutational signatures can be potentially
248 attributed to the activity of DNA translesion polymerases[39,40].
249
250 Interestingly, a number of mutational signatures exhibited cancer-type specific associations with
251 replication timing (**Figure S2**). For example, signature ID8 was enriched with replication timing
252 in 5 cancer types, depleted in 6 cancer types, and unaffected by replication timing in 7 cancer
253 types (**Figure 2*E***). Multiple aetiologies have been associated with ID8[4,41], including mutations
254 resulting in K743N amino acid change in TOP2A. All samples harbouring such mutations in
255 *TOP2A* exhibited an enrichment of ID8 in early replicating regions (**Figure 2*E***). The other
256 cancer-type specific mechanisms resulting in different associations with replication timing for
257 ID8 remain unknown.
258
259 Another notable example of cancer-type specific associations with replication timing is the
260 APOBEC3-associated SBS13 (**Figure S2**). SBS13 showed no dependence with replication
261 timing in 7/17 cancer types (*viz.*, bladder, breast, uterus, cervix, ovary, thyroid, and acute
262 lymphocytic leukaemia; **Figure 2*F***). This behaviour is consistent with prior reports where
263 SBS13 was attributed to uracil excision of deaminated cytosine followed by processing by DNA
264 translesion polymerases in breast cancer[39,40]. Surprisingly, in 10/17 cancer types, SBS13 was
265 highly enriched in late replicating regions. Using a previously defined approach for separating
266 the cancer samples on ones where SBS13 is not a hypermutator (low APOBEC3) and ones where
267 SBS13 is a hypermutator (high APOBEC3) revealed that the lack of dependence with replication
268 timing is predominately characteristic for hypermutated samples (**Figure 2*F***). This result

269 indicates that DNA translesion polymerases may play a significantly larger role in APOBEC3
270 hypermutators than previously anticipated.
271
272 **The Effect of Nucleosome Occupancy**
273 Nucleosomes are the basic packing units of chromatin with each nucleosome consisting of ~147
274 base-pair (bp) DNA wrapped around a histone octamer with 60 to 80 bp linker DNA between
275 consecutive nucleosomes[42,43]. Previous analyses have revealed dependencies between mutational
276 signatures operative in breast cancer and nucleosome occupancy[21] as well as a pan-cancer
277 periodicity of mutation rates within nucleosomes due to multiple substitution signatures[24].
278 However, beyond breast cancer, there has been no cancer-specific examination of the effect of
279 nucleosome occupancy on different mutational signatures.
280
281 Aggregated somatic mutations and mutations attributed to most mutational signatures were
282 depleted near nucleosomes compared to simulated data that mimic the mutational landscapes of
283 the examined cancer genomes (**Figure 3*A***). Remarkably, the majority of SBS, DBS, and ID
284 mutational signatures were similarly affected by nucleosome occupancy across most cancer
285 types (**Figure S3**). Some signatures were consistently enriched in the vicinity of nucleosomes.
286 For example, clock-like signature SBS1 exhibited a pattern closely mimicking simulated data
287 and showing higher number of mutations at nucleosomes in 36/36 cancer types, including
288 cancers of the lung, head and neck, liver, and oesophagus (**Figure 3*B***). In contrast, some
289 signatures were markedly different that simulated data (**Figure S3**), including signature DBS2
290 which was consistently depleted across 13/13 cancer types (**Figure 3*C***). Moreover, some
291 signatures were depleted in nucleosomes and strikingly appeared at linker DNA (**Figure S3**). For

292 example, clock-like signature ID1 was depleted when compared to simulated data and it
293 exhibited depletion in nucleosomes in 24/24 examined cancer types (**Figure 3*D***). The mutations
294 engraved by most flat mutational signatures (*e.g.*, SBS5, SBS8, SBS40) were generally
295 unaffected by nucleosomes (**Figure S3**).
296
297 Different types of periodicities of mutation rates around the nucleosome structure were observed
298 for signatures associated with tobacco smoking (SBS4 and ID3), ultraviolet light (SBS7a/b/c/d),
299 POLE deficiency (SBS10a), aristolochic acid (SBS22), and reactive oxygen species (SBS18,
300 SBS36, SBS38; **Figures 3*E* & S3**). Interestingly, signatures SBS17a/b also showed similar
301 periodic dependencies (**Figure 3*E*)** providing further circumstantial evidence for the hypothesis
302 that SBS17a/b may also be due to reactive oxygen species damage of the deoxyribonucleoside
303 triphosphate pools[23,44-48]. With the exception of signature SBS22 and ID3, all other periodic
304 signatures exhibited enrichment of mutations at nucleosomes (**Figure 3*E&G***). Further, for most
305 signatures periodicity of mutation rates was observed in each cancer types where the signature
306 was operative (**Figure S3)**. Nevertheless, signature SBS4 showed strong periodicity in cancers of
307 the lung and head and neck but not in cancers of the liver or oesophagus (**Figure 3*F***). Similarly,
308 signature ID3 exhibited periodic behaviour only in cancers of the lung but not in any other
309 cancer type (**Figure 3*G***).
310
311 **The Effect of CTCF Binding**
312 CCCTC-binding factor (CTCF) is a multi-purpose sequence-specific DNA-binding protein with
313 an essential role in transcriptional regulation, somatic recombination, and chromatin

314 architecture[49]. The human genome harbours many CTCF binding sites with prior studies
315 reporting that mutations due to ultraviolet light are enriched in CTCF binding sites[50].
316
317 Somatic mutations exhibited clear patterns of both enrichment and/or periodicity for multiple
318 mutational signatures and CTCF binding sites (**Figure 4**). While some signatures were
319 consistently depleted at CTCF biding sites across the majority of cancer types when compared to
320 simulated data (SBS1, SBS9, SBS10a/b, SBS15, SBS37, SBS84, and SBS85), others were
321 commonly enriched (SBS3, SBS5, SBS7a/b/d, SBS12, SBS17a/b, SBS18, SBS22, and SBS40;
322 DBS1; ID5, ID6, ID8, and ID9; **Figure 4*A***).
323
324 Aggregated single base substitutions exhibited an inconsistent behaviour across cancer types
325 with enrichment in some cancers (*e.g.,* liver cancers) and depletions in others (*e.g.*, lymphomas).
326 In contrast, indels were enriched at CTCF binding sites in the majority of cancer types (**Figure**
327 **4*A***). Remarkably, the effect of CTCF occupancy tended to be also consistent for many signatures
328 with similar aetiologies. Strong periodicities of mutation rates around CTCF binding sites were
329 observed for UV-associated signature SBS7a but not for UV-associated signatures DBS1 and
330 SBS7b/c/d (**Figure 4*B***).
331
332 Mutations due to SBS9, associated with defective polymerase eta driven replication errors, and
333 signatures SBS10a/b, found in samples with mutations in *POLE* and/or *POLD1*, were strikingly
334 depleted at CTCF binding sites. Signatures SBS15, associated microsatellite instability, was
335 strongly depleted at CTCF binding sites (**Figure 4*A***).
336

337 Only one of the clock-like signatures, SBS1, exhibited a depletion of mutations at CTCF binding
338 sites (**Figure 4*A***) while simulated data indicated that SBS1 should be enriched at these sites
339 (**Figure 4*B***). Signature SBS3, attributed to defective homologous recombination, was highly
340 elevated in CTCF binding sites for breast, ovarian, stomach, and oesophageal cancers. Signatures
341 SBS17a/b exhibited a striking enrichment at CTCF binding sites in all cancer types with
342 sufficient number of mutations from each signature (**Figure 4*A***). SBS17a showed enrichment in
343 stomach and oesophageal cancers, while SBS17b shows enrichment for stomach, oesophageal,
344 breast, pancreatic cancers, and non-Hodgkin's lymphomas. In contrast, simulated data indicate
345 that CTCF binding should have no effect on the accumulation of mutations from signatures
346 SBS17a/b (**Figure 4*B***).
347
348 **The Effect of Histone Modifications**
349 Each nucleosome consists of four pairs of core histones: H2A, H2B, H3, and H4. Post-
350 translational modifications of histone tails play a key role in regulating DNA replication, gene
351 transcription, and DNA damage response[51]. To evaluate the effect of histone modifications on
352 the accumulation of mutations from different mutational signatures, we mapped the depletion or
353 enrichment of mutations compared to simulated data in the context of the tissue specific
354 positions of 11 histone modifications: *(i)* H2AFZ, a replication-independent member of the
355 histone H2A family that renders chromatin accessible at enhancers and promoters
356 regulating transcriptional activation and repression[52]; *(ii)* H3K4me1, histone mark often
357 associated with enhancer activity[53] ; *(iii)* H3K4me2, a histone post-translational modification
358 enriched in *cis*-regulatory regions, including both enhancers and promoters[54]; *(iv)* H3K4me3,
359 post-translational modification enriched in active promoters near transcription start sites[55]; *(v)*

360 H3K9ac, associated with active gene promoters and active transcription[56]; *(vi)* H3K9me3,
361 silencer, typical mark of constitutive heterochromatin[57]; *(vii)* H3K27ac, histone modification
362 generally contained at nucleosomes flanking enhancers[55]; *(viii)* H3K27me3, repressive,
363 associated with silent genes[58]; *(ix)* H3K36me3, associated with transcribed regions and playing a
364 role in regulating DNA damage repair[59]; *(x)* H3K79me2, detected in the transcribed regions of
365 active genes[60]; and *(xi)* H4K20me1, found in gene promoters and associated with gene
366 transcriptional elongation and transcription activation[61].
367
368 Aggregated substitutions, dinucleotides, and indels exhibited dissimilar behaviour for different
369 histone modifications across cancer types (**Figure S4**). Aggregated substitutions were
370 predominately depleted around H2AFZ, H3K4me2, H3K4me3, and H3K27ac in approximately
371 half of the examined cancer types (**Figure S4*A*-*C***). Aggregated doublets and indels did not have
372 any clear pan-cancer preference but showed cancer-type specific enrichments and depletions. In
373 contrast, majority of mutational signatures had generally similar behaviour in vicinity of
374 different histone modifications revealing that histone modifications have similar effect on
375 mutagenesis across cancer types (**Figure S4**). Most SBS mutational signatures were either
376 unaffected or depleted near histone marks (**Figure S4*A***). Notable exceptions were APOBEC3-
377 associated signatures SBS2 and SBS13, AID-associated signatures SBS84 and SBS85, and
378 *POLH* attributed SBS9 which were generally enriched near most histone modifications (**Figure**
379 **S4*A***). Doublet signatures DBS1, DBS2, DBS3, DBS4, and DBS5 were also predominately
380 depleted near most histone marks (**Figure S4*B***). In contrast, signatures DBS7, DBS9, and
381 DBS11 were highly enriched near most histone marks. Most ID mutational signatures were
382 either unaffected or very highly enriched near histone marks (**Figure S4*C***) with the only

383 exceptions being depletions of: *(i)* ID1 and ID6 near H2AZ; *(ii)* ID3 in the vicinity of H3K4me3;
384 *(iii)* ID5 near H3K27me3; and *(iv)* ID14 in the vicinity H3K36me3. While enrichments and
385 depletions of somatic mutations in the vicinity of histone marks were commonly observed for
386 different mutational signatures (**Figure S4*A-C***), there was no specific pattern of mutations
387 within 1,000 base-pairs for any of the examined histone modifications. Exemplars of typically
388 observed patterns of enrichments, depletions, or no changes around different histone
389 modifications are provided for signatures SBS7a and ID1 across several histone modifications
390 (**Figure S5*D***).
391
392 Next, we examine two mutational signatures that exhibited inconsistent enrichments and
393 depletions near specific histone marks. Clock-like signature SBS1 was consistently depleted
394 across cancer-types for multiple histone marks, including H3K9me3 (**Figure 5*A***). Nevertheless,
395 SBS1 exhibited enrichment of mutations near H3K9me3 in two cancer types of the central
396 nervous system, depletion of mutations near H3K9me3 in three haematological malignancies,
397 and no effect in all other solid tumour types (**Figure 5*A***). Similarly, signature ID1 exhibited
398 dissimilar behaviour near H3K27ac with enrichments in medulloblastoma and lymphoma,
399 depletions in stomach and prostate cancer, and no change in most other cancer types (**Figure**
400 **5*B***).
401

402 **DISCUSSION**

403 Our analysis provides a comprehensive resource that maps the effects of topographical genomic
404 features on the cancer-specific accumulation of somatic mutations from distinct mutational
405 signatures. The reported results confirmed many of the prior observations for strand asymmetry,
406 replication timing, and nucleosome periodicity for some of the original 30 COSMICv2 SBS
407 signatures[21,23,24]. The examined larger dataset provided us with a greater resolution to identify
408 previously unobserved pan-cancer and cancer-specific dependencies for some of these 30
409 signatures as well as to reveal the effect of genome architecture on the accumulation of another
410 46 mutational signatures across human cancer. Importantly, this report also provides the first-
411 ever examination of the tissue-specific effect of CTCF binding and 11 different histone
412 modifications on the accumulation of somatic mutations from different mutational signatures. In
413 addition to the comprehensive global view in the results section, in this discussion, we zoom into
414 two specific case studies to further illustrate the power of using this resource for examining
415 topography of mutational signatures.

