Topography of mutational signatures in human cancer

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

Affiliations

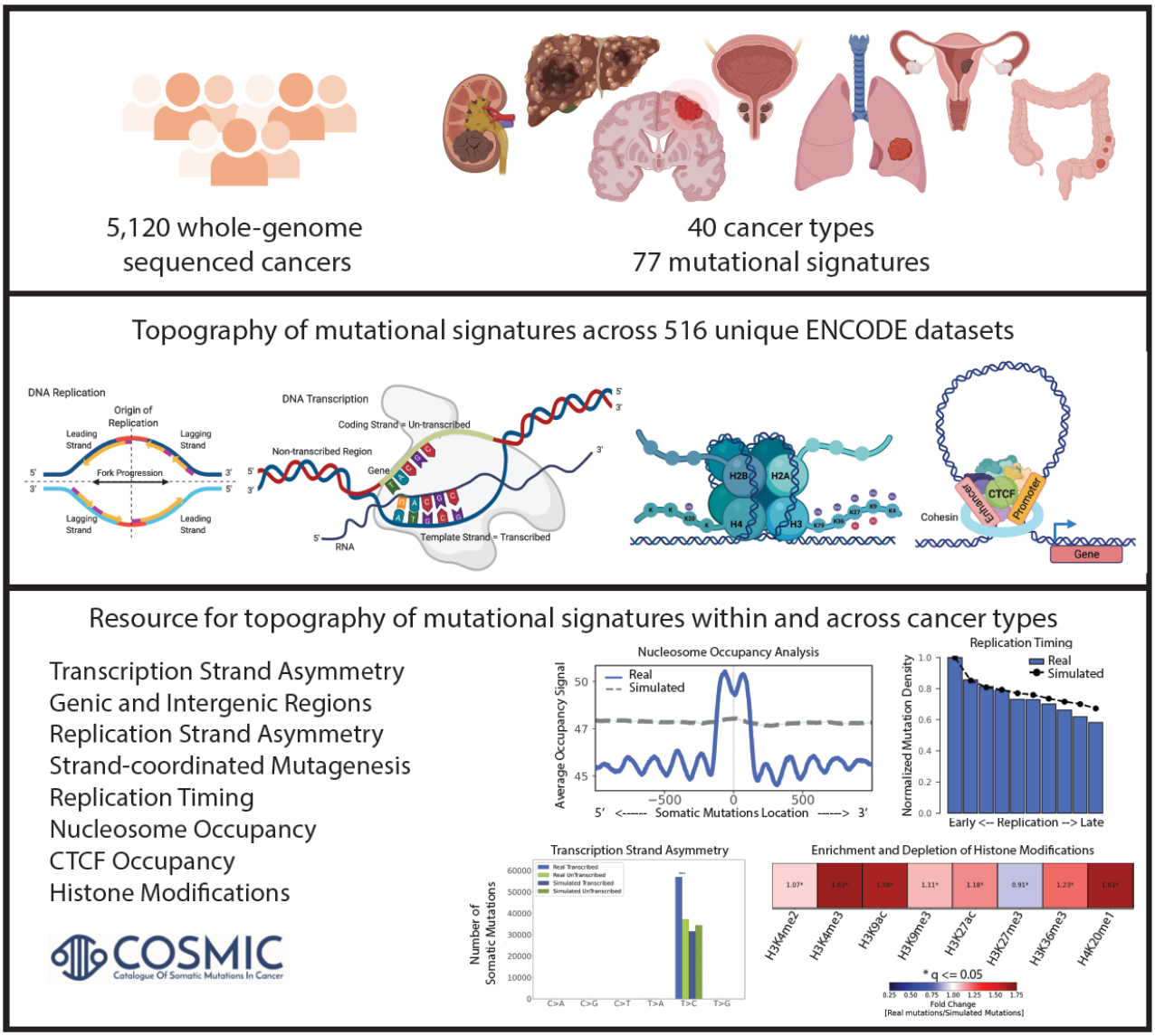
[1]Department of Cellular and Molecular Medicine, UC San Diego, La Jolla, CA, 92093, USA
[2]Department of Bioengineering, UC San Diego, La Jolla, CA, 92093, USA
[3]Moores Cancer Center, UC San Diego, La Jolla, CA, 92037, USA
[4]COSMIC, Wellcome Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA, UK

[*]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.

43 **GRAPHICAL ABSTRACT**



44

45

HIGHLIGHTS

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

- The accumulation of somatic mutations from most mutational signatures is affected by nucleosome occupancy, histone modifications, CTCF binding sites, transcribed regions, and replication strand/timing.