416

417 First, analysis of SBS28 in *POLE* deficient samples (*POLE*$^{-}$) and *POLE* proficient samples
418 (*POLE*$^{+}$) revealed a distinct behaviour (**Figure 6**). While the trinucleotide patterns of SBS28 in
419 *POLE*$^{+}$ and *POLE*$^{-}$ samples were similar (cosine similarity: 0.96), SBS28 in POLE^{-} samples
420 accounted for 97.7% mutations of all SBS28 mutations and it exhibited a clear enrichment in late
421 replicating regions as well as depletions at nucleosomes and at CTCF binding sites (**Figure 6*B-***
422 ***D,F***). Moreover, SBS28 in *POLE*$^{-}$ samples showed a strong replication strand bias on the leading
423 strand and exhibited a strand-coordinated mutagenesis with as many as 11 consecutively mutated
424 substitutions (**Figure 6*E,G***). In contrast, SBS28 in *POLE*$^{+}$ samples were enriched in early

425 replication regions, lacked depletion of mutations at nucleosomes or CTCF binding sites, had
426 weak replication strand bias on lagging strand, and did not exhibit much of a strand-coordinated
427 mutagenesis (**Figure 6**). Based on these topographical differences, we have now split SBS28 into
428 two distinct signatures: *(i)* SBS28a due to *POLE* deficiency found in ultra-hypermutate
429 colorectal and uterine cancers; and *(ii)* SBS28b with unknown aetiology found in lung and
430 stomach cancers.
431
432 Second, our analyses revealed striking difference in topographical features of clustered and non-
433 clustered somatic mutations in 288 whole-genome sequenced B-cell malignancies[4]. In particular,
434 the topographical behaviours of single base substitutions were examined after separating them
435 into non-clustered mutations, diffuse hypermutation of substitutions termed *omikli*[62], and longer
436 clusters of strand-coordinated substitutions termed *kataegis*[34,35,63]. In contrast to most cancer
437 types, where *omikli* and *kataegis* are predominately generated by APOBEC3 deaminases[64], in B-
438 cell malignancies, these clustered events are almost exclusively imprinted by the activity of
439 AID[64]. Further, the overall pattern of non-clustered mutations was very different than the ones of
440 *omikli* or *kataegis*. A representative example is provided using a single malignant B-cell
441 lymphoma (**Figure 7*A***) where non-clustered and clustered mutations have very different
442 trinucleotide patterns (**Figure 7*B-D***). Non-clustered mutations exhibited different topographical
443 features when compared to *omikli* or *kataegis*. Specifically, while non-clustered mutations had
444 some minor periodicity in regard to nucleosome occupancy, such periodicity was not observed
445 for any type of clustered events (**Figure 7*E***). Similarly, non-clustered mutations were slightly
446 depleted around CTCF binding sites while *omikli* and *kataegis* were very highly depleted
447 (**Figure 7*F&H***). Further, non-clustered and *omikli* events were clearly enriched in late

448 replication regions while *kataegis* was highly enriched in early replication regions (**Figure 7*G***).
449 Distinct patterns of enrichments were also observed for both *omikli* and *kataegis* mutations in the
450 vicinity of promoter and enhancer sites delineated by histone marks of H3K4me3, H3K9ac,
451 H3K27ac, H3K36me3, and H4K20me1(**Figure 7*H***). Only very minor differences were observed
452 for transcription or replication strand asymmetries between clustered and non-clustered somatic
453 mutations across the 288 whole-genome sequenced B-cell malignancies (**Figure S5**).
454
455 In summary, in this report we have performed a comprehensive topography analysis of
456 mutational signatures encompassing 82,890,857 somatic mutations in 5,120 whole-genome
457 sequenced tumours integrated with 516 tissue-matched topographical features from the
458 ENCODE project. Our evaluation encompassed examining the effects of nucleosome occupancy,
459 histone modifications, CTCF binding sites, replication timing, transcription strand asymmetry,
460 and replication strand asymmetry on the accumulation of somatic mutations from more than 70
461 distinct mutational signatures. The results from these analyses have been provided as an online
462 resource as a part of COSMIC signatures database, https://cancer.sanger.ac.uk/signatures/, where
463 researchers can explore each mutational signature as well as each topographical feature in a
464 cancer-specific manner.
465

466 **FIGURE LEGENDS**

467 **Figure 1. Strand asymmetries and strand-coordinated mutagenesis.** ***(A)*** Transcription strand
468 asymmetries of signatures of single base substitutions (SBSs). Rows represent the signatures,
469 where *n* reflects the number of cancer types in which each signature was found. Columns display
470 the six substitution subtypes based on the mutated pyrimidine base: C>A, C>G, C>T, T>A, T>C,
471 and T>G. SBS signatures with transcription strand asymmetries on the transcribed and/or
472 untranscribed strands with q-values ≤ 0.05 are shown in circles with blue and green colours,
473 respectively. The colour intensity reflects the odds ratio between the ratio of real mutations and
474 the ratio of simulated mutations, where each ratio is calculated using the number of mutations on
475 the transcribed strand and the number of mutations on the untranscribed strand. Only odds ratios
476 above 1.10 are shown. Circle sizes reflect the proportion of cancer types exhibiting a signature
477 with specific transcription strand asymmetry. ***(B)*** Replication strand asymmetries of SBS
478 signatures. Rows represent the signatures, where *n* reflects the number of cancer types in which
479 each signature was found. Columns display the six substitution subtypes based on the mutated
480 pyrimidine base: C>A, C>G, C>T, T>A, T>C, and T>G. SBS signatures with replicational
481 strand asymmetries on the lagging strand or leading strand with q-values ≤ 0.05 are shown in
482 circles with red and yellow colours, respectively. The colour intensity reflects the odds ratio
483 between the ratio of real mutations and the ratio of simulated mutations, where each ratio is
484 calculated using the number of mutations on the lagging strand and the number of mutations on
485 the leading strand. Only odds ratios above 1.10 are shown. Circle sizes reflect the proportion of
486 cancer types exhibiting a signature with specific replication strand asymmetry. ***(C)*** Transcription
487 strand asymmetries of signatures of SBS4 across cancer types. Data are presented in a format
488 similar to the one in panel *(A)*. ***(D)*** Replication strand asymmetries of signature SBS44 across

489 cancer types. Data are presented in a format similar to the one in panel *(B)*. ***(E)*** Transcription
490 strand asymmetries of signatures of doublet-base substitutions (DBSs) and of small
491 insertions/deletions (IDs). Data are presented in a format similar to the one in panel *(A)*. ***(F)***
492 Replication strand asymmetries of DBS and ID mutational signatures. Data are presented in a
493 format similar to the one in panel *(B)*. ***(G)*** Strand-coordinated mutagenesis of SBS signatures.
494 Rows represent SBS signatures and columns reflect the lengths, in numbers of consecutive
495 mutations, of strand-coordinated mutagenesis groups. SBS signatures with statistically
496 significant strand-coordinated mutagenesis (q-values ≤ 0.05) are shown as circles under the
497 respective group length with a minimum length of 5 consecutive mutations. The size of each
498 circle reflects the number of consecutive mutation groups for the specified group length
499 normalized for each signature. The colour of each circle reflects the statistical significance of the
500 number of subsequent mutation groups for each group length with respect to simulated
501 mutations.
502
503 **Figure 2. Interplay between replication timing and mutational signatures.** Replication time
504 data are separated into deciles, with each segment containing exactly 10% of the observed
505 replication time signal (x-axes). Normalized mutation densities per decile (y-axes) are presented
506 for early (left) to late (right) replication domains. Real data for single base substitutions (SBSs)
507 signatures are shown as blue bars, for doublet-base substitutions (DBSs) signatures as red bars,
508 and for small insertions/deletions (IDs) signatures as green bars. Simulated somatic mutations
509 are shown as dashed lines. Where applicable, the total number of evaluated cancer types for a
510 particular mutational signature is shown on top of each plot (*e.g.,* 18 cancer types were evaluated
511 for ID8 in panel E). For each signature, the number of cancer types where the mutation density

512 increases with replication timing is shown next to ➚ (*e.g.,* 5 cancer types for ID8). Similarly, the
513 number of cancer types where the mutation density decreases with replication timing is shown
514 next to ➘ (*e.g.,* 6 cancer types for ID8). Lastly, the number of cancer types where the mutation
515 density is not affected by replication timing is shown next to ➞ (*e.g.,* 7 cancer types for ID8).
516 ***(A)*** All SBSs, DBSs, and IDs across all examined cancer types with each cancer type examined
517 separately. ***(B)*** Exemplar signatures consistently associated with late replication timing. ***(C)***
518 Exemplar signatures consistently associated with early replication timing. ***(D)*** Exemplar
519 signatures consistently unaffected by replication timing. ***(E)*** ID8 as a mutational signature
520 inconsistently affected by replication timing. ***(F)*** The effect of replication timing on APOBEC3-
521 associated signature SBS13 in samples with low and high APOBEC3 mutational burden.
522
523 **Figure 3. Relationship between mutational signatures and nucleosome occupancy.** In all
524 cases, solid lines correspond to real somatic mutations with blue solid lines reflecting single base
525 substitutions (SBSs), red doublet-base substitutions (DBSs), and green small insertions/deletions
526 (IDs). Simulated somatic mutations are shown as dashed lines. Solid lines and dashed lines
527 display the average nucleosome signal (y-axes) along a 2 kilobase window (x-axes) centred at
528 the mutation start site for real and simulated mutations, respectively. The mutation site is
529 annotated in the middle of each plot and denoted as 0. The 2 kilobase window encompasses
530 1,000 base-pairs 5' adjacent to each mutation as well as 1,000 base-pairs 3' adjacent to each
531 mutation. Where applicable, the total number of similar and considered cancer types using an
532 *X/Y* format, with *X* being the number of cancer types where a signature has similar nucleosome
533 behaviour (Pearson correlation ≥ 0.5 and q-value ≤ 0.05) and *Y* representing the total number of
534 examined cancer types for that signature. For example, signature ID3 in panel G annotated with

535 6/9 reflects a total of 9 examined cancer types with similar nucleosome behaviour observed in 6
536 cancer types. ***(A)*** All SBSs, DBSs, and IDs across all examined cancer types with each cancer
537 type examined separately. The nucleosome occupancy of signatures SBS1 ***(B)***, DBS2 ***(C)***, and
538 ID1 ***(D)*** are shown across all cancer types as well as within cancers of the lung, head and neck,
539 liver, and oesophagus. ***(E)*** Signatures with consistent periodicities of mutation rates around the
540 nucleosome. Tobacco-associated SBS4 ***(F)*** and ID3 ***(G)*** exhibiting periodicities of mutation rates
541 only in certain cancer types.
542
543 **Figure 4. Relationship between mutational signatures and CTCF binding sites.** ***(A)***
544 Enrichments and depletions of somatic mutations within CTCF binding sites. Heatmaps display
545 only mutational signatures and cancer types that have at least one statistically significant
546 enrichment or depletion of somatic mutations attributed to signatures of either single base
547 substitutions (SBSs), doublet-base substitutions (DBSs), or small insertions/deletions (IDs). Red
548 colours correspond to enrichments of real somatic mutations when compared to simulated data.
549 Blue colours correspond to depletions of real somatic mutations when compared to simulated
550 data. The intensities of red and blue colours reflect the degree of enrichments or depletions based
551 on the fold change. White colours correspond to lack of data for performing statistical
552 comparisons *(e.g.*, signature not being detected in a cancer type). Statistically significant
553 enrichments and depletions are annotated with * (q-value ≤ 0.05). ***(B)*** The top three panels
554 reflect average CTCF occupancy signal for all SBSs, DBS, and IDs across all examined cancer
555 types. Bottom panels reflect all somatic mutations attributed for several exemplar mutational
556 signatures across all cancer types. In all cases, solid lines correspond to real somatic mutations
557 with blue solid lines reflecting SBSs, red solid lines reflecting DBSs, and green solid lines

558 reflecting IDs. Solid lines and dashed lines display the average CTCF binding signal (y-axes)
559 along a 2 kilobase window (x-axes) centred at the mutation start site for real and simulated
560 mutations, respectively. The mutation start site is annotated in the middle of each plot and
561 denoted as 0. The 2 kilobase window encompasses 1,000 base-pairs 5' adjacent to each mutation
562 as well as 1,000 base-pairs 3' adjacent to each mutation.
563
564 **Figure 5. Exemplar relationships between mutational signatures and histone modifications.**
565 The effect of histone modifications is shown for signatures SBS1 ***(A)*** and ID1 ***(B)***. For each
566 signature, an evaluation was made for each of the 11 histone marks across all cancer types with
567 sufficient numbers of somatic mutations with results presented as circles. Each circle is separated
568 in red, blue, and grey segments proportionate to the cancer types in which the signature has a
569 specific behaviour. A red segment in a circle reflects the signature being enriched in the vicinity
570 of a histone modification (q-value ≤ 0.05 and at least 5% enrichment). A blue segment in a circle
571 reflects the signature being depleted in the vicinity of a histone modification (q-value ≤ 0.05 and
572 at least 5% depletion). A grey segment in a circle corresponds to neither depletion nor
573 enrichment of the signature in the vicinity of a histone modification. The figure zooms into the
574 effect of H3K9me3 on SBS1 ***(A)*** and H3K27ac on ID1 ***(B)***. Red colours correspond to
575 enrichments of real somatic mutations when compared to simulated data. Blue colours
576 correspond to depletions of real somatic mutations when compared to simulated data. The
577 intensities of red and blue colours reflect the degree of enrichments or depletions based on the
578 fold change. We further zoom into the patterns of mutations around H3K9me3 and H3K27ac.
579 Solid lines correspond to real somatic mutations with blue solid lines reflecting SBSs and green
580 solid lines reflecting IDs. Solid lines and dashed lines display the average histone mark signal (y-

581 axes) along a 2 kilobase window (x-axes) centred at the mutation start site for real and simulated
582 mutations, respectively. The mutation start site is annotated in the middle of each plot and
583 denoted as 0. The 2 kilobase window encompasses 1,000 base-pairs 5' adjacent to each mutation
584 as well as 1,000 base-pairs 3' adjacent to each mutation.
585
586 **Figure 6. Topography of signature SBS28 in *POLE* deficient (*POLE*$^-$) and *POLE* proficient**
587 **(*POLE*$^+$) samples.** ***(A)*** Mutational patterns of signature SBS28 in *POLE*$^-$ and *POLE*$^+$ samples
588 displayed using the conventional 96 mutational classification schema for single base
589 substitutions. ***(B)*** Nucleosome occupancy of SBS28 in *POLE*$^-$ and *POLE*$^+$ samples. Blue solid
590 lines and grey dashed lines display the average nucleosome signal (y-axes) along a 2 kilobase
591 window (x-axes) centred at the mutation start site for real and simulated mutations, respectively.
592 The mutation start site is annotated in the middle of each plot and denoted as 0. The 2 kilobase
593 window encompasses 1,000 base-pairs 5' adjacent to each mutation as well as 1,000 base-pairs
594 3' adjacent to each mutation. ***(C)*** CTCF occupancy of SBS28 in *POLE*$^-$ and *POLE*$^+$ samples.
595 Blue solid lines and grey dashed lines display the average CTCF binding signal (y-axes) along a
596 2 kilobase window (x-axes) centred at the mutation start site for real and simulated mutations,
597 respectively. The mutation start site is annotated in the middle of each plot and denoted as 0. The
598 2 kilobase window encompasses 1,000 base-pairs 5' adjacent to each mutation as well as 1,000
599 base-pairs 3' adjacent to each mutation. ***(D)*** Replication timing of SBS28 mutations in *POLE*$^-$
600 and *POLE*$^+$ samples. Replication time data are separated into deciles, with each segment
601 containing exactly 10% of the observed replication time signal (x-axes). Normalized mutation
602 densities per decile (y-axes) are presented for early (left) to late (right) replication domains.
603 Normalized mutation densities of real somatic mutations and simulated somatic mutations from

604 early to late replicating regions are shown as blue bars and dashed lines, respectively. ***(E)***
605 Replication strand asymmetry of SBS28 mutations in *POLE*$^-$ and *POLE*$^+$ samples. Bar plots
606 display the number of mutations accumulated on the lagging strand and leading strand for six
607 substitution subtypes based on the mutated pyrimidine base: C>A, C>G, C>T, T>A, T>C, and
608 T>G in red and yellow colours, respectively. Simulated mutations on lagging and leading strands
609 are displayed in shaded bar plots. Statistically significant strand asymmetries are shown with
610 stars: * q-value ≤ 0.05; ** q-value ≤ 0.01; *** q-value ≤ 0.001. ***(F)*** Enrichments and depletions
611 of SBS28 somatic mutations in *POLE*$^-$ and *POLE*$^+$ samples within CTCF binding sites, histone
612 modifications, and nucleosome occupied regions. Red colours correspond to enrichments of real
613 somatic mutations when compared to simulated data. Blue colours correspond to depletions of
614 real somatic mutations when compared to simulated data. The intensities of red and blue colours
615 reflect the degree of enrichments or depletions based on the fold change. White colours
616 correspond to lack of data for performing statistical comparisons. Statistically significant
617 enrichments and depletions are annotated with * (q-value ≤ 0.05). ***(G)*** Strand-coordinated
618 mutagenesis of SBS28 mutations in *POLE*$^-$ and *POLE*$^+$ samples. Rows represent SBS28
619 signature in *POLE*$^-$ and *POLE*$^+$ samples across all cancer types and columns reflect the lengths,
620 in numbers of consecutive mutations, of strand-coordinated mutagenesis groups. Statistically
621 significant strand-coordinated mutagenesis (q-value ≤ 0.05) are shown as circles under the
622 respective group length with a length starting from 2 to 11 consecutive mutations. The size of
623 each circle reflects the number of consecutive mutation groups for the specified group length
624 normalized for each SBS28 signature in *POLE*$^-$ and *POLE*$^+$ samples. The colour of each circle
625 reflects the statistical significance of the number of subsequent mutation groups for each group
626 length with respect to the simulated mutations using $-\log_{10}$(q-value).

627 **Figure 7. Topography of non-clustered, *omikli*, and *kataegis* substitutions across 288 whole-**
628 **genome sequenced B-cell malignancies**. ***(A)*** A rainfall plot of an example B-cell malignancy
629 sample, MALY-DE_SP116612, depicting the intra-mutational distance (IMD) distributions of
630 substitutions across genomic coordinates. Each dot represents the minimum distance between
631 two adjacent mutations. Dots are coloured based on their corresponding classifications.
632 Specifically, non-clustered mutations are shown in grey, doublet-base substitutions (DBSs) in
633 red, multi-base substitutions (MBSs) in black, *omikli* events in green, *kataegis* events in orange,
634 and all other clustered events in blue. The red line depicts the sample-dependent IMD threshold
635 for each sample. Specific clustered mutations may be above this threshold due to corrections for
636 regional mutation density. ***(B-D)*** The trinucleotide mutational spectra for the different catalogues
637 of non-clustered, *omikli,* and *kataegis* mutations for the exemplar sample (DBSs and MBSs are
638 not shown). ***(E)*** Nucleosome occupancy of non-clustered, *omikli*, and *kataegis* mutations of B-
639 cell malignancies. Blue solid lines and grey dashed lines display the average nucleosome signal
640 (y-axes) along a 2 kilobase window (x-axes) centred at the mutation start site for real and
641 simulated mutations, respectively. The mutation start site is annotated in the middle of each plot
642 and denoted as 0. The 2 kilobase window encompasses 1,000 base-pairs 5' adjacent to each
643 mutation as well as 1,000 base-pairs 3' adjacent to each mutation. ***(F)*** CTCF occupancy of non-
644 clustered, *omikli*, and *kataegis* mutations of B-cell malignancies. Blue solid lines and grey
645 dashed lines display the average CTCF signal (y-axes) along a 2 kilobase window (x-axes)
646 centred at the mutation start site for real and simulated mutations, respectively. The mutation
647 start site is annotated in the middle of each plot and denoted as 0. The 2 kilobase window
648 encompasses 1,000 base-pairs 5' adjacent to each mutation as well as 1,000 base-pairs 3'
649 adjacent to each mutation. ***(G)*** Replication timing of non-clustered, *omikli*, and *kataegis*

650 mutations of B-cell malignancies. Replication time data are separated into deciles, with each
651 segment containing exactly 10% of the observed replication time signal (x-axes). Normalized
652 mutation densities per decile (y-axes) are presented for early (left) to late (right) replication
653 domains. Normalized mutation densities of real somatic mutations and simulated somatic
654 mutations from early to late replicating regions are shown as blue bars and dashed lines,
655 respectively. ***(H)*** Enrichments and depletions of non-clustered, *omikli*, and *kataegis* mutations of
656 B-cell malignancies within CTCF binding sites and histone modifications. Red colours
657 correspond to enrichments of real somatic mutations when compared to simulated data. Blue
658 colours correspond to depletions of real somatic mutations when compared to simulated data.
659 The intensities of red and blue colours reflect the degree of enrichments or depletions based on
660 the fold change. White colours correspond to lack of data for performing statistical comparisons.
661 Statistically significant enrichments and depletions are annotated with * (q-value ≤ 0.05).
662

663 **SUPPLEMENTARY FIGURE LEGENDS**

664 **Figure S1. Somatic mutations in genic and intergenic regions imprinted by different**
665 **mutational signatures. Related to Figure 1.** ***(A)*** Somatic mutations in genic and intergenic
666 regions for signatures of single base substitutions (SBSs). Rows represent the signatures, where *n*
667 reflects the number of cancer types in which each signature was found. Columns display the six
668 substitution subtypes based on the mutated pyrimidine base: C>A, C>G, C>T, T>A, T>C, and
669 T>G. SBS signatures with genic and intergenic regions asymmetries with q-values ≤ 0.05 are
670 shown in circles with cyan and grey colours, respectively. The colour intensity reflects the odds
671 ratio between the ratio of real mutations and the ratio of simulated mutations, where each ratio is
672 calculated using the number of mutations in the genic regions and the number of mutations in the
673 intergenic regions. Only odds ratios above 1.10 are shown. Circle sizes reflect the proportion of
674 cancer types exhibiting a signature with specific genic versus intergenic regions asymmetry. ***(B)***
675 Somatic mutations in genic and intergenic regions for signatures of doublet-base substitutions
676 (DBSs). Data are presented in a format similar to the one in panel *(A)*. ***(C)*** Somatic mutations in
677 genic and intergenic regions for small insertions/deletions (IDs). Data are presented in a format
678 similar to the one in panel *(A)*. ***(D)*** Histogram of fold enrichment as odds ratio between the ratio
679 of real mutations and the ratio of simulated mutations, where each ratio is calculated using the
680 number of mutations in the genic regions and the number of mutations in the intergenic regions.
681 Frequency of fold enrichments (y-axis) are presented for discreet bins of fold enrichments (x-
682 axis). Each fold enrichment reflects the odds ratio between real and simulated mutations where
683 each ratio is the number of mutations in intergenic regions divided by the number of mutations in
684 genic regions. Total number of fold enrichments, mean, and standard deviation of fold
685 enrichments are shown in the upper right corner of the histogram. ***(E)*** Same format as panel *(D)*