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

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

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

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

INTRODUCTION

Cancer genomes are peppered with somatic mutations imprinted by the activities of different endogenous and exogenous mutational processes (Martincorena and Campbell, 2015; Stratton et al., 2009). Due to their intrinsic biophysical and biochemical properties, each mutational process engraves a characteristic pattern of somatic mutations, known as a mutational signature (Alexandrov et al., 2013). 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 (Alexandrov et al., 2020; Islam et al., 2022; Moody et al., 2021; Zhang et al., 2021). Through statistical associations and further experimental characterizations, aetiology has been proposed for approximately half of the identified signatures (Alexandrov et al., 2016; Alexandrov et al., 2020; Huang et al., 2017; Jager et al., 2019; Kucab et al., 2019; Meier et al., 2018; Petljak et al., 2019; Phillips, 2018; Zhivagui et al., 2019). 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 (Lawrence et al., 2013; Polak et al., 2015; Schuster-Bockler and Lehner, 2012; Stamatoyannopoulos et al., 2009). Subsequent analyses of hundreds of cancer

genomes have revealed that differential DNA repair can explain variations in mutation rates across some cancer genomes (Supek and Lehner, 2015) 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 (Polak et al., 2015). Recently, Morganella *et al.* examined the effect of the genomic and the epigenomic architecture on the activity of 12 SBS signatures in breast cancer (Morganella et al., 2016). These analyses demonstrated that mutations generated by different mutational processes exhibit distinct strand asymmetries and that mutational signatures are differently affected by replication timing and nucleosome occupancy (Morganella et al., 2016). Pan-cancer exploration of strand asymmetries was also conducted for different mutation types across multiple cancer types (Haradhvala et al., 2016) as well as for different mutational signatures (Tomkova et al., 2018). In particular, pan-cancer analyses of more than 3,000 cancers have revealed the strand asymmetries and replication timings of the 30 SBS mutational signatures from COSMICv2 signatures database (Tomkova et al., 2018). Similarly, more than 3,000 cancer genomes were used to elucidate the mutation periodicity of the 30 SBS COSMIC signatures database v2 signatures in regard to nucleosome occupancy (Pich et al., 2018). More recently, a study has also shown the interplay between the three-dimensional genome organization and the activity of different mutational signatures (Akdemir et al., 2020; Vohringer et al., 2021).

Here, we report the most comprehensive evaluation of the effect of nucleosome occupancy, histone modifications, CCCTC-binding factor (CTCF) binding sites, replication timing, transcription strand asymmetry, and replication strand asymmetry on the cancer-specific accumulation of somatic mutations from distinct mutational signatures. Our analysis leverages

the complete set of known COSMICv2 mutational signatures (78 SBS, 11 DBS, and 18 ID) and it examines 5,120 whole-genome sequenced cancers while simultaneously utilizing 516 unique tissue-matched topographical features from the ENCODE project (**Table S1**) (Consortium et al., 2007). In all analyses, the observed patterns of somatic mutations are compared to background simulation models of mutational signatures that mimic both the trinucleotide pattern of these signatures as well as their mutational burden within each chromosome in each examined sample (**Methods**). Our results confirm many of the observations previously reported for strand asymmetry, replication timing, and nucleosome periodicity for the original 30 COSMICv2 SBS signatures. Further, the richer and larger dataset allowed us to elucidate novel biological findings for some of these 30 SBS signatures revealing previously unobserved pan-cancer and cancer-specific dependencies. Additionally, this report provides the first-ever map of the genome topography of indel, doublet-base, and another 24 substitution signatures in human cancer. Moreover, our study examines, for the first time, the tissue-specific effect of CTCF binding and 11 different histone modifications on the accumulation of somatic mutations from different mutational signatures. As part of the results, we provide a global view of the topography of mutational signatures across 5,120 whole-genome sequenced tumours from 40 types of human cancer. As part of the discussion, we zoom into two distinct case studies: *(i)* the topography of different types of clustered somatic mutations; and *(ii)* using the topography of mutational signatures to separate mutational signatures with similar patterns. Lastly, the reported results are released as part of the COSMIC signatures database, https://cancer.sanger.ac.uk/signatures, providing an unprecedented online resource for examining the topography of mutational signatures within and across human cancer types.

RESULTS

Transcription Strand Asymmetries of Mutational Signatures

Transcription strand asymmetries have been generally attributed to transcription-coupled nucleotide excision repair (TC-NER) since bulky adducts (*e.g.*, ones due to tobacco carcinogens) in actively transcribed regions of the genome will be preferentially repaired by TC-NER (Sancar, 2016). Additionally, transcription-coupled damage may also lead to transcription strand asymmetry due to one of the strands being preferentially damaged during transcription (Haradhvala et al., 2016).