686 with the underlaying data reflecting fold enrichments after inflating the number of somatic
687 mutations in genic regions to remove any transcription strand asymmetry.
688
689 **Figure S2. The effect of replication timing on mutational signatures. Related to Figure 2.**
690 Top three panels reflect results for all single base substitutions (SBSs), all doublet-base
691 substitutions (DBSs), and all small insertions/deletions (IDs) across all examined cancer types
692 with each cancer type examined separately. Bottom panels reflect all somatic mutations
693 attributed to a particular signature across all cancer types. Replication time data are separated
694 into deciles, with each segment containing exactly 10% of the observed replication time signal
695 (x-axes). Normalized mutation densities per decile (y-axes) are presented for early (left) to late
696 (right) replication domains. Real data for SBS signatures are shown as blue bars, for DBS
697 signatures as red bars, and for ID signatures as green bars. In all cases, simulated somatic
698 mutations are shown as dashed lines. The total number of evaluated cancer types for a particular
699 mutational signature is shown on top of each plot (*e.g.,* 36 cancer types were evaluated for
700 SBS1). For each signature, the number of cancer types where the mutation density increases with
701 replication timing is shown next to ↗ (*e.g.,* 23 cancer types for SBS1). Similarly, the number of
702 cancer types where the mutation density decreases with replication timing is shown next to ↘
703 (*e.g.,* 0 cancer types for SBS1). Lastly, the number of cancer types where the mutation density is
704 not affected by replication timing is shown next to → (*e.g.,* 13 cancer types for SBS1).
705
706 **Figure S3. The effect of nucleosome occupancy on mutational signatures. Related to Figure**
707 **3.** Top three panels reflect results for all single base substitutions (SBSs), all doublet-base
708 substitutions (DBSs), and all small insertions/deletions (IDs) across all examined cancer types

709 with each cancer type examined separately. Bottom panels reflect all somatic mutations
710 attributed to a particular signature across all cancer types. In all cases, solid lines correspond to
711 real somatic mutations with blue solid lines reflecting SBSs, red solid lines reflecting DBSs, and
712 green solid lines reflecting IDs. Solid lines and dashed lines display the average nucleosome
713 signal (y-axes) along a 2 kilobase window (x-axes) centred at the mutation start site for real and
714 simulated mutations, respectively. The mutation start site is annotated in the middle of each plot
715 and denoted as 0. The 2 kilobase window encompasses 1,000 base-pairs 5' adjacent to each
716 mutation as well as 1,000 base-pairs 3' adjacent to each mutation. For each mutational
717 signatures, the total number of similar and considered cancer types using an *X/Y* format, with *X*
718 being the number of cancer types where a signature has similar nucleosome behaviour (Pearson
719 correlation ≥ 0.5 and q-value ≤ 0.05) and *Y* representing the total number of examined cancer
720 types for that signature. For example, signature SBS3 annotated with 11/14 reflects a total of 14
721 examined cancer types with similar nucleosome behaviour observed in 11 of these 14 cancer
722 types.
723
724 **Figure S4. Relationships between mutational signatures and histone modifications. Related**
725 **to Figure 5.** ***(A-C)*** Relationships between 11 histone modifications and signatures of single base
726 substitutions (SBSs) in panel *(A)*, doublet-base substitutions (DBSs) in panel *(B)*, and small
727 insertions/deletions (IDs) in panel *(C)*. The examined histone modifications encompass H2AFZ,
728 H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, H3K27ac, H3K27me3, H3K36me3,
729 H3K79me2, and H4K20me1. Rows and columns reflect the mutational signatures and histone
730 modifications, respectively. The circle in each cell is separated in red, blue, and grey segments
731 proportionate to the cancer types in which the signature has a specific behaviour. A red segment

732 in a circle reflects the signature being enriched in the vicinity of a histone modification (q-value
733 ≤ 0.05 and at least 5% enrichment). A blue segment in a circle reflects the signature being
734 depleted in the vicinity of a histone modification (q-value ≤ 0.05 and at least 5% depletion). A
735 grey segment in a circle corresponds to neither depletion nor enrichment of the signature in the
736 vicinity of a histone modification. Cells without a circle correspond to insufficient data to
737 perform any statistical comparisons. ***(D)*** Exemplars of enrichment, depletions, or no effect for
738 several histone modifications and signatures SBS7a and ID1. Solid lines and dashed lines display
739 the average signal for a particular histone modification (y-axes) along a 2 kilobase window (x-
740 axes) centred at the mutation start site for real and simulated mutations, respectively. The
741 mutation start site is annotated in the middle of each plot and denoted as 0. The 2 kilobase
742 window encompasses 1,000 base-pairs 5' adjacent to each mutation as well as 1,000 base-pairs
743 3' adjacent to each mutation.
744
745 **Figure S5. Strand asymmetries of non-clustered, *omikli*, and *kataegis* substitutions across**
746 **288 whole-genome sequenced B-cell malignancies. Related to Figure 7.** Transcription strand
747 asymmetries are shown in the left panels where bars display the six substitution subtypes based
748 on the mutated pyrimidine base: C>A, C>G, C>T, T>A, T>C, and T>G (depicted on the x-axes).
749 Y-axes correspond to the numbers of single base substitutions. Blue bars reflect real transcribed
750 substitutions, while shaded blue bars correspond to simulated transcribed substitutions.
751 Similarly, green bars reflect real untranscribed mutations, whereas shaded green bars correspond
752 to simulated untranscribed substitutions. Replication strand asymmetries are shown in the middle
753 panels where bars display the six substitution subtypes based on the mutated pyrimidine base:
754 C>A, C>G, C>T, T>A, T>C, and T>G (depicted on the x-axes). Y-axes correspond to the

755 numbers of single base substitutions. Red bars reflect real substitutions on the lagging strand,
756 while shaded red bars correspond to simulated substitutions on the lagging strand. Similarly,
757 yellow bars reflect real substitutions on the leading strand, whereas shaded yellow bars
758 correspond to simulated substitutions on the leading strand. Comparisons of genic and intergenic
759 regions are shown in the right panels where bars display the six substitution subtypes based on
760 the mutated pyrimidine base: C>A, C>G, C>T, T>A, T>C, and T>G (depicted on the x-axes). Y-
761 axes correspond to the numbers of single base substitutions. Cyan bars reflect real substitutions
762 in genic regions, while shaded cyan bars correspond to simulated substitutions in genic regions.
763 Similarly, grey bars reflect real substitutions in intergenic regions, whereas shaded grey bars
764 correspond to simulated substitutions in intergenic regions. Results for non-clustered mutations
765 are shown in panel ***(A)***, *omikli* mutations in panel ***(B)***, and *kataegis* mutations in panel ***(C)***.
766 Statistically significant strand asymmetries are shown with stars: * q-value ≤ 0.05; ** q-value ≤
767 0.01; *** q-value ≤ 0.001.
768

769 **SUPPLEMENTARY TABLES**

770 **Supplementary Table 1: External Datasets Utilized in Performing the Topography Analyses.**
771 The table provides information on the ENCODE (https://www.encodeproject.org/) or GEO
772 (https://www.ncbi.nlm.nih.gov/gds) datasets used for examining every topography feature within
773 each evaluated cancer type.

ACKNOWLEDGEMENTS

The authors would like to thank the COSMIC team for assistance in developing and deploying the COSMIC signatures topography database. BO and LBA would like to thank Prof. Steven Rozen (Duke-NUS) and Prof. Michael Stratton (Sanger Institute) for the many useful discussion in regard to examining the topography of mutational signatures as well as Mariya Kazachkova (UC San Diego) for her feedback and comments on improving the readability of the methods. This was work was funded by Cancer Research UK Grand Challenge Award [C98/A24032] as well as US National Institute of Health grants R01ES030993, R01ES032547, and R01CA269919. Work at the Wellcome Sanger Institute was also supported by the Wellcome Trust [grant number 108413/A/15/D]. LBA is also supported by a Packard Fellowship for Science and Engineering. The funders had no roles in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

DECLARATION OF INTERESTS

LBA is a compensated consultant and has equity interest in io9, LLC and Genome Insight. His spouse is an employee of Biotheranostics, Inc. LBA is also an inventor of a US Patent 10,776,718 for source identification by non-negative matrix factorization. ENB and LBA declare U.S. provisional applications with serial numbers: 63/289,601; 63/269,033; and 63/483,237. LBA also declares U.S. provisional applications with serial numbers: 63/366,392; 63/367,846; 63/412,835; and 63/492,348. All other authors declare they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

797 **AUTHOR CONTRIBUTIONS**

798 BO and LBA conceived the performed computational analyses and wrote the manuscript with
799 assistance from MDG. BO developed the Python code and performed the bioinformatics
800 analyses with assistance from MDG, ENB, MZ, and MB. The online COSMIC signatures
801 topography database was designed by BO, IV, and LBA with assistance from MDG, MZ, and
802 MB. The COSMIC signatures topography database was implemented by IV with feedback from
803 all authors. LBA supervised the overall development of the code, website, analysis, and writing
804 of the manuscript. All authors read and approved the final manuscript.

805

806 **REFERENCES**

807 1. Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*,
808 719-724. 10.1038/nature07943.
809 2. Martincorena, I., and Campbell, P.J. (2015). Somatic mutation in cancer and normal cells.
810 Science *349*, 1483-1489. 10.1126/science.aab4082.
811 3. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin,
812 A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L., et al. (2013). Signatures of
813 mutational processes in human cancer. Nature *500*, 415-421. 10.1038/nature12477.
814 4. Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot,
815 A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N., et al. (2020). The repertoire of
816 mutational signatures in human cancer. Nature *578*, 94-101. 10.1038/s41586-020-1943-3.
817 5. Islam, S.M.A., Díaz-Gay, M., Wu, Y., Barnes, M., Vangara, R., Bergstrom, E.N., He, Y.,
818 Vella, M., Wang, J., Teague, J.W., et al. (2022). Uncovering novel mutational signatures
819 by *de novo* extraction with SigProfilerExtractor. bioRxiv, 2020.2012.2013.422570.
820 10.1101/2020.12.13.422570.
821 6. Zhang, T., Joubert, P., Ansari-Pour, N., Zhao, W., Hoang, P.H., Lokanga, R., Moye,
822 A.L., Rosenbaum, J., Gonzalez-Perez, A., Martinez-Jimenez, F., et al. (2021). Genomic
823 and evolutionary classification of lung cancer in never smokers. Nat Genet *53*, 1348-
824 1359. 10.1038/s41588-021-00920-0.
825 7. Moody, S., Senkin, S., Islam, S.M.A., Wang, J., Nasrollahzadeh, D., Penha, R.C.C.,
826 Fitzgerald, S., Bergstrom, E.N., Atkins, J., He, Y., et al. (2021). Mutational signatures in
827 esophageal squamous cell carcinoma from eight countries of varying incidence.
828 medRxiv, 2021.2004.2029.21255920. 10.1101/2021.04.29.21255920.
829 8. Alexandrov, L.B., Ju, Y.S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S.,
830 Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T., et al. (2016). Mutational signatures
831 associated with tobacco smoking in human cancer. Science *354*, 618-622.
832 10.1126/science.aag0299.
833 9. Petljak, M., Alexandrov, L.B., Brammeld, J.S., Price, S., Wedge, D.C., Grossmann, S.,
834 Dawson, K.J., Ju, Y.S., Iorio, F., Tubio, J.M.C., et al. (2019). Characterizing Mutational
835 Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. Cell
836 *176*, 1282-1294 e1220. 10.1016/j.cell.2019.02.012.
837 10. Zhivagui, M., Ng, A.W.T., Ardin, M., Churchwell, M.I., Pandey, M., Renard, C., Villar,
838 S., Cahais, V., Robitaille, A., Bouaoun, L., et al. (2019). Experimental and pan-cancer
839 genome analyses reveal widespread contribution of acrylamide exposure to
840 carcinogenesis in humans. Genome Res *29*, 521-531. 10.1101/gr.242453.118.
841 11. Phillips, D.H. (2018). Mutational spectra and mutational signatures: Insights into cancer
842 aetiology and mechanisms of DNA damage and repair. DNA Repair (Amst) *71*, 6-11.
843 10.1016/j.dnarep.2018.08.003.
844 12. Kucab, J.E., Zou, X., Morganella, S., Joel, M., Nanda, A.S., Nagy, E., Gomez, C.,
845 Degasperi, A., Harris, R., Jackson, S.P., et al. (2019). A Compendium of Mutational
846 Signatures of Environmental Agents. Cell *177*, 821-836 e816.
847 10.1016/j.cell.2019.03.001.
848 13. Huang, M.N., Yu, W., Teoh, W.W., Ardin, M., Jusakul, A., Ng, A.W.T., Boot, A.,
849 Abedi-Ardekani, B., Villar, S., Myint, S.S., et al. (2017). Genome-scale mutational
850 signatures of aflatoxin in cells, mice, and human tumors. Genome Res *27*, 1475-1486.
851 10.1101/gr.220038.116.

852 14. Jager, M., Blokzijl, F., Kuijk, E., Bertl, J., Vougioukalaki, M., Janssen, R., Besselink, N.,
853 Boymans, S., de Ligt, J., Pedersen, J.S., et al. (2019). Deficiency of nucleotide excision
854 repair is associated with mutational signature observed in cancer. Genome Res *29*, 1067-
855 1077. 10.1101/gr.246223.118.
856 15. Meier, B., Volkova, N.V., Hong, Y., Schofield, P., Campbell, P.J., Gerstung, M., and
857 Gartner, A. (2018). Mutational signatures of DNA mismatch repair deficiency in C.
858 elegans and human cancers. Genome Res *28*, 666-675. 10.1101/gr.226845.117.
859 16. Schuster-Bockler, B., and Lehner, B. (2012). Chromatin organization is a major influence
860 on regional mutation rates in human cancer cells. Nature *488*, 504-507.
861 10.1038/nature11273.
862 17. Polak, P., Karlic, R., Koren, A., Thurman, R., Sandstrom, R., Lawrence, M., Reynolds,
863 A., Rynes, E., Vlahovicek, K., Stamatoyannopoulos, J.A., and Sunyaev, S.R. (2015).
864 Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature
865 *518*, 360-364. 10.1038/nature14221.
866 18. Stamatoyannopoulos, J.A., Adzhubei, I., Thurman, R.E., Kryukov, G.V., Mirkin, S.M.,
867 and Sunyaev, S.R. (2009). Human mutation rate associated with DNA replication timing.
868 Nat Genet *41*, 393-395. 10.1038/ng.363.
869 19. Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A.,
870 Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational
871 heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214-
872 218. 10.1038/nature12213.
873 20. Supek, F., and Lehner, B. (2015). Differential DNA mismatch repair underlies mutation
874 rate variation across the human genome. Nature *521*, 81-84. 10.1038/nature14173.
875 21. Morganella, S., Alexandrov, L.B., Glodzik, D., Zou, X., Davies, H., Staaf, J., Sieuwerts,
876 A.M., Brinkman, A.B., Martin, S., Ramakrishna, M., et al. (2016). The topography of
877 mutational processes in breast cancer genomes. Nat Commun *7*, 11383.
878 10.1038/ncomms11383.
879 22. Haradhvala, N.J., Polak, P., Stojanov, P., Covington, K.R., Shinbrot, E., Hess, J.M.,
880 Rheinbay, E., Kim, J., Maruvka, Y.E., Braunstein, L.Z., et al. (2016). Mutational Strand
881 Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. Cell
882 *164*, 538-549. 10.1016/j.cell.2015.12.050.
883 23. Tomkova, M., Tomek, J., Kriaucionis, S., and Schuster-Bockler, B. (2018). Mutational
884 signature distribution varies with DNA replication timing and strand asymmetry. Genome
885 Biol *19*, 129. 10.1186/s13059-018-1509-y.
886 24. Pich, O., Muinos, F., Sabarinathan, R., Reyes-Salazar, I., Gonzalez-Perez, A., and Lopez-
887 Bigas, N. (2018). Somatic and Germline Mutation Periodicity Follow the Orientation of
888 the DNA Minor Groove around Nucleosomes. Cell *175*, 1074-1087 e1018.
889 10.1016/j.cell.2018.10.004.
890 25. Akdemir, K.C., Le, V.T., Kim, J.M., Killcoyne, S., King, D.A., Lin, Y.P., Tian, Y.,
891 Inoue, A., Amin, S.B., Robinson, F.S., et al. (2020). Somatic mutation distributions in
892 cancer genomes vary with three-dimensional chromatin structure. Nat Genet *52*, 1178-
893 1188. 10.1038/s41588-020-0708-0.
894 26. Vohringer, H., Hoeck, A.V., Cuppen, E., and Gerstung, M. (2021). Learning mutational
895 signatures and their multidimensional genomic properties with TensorSignatures. Nat
896 Commun *12*, 3628. 10.1038/s41467-021-23551-9.

897 27. Consortium, E.P., Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras,
898 T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., et al. (2007).
899 Identification and analysis of functional elements in 1% of the human genome by the
900 ENCODE pilot project. Nature *447*, 799-816. 10.1038/nature05874.
901 28. Sancar, A. (2016). Mechanisms of DNA Repair by Photolyase and Excision Nuclease
902 (Nobel Lecture). Angew Chem Int Ed Engl *55*, 8502-8527. 10.1002/anie.201601524.
903 29. Li, X.C., Wang, M.Y., Yang, M., Dai, H.J., Zhang, B.F., Wang, W., Chu, X.L., Wang,
904 X., Zheng, H., Niu, R.F., et al. (2018). A mutational signature associated with alcohol
905 consumption and prognostically significantly mutated driver genes in esophageal
906 squamous cell carcinoma. Ann Oncol *29*, 938-944. 10.1093/annonc/mdy011.
907 30. Letouze, E., Shinde, J., Renault, V., Couchy, G., Blanc, J.F., Tubacher, E., Bayard, Q.,
908 Bacq, D., Meyer, V., Semhoun, J., et al. (2017). Mutational signatures reveal the dynamic
909 interplay of risk factors and cellular processes during liver tumorigenesis. Nat Commun
910 *8*, 1315. 10.1038/s41467-017-01358-x.
911 31. Chang, J., Tan, W., Ling, Z., Xi, R., Shao, M., Chen, M., Luo, Y., Zhao, Y., Liu, Y.,
912 Huang, X., et al. (2017). Genomic analysis of oesophageal squamous-cell carcinoma
913 identifies alcohol drinking-related mutation signature and genomic alterations. Nat
914 Commun *8*, 15290. 10.1038/ncomms15290.
915 32. Petljak, M., Dananberg, A., Chu, K., Bergstrom, E.N., Striepen, J., von Morgen, P.,
916 Chen, Y., Shah, H., Sale, J.E., Alexandrov, L.B., et al. (2022). Mechanisms of APOBEC3
917 mutagenesis in human cancer cells. Nature *607*, 799-807. 10.1038/s41586-022-04972-y.
918 33. Saini, N., and Gordenin, D.A. (2020). Hypermutation in single-stranded DNA. DNA
919 Repair (Amst) *91-92*, 102868. 10.1016/j.dnarep.2020.102868.
920 34. Roberts, S.A., Sterling, J., Thompson, C., Harris, S., Mav, D., Shah, R., Klimczak, L.J.,
921 Kryukov, G.V., Malc, E., Mieczkowski, P.A., et al. (2012). Clustered mutations in yeast
922 and in human cancers can arise from damaged long single-strand DNA regions. Mol Cell
923 *46*, 424-435. 10.1016/j.molcel.2012.03.030.
924 35. Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K.,
925 Jones, D., Hinton, J., Marshall, J., Stebbings, L.A., et al. (2012). Mutational processes
926 molding the genomes of 21 breast cancers. Cell *149*, 979-993.
927 10.1016/j.cell.2012.04.024.
928 36. Koren, A., Polak, P., Nemesh, J., Michaelson, J.J., Sebat, J., Sunyaev, S.R., and
929 McCarroll, S.A. (2012). Differential relationship of DNA replication timing to different
930 forms of human mutation and variation. Am J Hum Genet *91*, 1033-1040.
931 10.1016/j.ajhg.2012.10.018.
932 37. Chen, C.L., Rappailles, A., Duquenne, L., Huvet, M., Guilbaud, G., Farinelli, L., Audit,
933 B., d'Aubenton-Carafa, Y., Arneodo, A., Hyrien, O., and Thermes, C. (2010). Impact of
934 replication timing on non-CpG and CpG substitution rates in mammalian genomes.
935 Genome Res *20*, 447-457. 10.1101/gr.098947.109.
936 38. Watanabe, Y., Fujiyama, A., Ichiba, Y., Hattori, M., Yada, T., Sakaki, Y., and Ikemura,
937 T. (2002). Chromosome-wide assessment of replication timing for human chromosomes
938 11q and 21q: disease-related genes in timing-switch regions. Hum Mol Genet *11*, 13-21.
939 10.1093/hmg/11.1.13.
940 39. Petljak, M., and Maciejowski, J. (2020). Molecular origins of APOBEC-associated
941 mutations in cancer. DNA Repair (Amst) *94*, 102905. 10.1016/j.dnarep.2020.102905.