Mutational signatures with similar aetiologies generally exhibited consistent patterns of transcription strand asymmetries. Specifically, signatures attributed to exogenous mutational processes, including ones due to environmental mutagens or chemotherapy, in most cases, showed transcription strand bias with mutations usually enriched on the transcribed strand (**Figure 1*A&C***). Mutational signatures due to tobacco smoking had strong transcription strand bias for C>A and T>A mutations (SBS4) as well as CC>AA mutations (DBS2) on the transcribed strand. Similarly, signature SBS29 (tobacco chewing) exhibited enrichment on the transcribed strand for C>A. SBS22 (aristolochic acid) had strong transcription strand bias for T>A on the transcribed strand, while SBS24 (aflatoxin) showed C>A transcription asymmetry with enrichment on the transcribed strand. Signatures SBS31, SBS35, DBS5 (platinum drugs), SBS32 (prior treatment with azathioprine), and SBS25 (likely due to a chemotherapy agent) also showed transcription strand bias with strong enrichment of mutations on the transcribed strand. SBS16 (alcohol consumption) had extreme transcription strand bias with almost all mutations occurring on the transcribed strand (**Figure 1*A***). In contrast, mutational signatures due to direct

damage from ultraviolet light (*viz.*, SBS7a/b/c/d and DBS1) were the only known exogenous mutational processes to exhibit transcription strand asymmetry with strong enrichment of mutations on the untranscribed strand, consistent with damage from ultraviolet light on cytosine (**Figure 1*A&C***).

Transcription strand asymmetry with consistent enrichment of mutations on the transcribed strand was also observed for clock-like signature SBS5 (mainly for T>C mutations at ApTpN), as well as for multiple mutational signatures with unknown aetiology, including: SBS12 (biliary, kidney, and liver cancers), SBS19 (liver cancer), and ID14 (oesophageal, colorectal, and stomach adenocarcinomas; **Figure 1*A&C***). Strand bias with preferences for the untranscribed strand was observed for SBS33 (unknown aetiology) in oesophageal as well as head and neck cancers. Similarly, ID11 (oesophageal, liver, and head and neck cancers) had transcription strand asymmetry on the untranscribed strand. Lastly, other mutational signatures exhibited transcription strand asymmetry in only a small subset of cancer types (**Figure 1*A&C***).

Distribution of Mutational Signatures in Genic and Intergenic Regions

Except for signatures SBS16 and ID11, all other mutational signatures showed statistically significant enrichment of mutation in intergenic regions across most cancer types (**Supplementary Figure S1*A-C***). Excluding SBS16 and ID11, this enrichment ranged from 1.30-fold, for example, for signature SBS24, to more than 2-fold, for example, for signatures SBS17a/b. To quantify whether the observed depletion of mutations in genic regions can be attributed to transcription strand asymmetries, we nullified the asymmetry by assigning the number of mutations on both transcribed and untranscribed strands to their highest value.

Effectively, this removed the transcription strand asymmetries and inflated the number of mutations in genic regions without affecting the number of mutations in intergenic regions. Nevertheless, this inflation resulted in only a minor change from 1.37 average odds ratio of mutations in intergenic regions for real somatic mutations (0.30 standard deviation) to 1.31 average odds ratio of mutations in intergenic regions after inflating the number of mutations in genic regions by removing strand bias (0.30 standard deviation; **Supplementary Figure S1*D-E***). Overall, these results suggest that transcription strand asymmetry, usually attributed to the activity of TC-NER, do not account for the high enrichment of somatic mutations in intergenic regions.

SBS16 and ID11 showed statistically significant enrichment of mutation in genic regions in liver and oesophageal cancers, while ID11 was also enriched in genic regions in cancers of the liver. SBS16 has been previously associated with exposure to alcohol (Chang et al., 2017; Letouze et al., 2017; Li et al., 2018) and attributed to the activity of transcription-coupled damage (Haradhvala et al., 2016). Prior studies have also associated ID11 to alcohol consumption in oesophageal cancers (Moody *et al.*, 2021). Re-examining ID11 in the current cohort of whole-genome sequenced liver cancers, by comparing the mutations attributed to ID11 in 32 heavy drinkers to the mutations attributed to ID11 in 94 light drinkers, reveals a 2-fold enrichment in heavy drinkers (p-value: 1.31×10^{-3}; Mann-Whitney U test). This and the prior associations in oesophageal cancers (Moody *et al.*, 2021) strongly suggest a similar exogenous mutational processes, related to alcohol consumption, accounting for the enrichment of mutation in genic regions for both signatures SBS16 and ID11.