942 40. Helleday, T., Eshtad, S., and Nik-Zainal, S. (2014). Mechanisms underlying mutational
943 signatures in human cancers. Nat Rev Genet *15*, 585-598. 10.1038/nrg3729.
944 41. Hwang, T., Reh, S., Dunbayev, Y., Zhong, Y., Takata, Y., Shen, J., McBride, K.M.,
945 Murnane, J.P., Bhak, J., Lee, S., et al. (2020). Defining the mutation signatures of DNA
946 polymerase theta in cancer genomes. NAR Cancer *2*, zcaa017. 10.1093/narcan/zcaa017.
947 42. Davey, C.A., Sargent, D.F., Luger, K., Maeder, A.W., and Richmond, T.J. (2002).
948 Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a
949 resolution. J Mol Biol *319*, 1097-1113. 10.1016/S0022-2836(02)00386-8.
950 43. Richmond, T.J., and Davey, C.A. (2003). The structure of DNA in the nucleosome core.
951 Nature *423*, 145-150. 10.1038/nature01595.
952 44. Christensen, S., Van der Roest, B., Besselink, N., Janssen, R., Boymans, S., Martens,
953 J.W.M., Yaspo, M.L., Priestley, P., Kuijk, E., Cuppen, E., and Van Hoeck, A. (2019). 5-
954 Fluorouracil treatment induces characteristic T>G mutations in human cancer. Nat
955 Commun *10*, 4571. 10.1038/s41467-019-12594-8.
956 45. Hidaka, K., Yamada, M., Kamiya, H., Masutani, C., Harashima, H., Hanaoka, F., and
957 Nohmi, T. (2008). Specificity of mutations induced by incorporation of oxidized dNTPs
958 into DNA by human DNA polymerase eta. DNA Repair (Amst) *7*, 497-506.
959 10.1016/j.dnarep.2007.12.005.
960 46. Inoue, M., Kamiya, H., Fujikawa, K., Ootsuyama, Y., Murata-Kamiya, N., Osaki, T.,
961 Yasumoto, K., and Kasai, H. (1998). Induction of chromosomal gene mutations in
962 Escherichia coli by direct incorporation of oxidatively damaged nucleotides. New
963 evaluation method for mutagenesis by damaged DNA precursors in vivo. J Biol Chem
964 *273*, 11069-11074. 10.1074/jbc.273.18.11069.
965 47. Dvorak, K., Payne, C.M., Chavarria, M., Ramsey, L., Dvorakova, B., Bernstein, H.,
966 Holubec, H., Sampliner, R.E., Guy, N., Condon, A., et al. (2007). Bile acids in
967 combination with low pH induce oxidative stress and oxidative DNA damage: relevance
968 to the pathogenesis of Barrett's oesophagus. Gut *56*, 763-771. 10.1136/gut.2006.103697.
969 48. Focaccetti, C., Bruno, A., Magnani, E., Bartolini, D., Principi, E., Dallaglio, K., Bucci,
970 E.O., Finzi, G., Sessa, F., Noonan, D.M., and Albini, A. (2015). Effects of 5-fluorouracil
971 on morphology, cell cycle, proliferation, apoptosis, autophagy and ROS production in
972 endothelial cells and cardiomyocytes. PLoS One *10*, e0115686.
973 10.1371/journal.pone.0115686.
974 49. Ong, C.T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome
975 topology and function. Nat Rev Genet *15*, 234-246. 10.1038/nrg3663.
976 50. Poulos, R.C., Thoms, J.A.I., Guan, Y.F., Unnikrishnan, A., Pimanda, J.E., and Wong,
977 J.W.H. (2016). Functional Mutations Form at CTCF-Cohesin Binding Sites in Melanoma
978 Due to Uneven Nucleotide Excision Repair across the Motif. Cell Rep *17*, 2865-2872.
979 10.1016/j.celrep.2016.11.055.
980 51. Kouzarides, T. (2007). Chromatin modifications and their function. Cell *128*, 693-705.
981 10.1016/j.cell.2007.02.005.
982 52. Giaimo, B.D., Ferrante, F., Herchenrother, A., Hake, S.B., and Borggrefe, T. (2019). The
983 histone variant H2A.Z in gene regulation. Epigenetics Chromatin *12*, 37.
984 10.1186/s13072-019-0274-9.
985 53. Kang, Y., Kim, Y.W., Kang, J., and Kim, A. (2021). Histone H3K4me1 and H3K27ac
986 play roles in nucleosome eviction and eRNA transcription, respectively, at enhancers.
987 FASEB J *35*, e21781. 10.1096/fj.202100488R.

988 54. Wang, Y., Li, X., and Hu, H. (2014). H3K4me2 reliably defines transcription factor
989 binding regions in different cells. Genomics *103*, 222-228. 10.1016/j.ygeno.2014.02.002.
990 55. Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and
991 why? Mol Cell *49*, 825-837. 10.1016/j.molcel.2013.01.038.
992 56. Gates, L.A., Shi, J., Rohira, A.D., Feng, Q., Zhu, B., Bedford, M.T., Sagum, C.A., Jung,
993 S.Y., Qin, J., Tsai, M.J., et al. (2017). Acetylation on histone H3 lysine 9 mediates a
994 switch from transcription initiation to elongation. J Biol Chem *292*, 14456-14472.
995 10.1074/jbc.M117.802074.
996 57. Saksouk, N., Simboeck, E., and Dejardin, J. (2015). Constitutive heterochromatin
997 formation and transcription in mammals. Epigenetics Chromatin *8*, 3. 10.1186/1756-
998 8935-8-3.
999 58. Cai, Y., Zhang, Y., Loh, Y.P., Tng, J.Q., Lim, M.C., Cao, Z., Raju, A., Lieberman Aiden,
1000 E., Li, S., Manikandan, L., et al. (2021). H3K27me3-rich genomic regions can function
1001 as silencers to repress gene expression via chromatin interactions. Nat Commun *12*, 719.
1002 10.1038/s41467-021-20940-y.
1003 59. Sun, Z., Zhang, Y., Jia, J., Fang, Y., Tang, Y., Wu, H., and Fang, D. (2020). H3K36me3,
1004 message from chromatin to DNA damage repair. Cell Biosci *10*, 9. 10.1186/s13578-020-
1005 0374-z.
1006 60. Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K.,
1007 Roh, T.Y., Peng, W., Zhang, M.Q., and Zhao, K. (2008). Combinatorial patterns of
1008 histone acetylations and methylations in the human genome. Nat Genet *40*, 897-903.
1009 10.1038/ng.154.
1010 61. Li, Z., Nie, F., Wang, S., and Li, L. (2011). Histone H4 Lys 20 monomethylation by
1011 histone methylase SET8 mediates Wnt target gene activation. Proc Natl Acad Sci U S A
1012 *108*, 3116-3123. 10.1073/pnas.1009353108.
1013 62. Mas-Ponte, D., and Supek, F. (2020). DNA mismatch repair promotes APOBEC3-
1014 mediated diffuse hypermutation in human cancers. Nat Genet *52*, 958-968.
1015 10.1038/s41588-020-0674-6.
1016 63. Bergstrom, E.N., Kundu, M., Tbeileh, N., and Alexandrov, L.B. (2022). Examining
1017 clustered somatic mutations with SigProfilerClusters. Bioinformatics.
1018 10.1093/bioinformatics/btac335.
1019 64. Bergstrom, E.N., Luebeck, J., Petljak, M., Khandekar, A., Barnes, M., Zhang, T., Steele,
1020 C.D., Pillay, N., Landi, M.T., Bafna, V., et al. (2022). Mapping clustered mutations in
1021 cancer reveals APOBEC3 mutagenesis of ecDNA. Nature *602*, 510-517.
1022 10.1038/s41586-022-04398-6.

1023

Figure 1. Strand asymmetries and strand-coordinated mutagenesis.

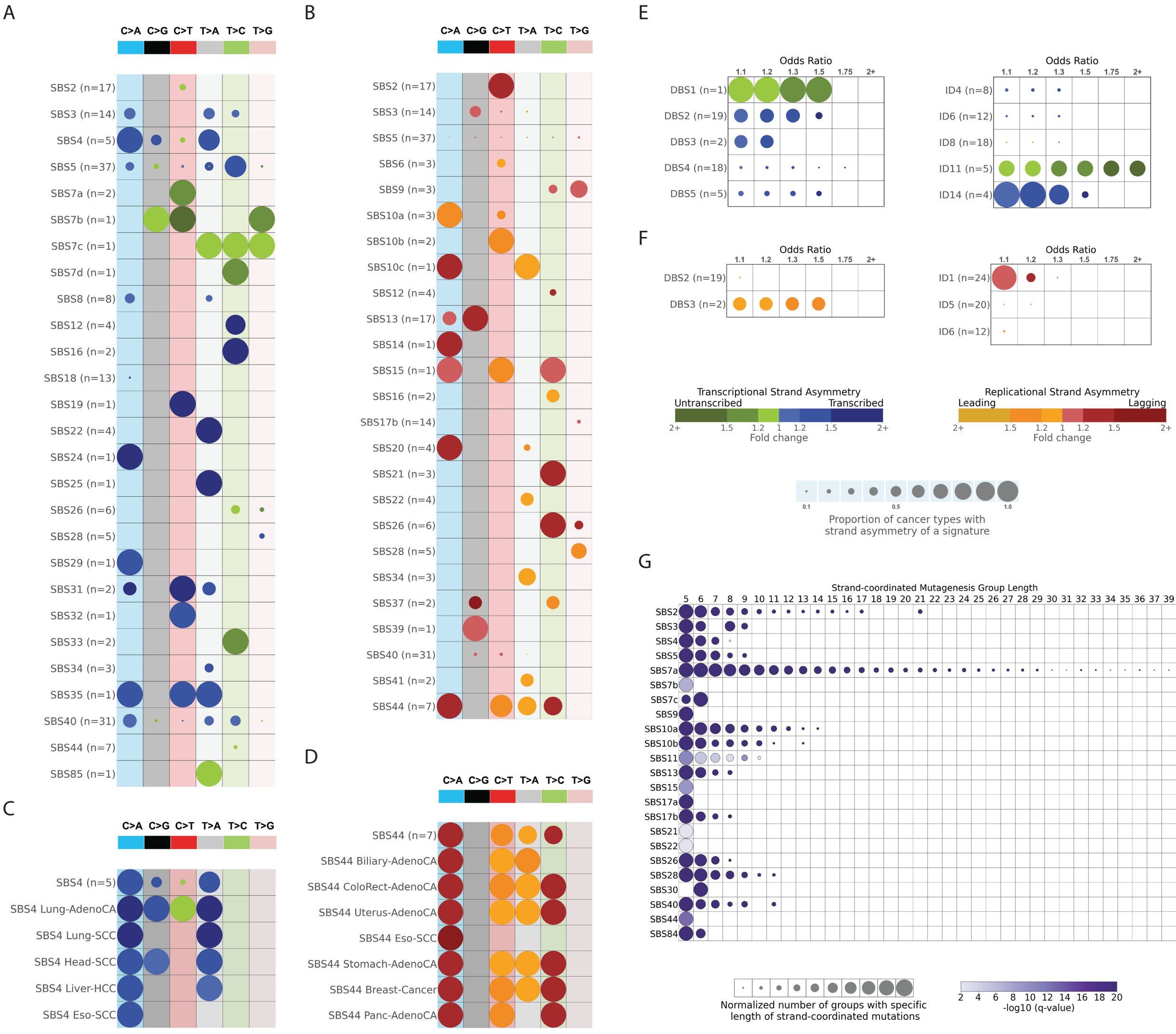


Figure 2. Interplay between replication timing and mutational signatures.

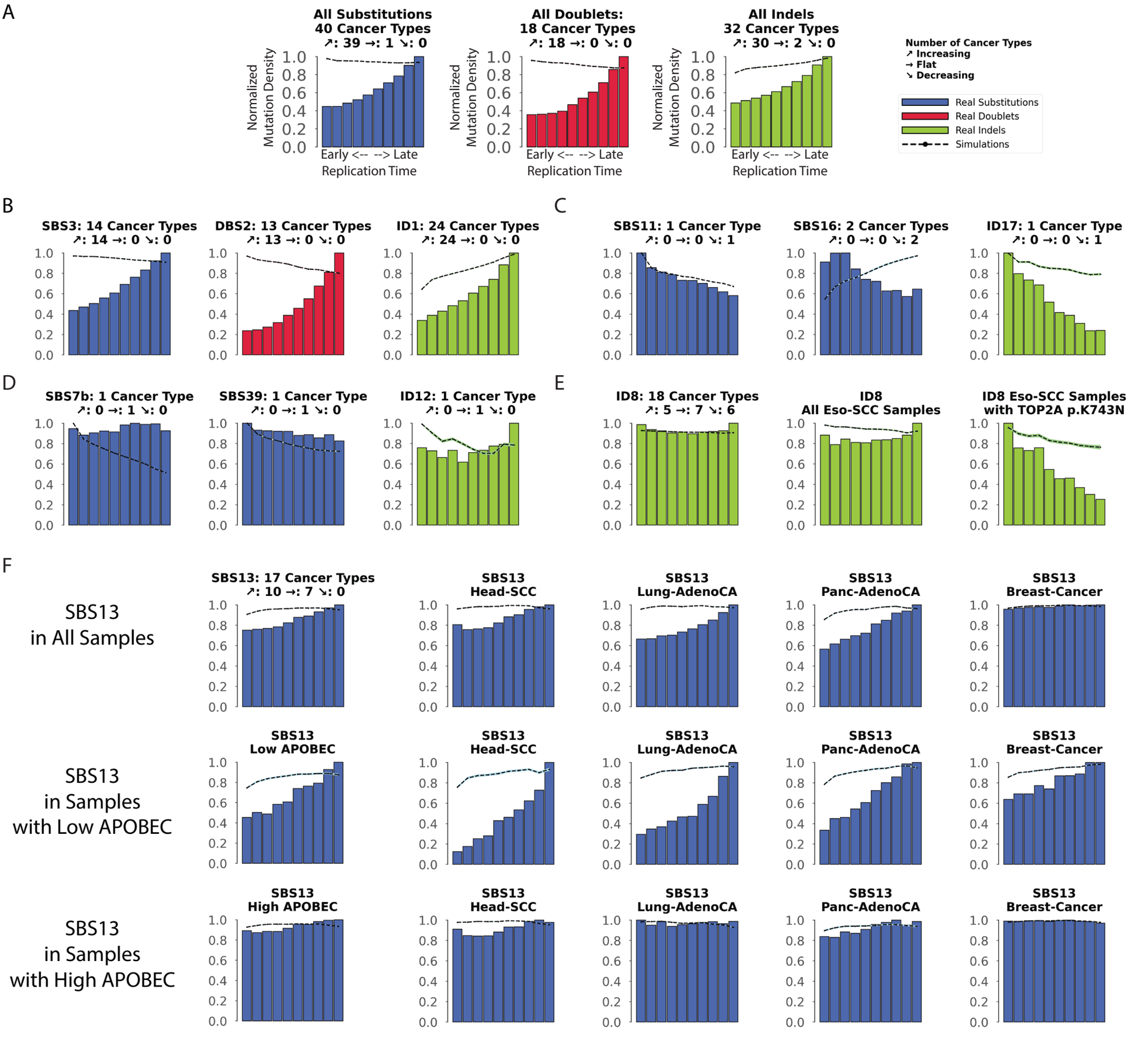


Figure 3. Relationship between mutational signatures and nucleosome occupancy.

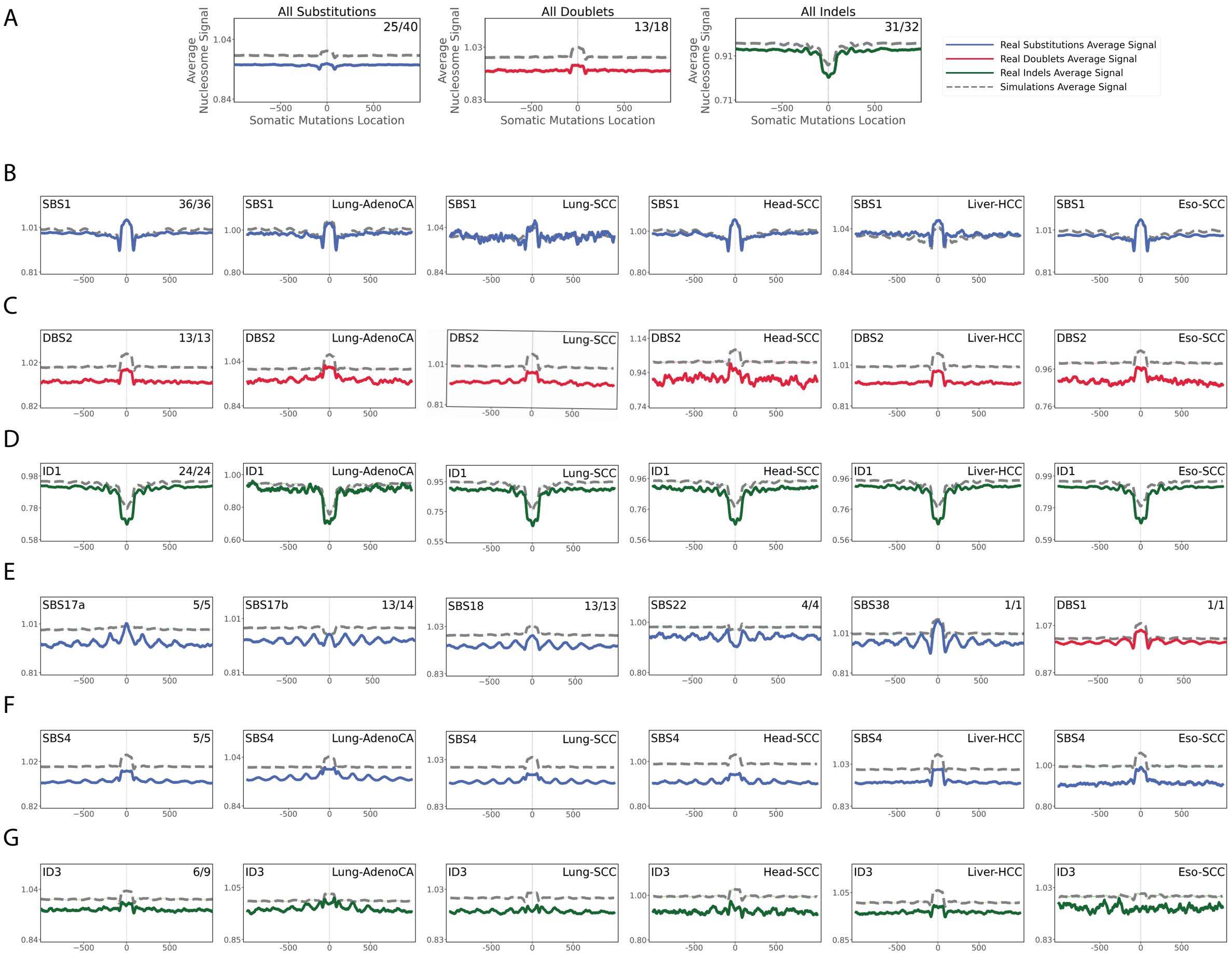


Figure 4. Relationship between mutational signatures and CTCF binding sites.

A

	Biliary-AdenoCA	Bone-Osteosarc	Breast-Cancer	CNS-GBM	CNS-Medullo	CNS-PiloAstro	ColoRect-AdenoCA	Eso-SCC	Eso-AdenoCA	Head-SCC	Kidney-ChRCC	Kidney-RCC	Liver-HCC	Lung-AdenoCA	Lymph-BNHL	Lymph-CLL	Ovary-AdenoCA	Panc-AdenoCA	Panc-Endocrine	Prost-AdenoCA	Skin-Melanoma	SoftTissue-Liposarc	Stomach-AdenoCA	Uterus-AdenoCA
All Substitutions	1.12*	1.06*	1.04*	0.96*	1.01	0.83*	0.91*	0.98*	1.19*	0.98	1.03	1.05*	1.24*	1.05*	0.83*	0.85*	1.02	0.96*	1.04*	0.97*	1.04*	1.07*	1.07*	0.95*
SBS1	0.65*	0.76*	0.89*	0.72*	0.79*		0.78*	0.82*	0.82*	0.87*	0.67*	0.84*		0.86*	0.87*	0.75*	0.92*	0.74*	0.90*	0.77*	0.89*	0.74*	0.77*	0.90*
SBS2			1.06*					1.01	0.92	1.02				1.05				1.10						1.02
SBS3			1.11*					1.03*	1.09								1.09*	1.05					1.23*	
SBS4								0.76*		0.94*			1.03	1.06*										
SBS5	1.18*	1.03	1.06*	0.99	1.04		0.93*	1.02*	1.13*	1.01		1.12*	1.17*	1.07*	0.90*	0.97		1.00*	1.07*	0.98	1.06*		1.07*	0.98
SBS7a																					1.04*			
SBS7b																					1.15*			
SBS7c																					1.03*			
SBS7d																					1.05*			
SBS9															0.98	0.80*								
SBS10a							0.91*																	0.89*
SBS10b							0.76*																	0.74*
SBS11				0.99																				
SBS12													1.48*											
SBS15								0.82*																
SBS17a									1.45*														1.43*	
SBS17b			1.41*					1.74*	1.38*				1.47		1.22*			1.32*					1.36*	
SBS18			0.98				1.00	1.06*	1.03									1.03					1.00	
SBS22												1.08	1.15*											
SBS37							0.82								0.56*									
SBS40	1.19*	1.08*		0.95	1.12*		1.04*	0.98	1.06*	0.98	1.08	1.06*	1.44*	1.01*	0.94*		0.99	0.99		1.01	1.08*	1.02	1.03	0.99
SBS44	1.15*		0.94*				0.97											0.86*					1.01	1.04
SBS84															0.51*									
SBS85															0.65*									
All Doublets								1.07					1.02	0.86						1.12	1.10*			
DBS1																					1.10*			
All Indels	1.07		1.10*	1.28*	0.90		0.99	1.07*	1.07*	1.12*		1.08	1.23*	1.11*	0.97		1.14*	1.06*	1.15	1.17*	1.08*		1.04*	1.05
ID2	1.02						0.97	1.07*	0.98						1.00			1.14*					1.02	1.05
ID3													0.97	1.10*										
ID5									1.26*			1.08	1.29*							1.08	1.31*			
ID6			1.08					1.17*									1.22*	0.92		1.23*				
ID8			1.19*					1.25				1.10	1.37*					1.05		1.16*				
ID9								1.15*																

* q <= 0.05

0.25 0.50 0.75 1.00 1.25 1.50 1.75

Fold Change
[Real mutations/Simulated Mutations]

B

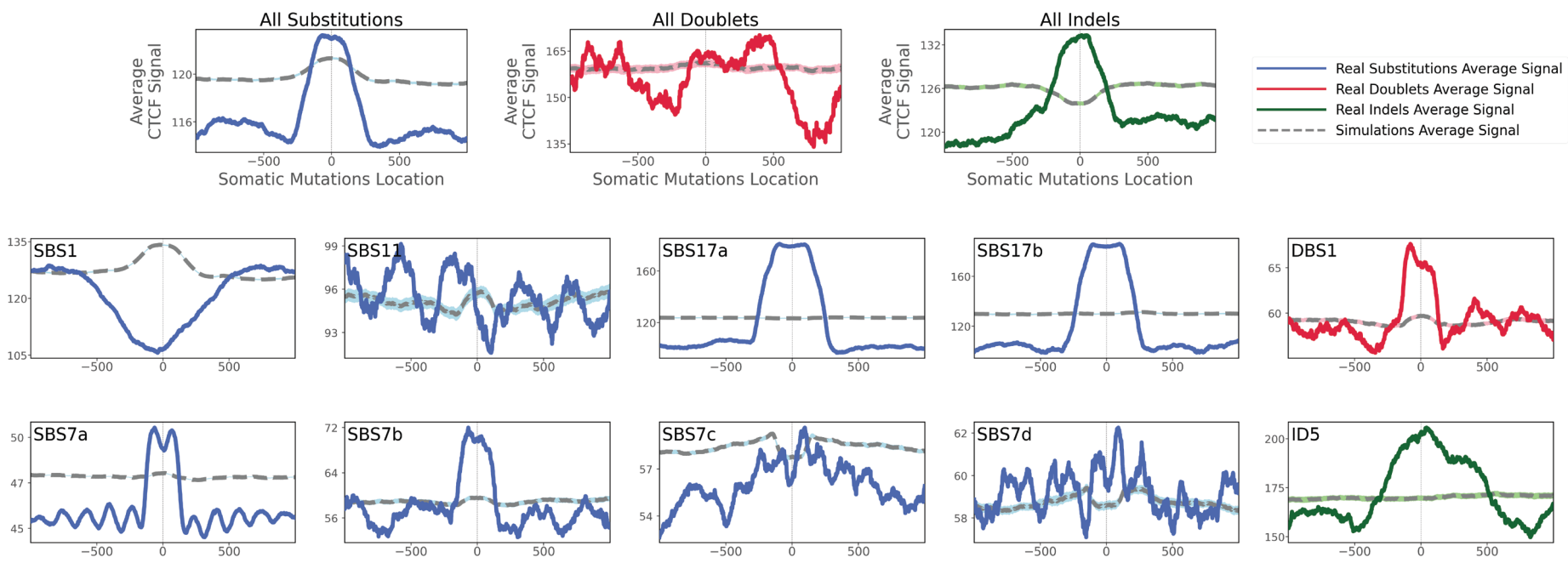


Figure 5. Exemplar relationships between mutational signatures and histone modifications.

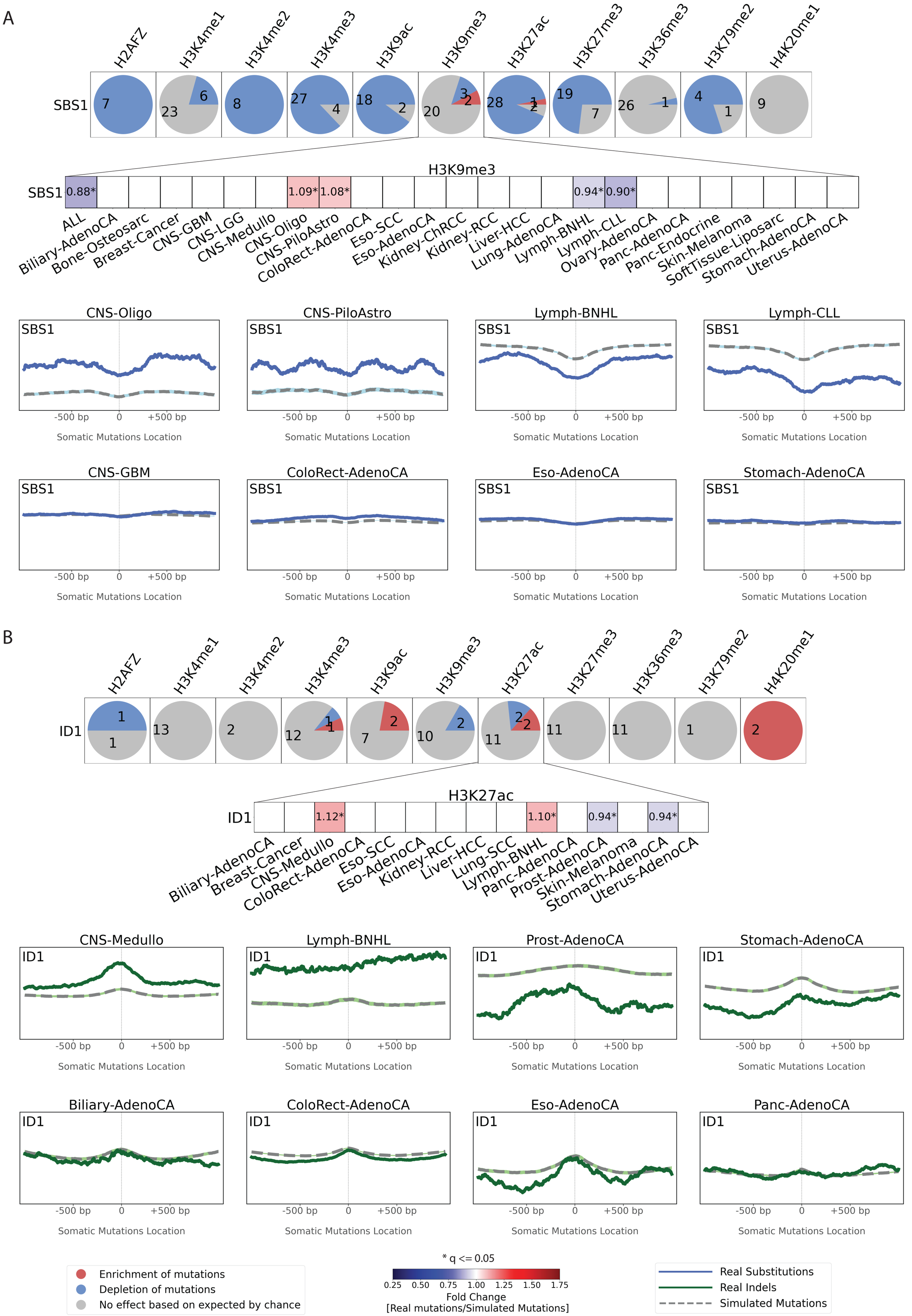


Figure 6. Topography of signature SBS28 in POLE deficient (POLE-) and POLE proficient (POLE+) samples.

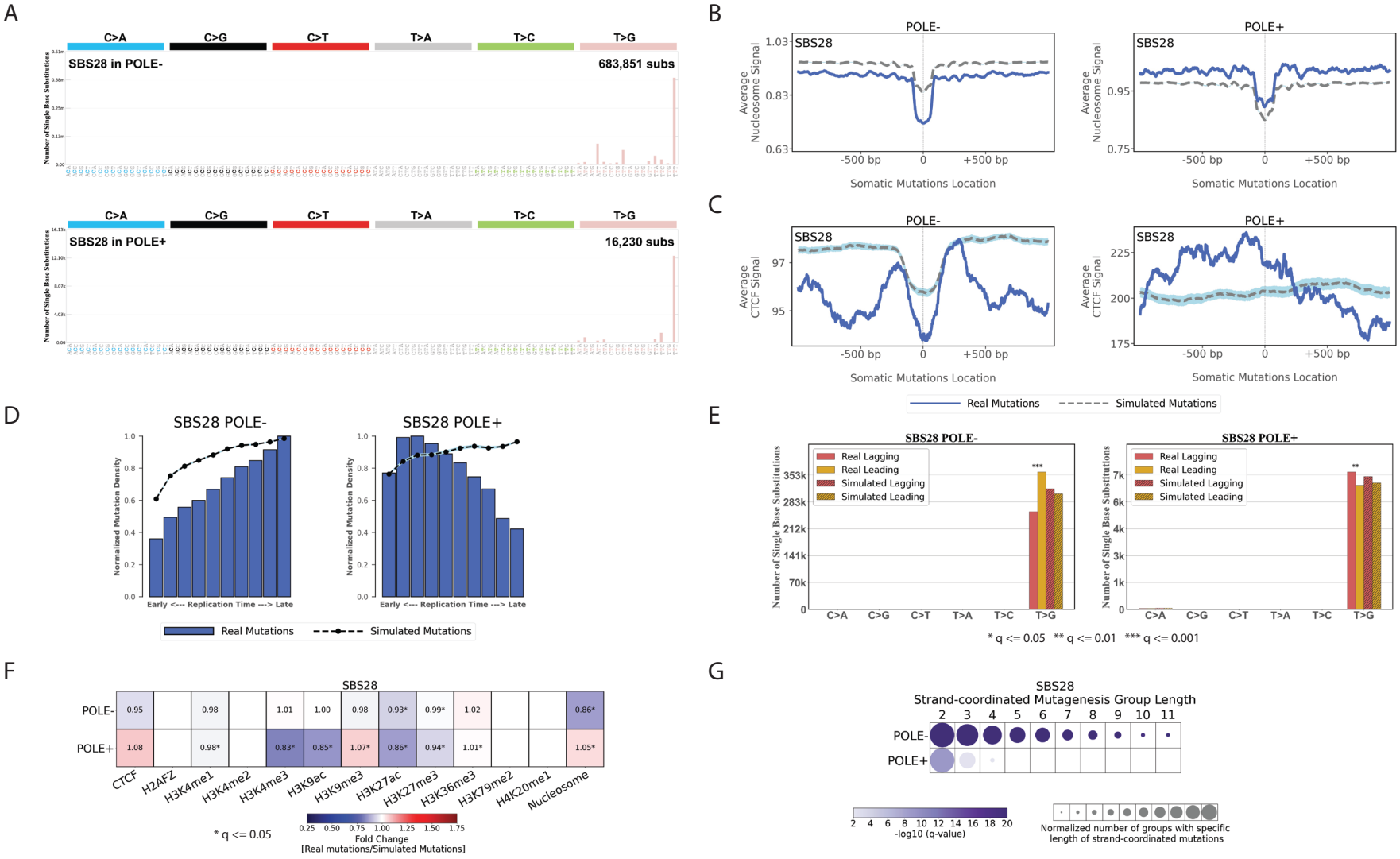
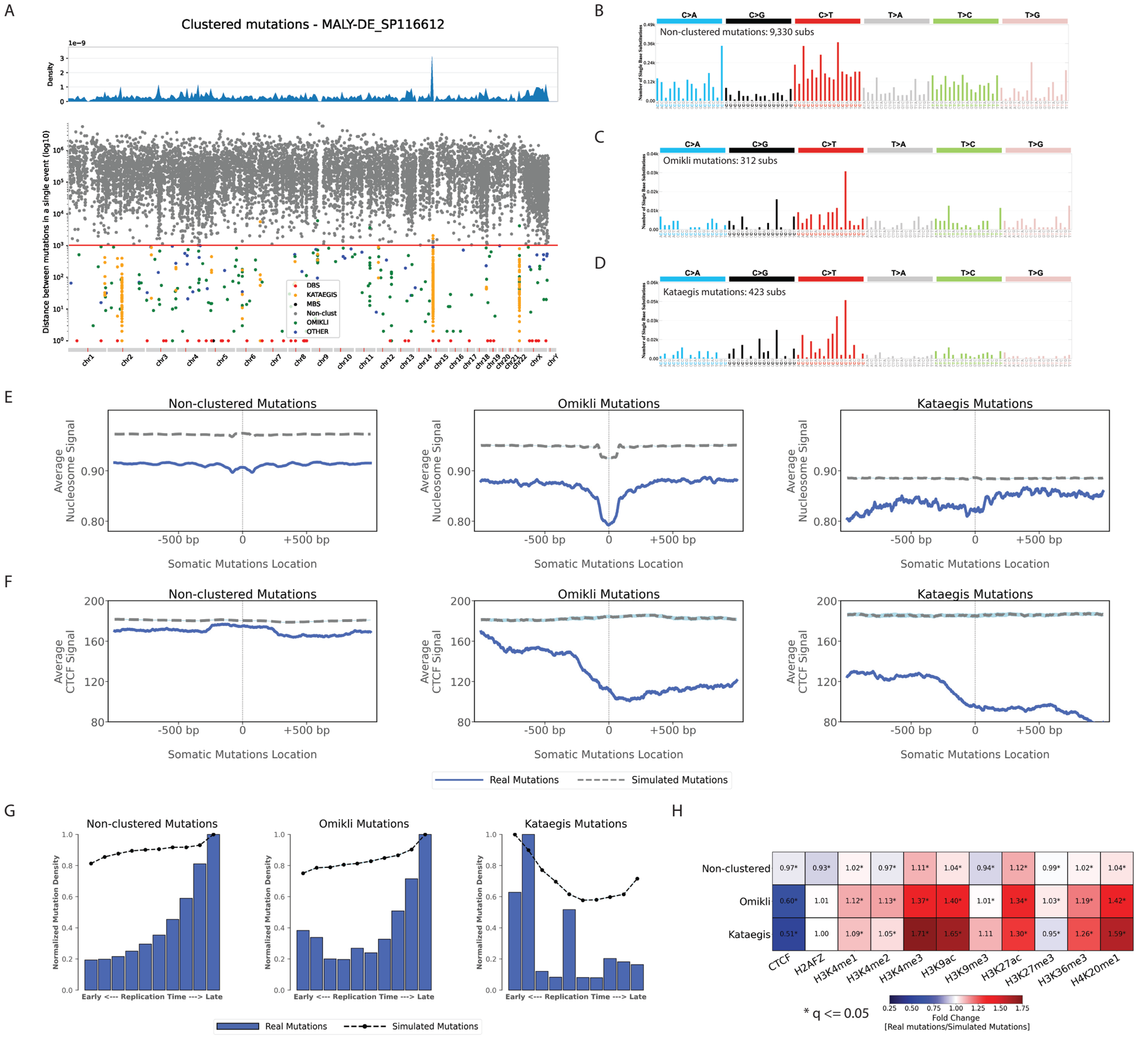



Figure 7. Topography of non-clustered, omikli, and kataegis substitutions across 288 whole-genome sequenced B-cell malignancies.