206 **Replication Strand Asymmetries of Mutational Signatures**

207 Replication strand bias was consistently observed in most signatures attributed to aberrant or
208 defective endogenous mutational processes with strand bias either on the leading or on the
209 lagging strand (**Figure 1*B&D***). Strong replication strand asymmetries with enrichment of
210 mutations on the leading strand was observed for signatures SBS10a, SBS10b, and DBS3 which
211 are exclusively found at extremely high levels in samples with exonuclease domain mutations in
212 DNA polymerase epsilon (POLE). This strand bias is consistent with recent observations
213 suggesting that POLE plays a major role in leading strand DNA synthesis (Daigaku et al., 2015;
214 Pursell et al., 2007; Shinbrot et al., 2014). Interestingly, SBS28 (unknown aetiology) exhibited a
215 strong replication strand bias when found at high levels in POLE deficient samples. Additionally,
216 replication strand asymmetries were also observed for SBS9, attributed to infidelity of
217 polymerase eta (POLH), and SBS10c, due to defective polymerase delta (POLD1).
218
219 Mutational signatures associated with defective DNA mismatch repair exhibited statistically
220 significant replication strand bias either on the leading or the lagging strand (**Figure 1*B***).
221 Signatures SBS14, SBS20, SBS21, and SBS26 (all attributed to mismatch repair deficiencies)
222 caused mutations preferentially on the lagging strand. Signatures SBS6, SBS15, and SBS44 (also
223 attributed to mismatch repair deficiencies) exhibited C>A and T>C substitutions on the lagging
224 strand as well as C>T preferentially found on the leading strand. ID1 and DBS2 also exhibited
225 replication strand bias on the lagging and leading strands, respectively.
226
227 Somatic mutations due to signatures SBS2 and SBS13, both attributed to the aberrant behaviour
228 of the APOBEC3 family of deaminases, were found enriched on the lagging strand in all cancer

types. This result is consistent with the observation that single-stranded DNA formed during DNA replication on the lagging strand is a major substrate for the APOBEC3 family of deaminases (Roberts et al., 2012; Saini and Gordenin, 2020). Lastly, several other mutational signatures, most with unknown aetiology, exhibited replication strand bias within a small set of cancer types (**Figure 1*B&D***).

Mutational Signatures with Strand-coordinated Mutagenesis

Prior analyses have shown that certain types of mutations on the same reference allele were observed on the same strand more frequently than expected by chance (Morganella *et al.*, 2016; Nik-Zainal et al., 2012; Roberts *et al.*, 2012). This strand-coordinated clustered mutations usually arise due to damage on single-stranded DNA, and they are often indicative of the formation of hypermutable loci in the genome (Roberts *et al.*, 2012; Saini and Gordenin, 2020).

SBS7a (UV light) attained the highest strand-coordinated mutagenesis with lengths of subsequent mutations up to 40 consecutive mutations (**Figure 1*E***). In contrast, other mutational signatures attributed to ultraviolet light, mainly, SBS7b/c/d, either did not exhibit or exhibited much lower strand-coordinated mutagenesis. APOBEC3-attributed SBS2 and SBS13 showed strand-coordinated mutagenesis with as many as 21 consecutive strand-coordinated mutations (**Figure 1*E***). SBS17b (unknown aetiology) also exhibited processive groups with as many eight strand-coordinated mutations. In ultra-hypermutated samples with deficiency in the *POLE* proofreading domain, SBS10a and SBS10b also showed strand-coordinated mutagenesis with up to 14 consecutive strand-coordinated mutations. Similarly, consecutive strand-coordinated mutations were observed for SBS4 (associated with tobacco smoking; up to seven consecutive

mutations), SBS26 (defective mismatch repair; up to eight mutations), and SBS28 (unknown aetiology; up to 11 mutations; **Figure 1*E***).

The Effect of DNA Replication Timing on Somatic Mutagenesis

Consistent with prior reports (Chen et al., 2010; Koren et al., 2012; Stamatoyannopoulos et al., 2009; Watanabe et al., 2002), the aggregated set of somatic mutations was shown to be enriched in late replicating regions for most cancer types (**Figure 2**). Specifically, from the examined 40 cancer types, SBSs were found more common in regions of the genome that undergo late replication in 39/40 cancer types (indicated as ↗:39 in **Figure 2**). In one cancer type, SBSs were not associated with replication time (indicated as →:1) and there were no cancer types in which SBSs were enriched in early replication regions (indicated as ↘:0). Similarly, DBSs and IDs were enriched in late replicating regions in 18/18 and 30/32 cancer types, respectively. Note that due to their lower mutational burdens, we could confidently evaluate DBSs only in 18 of the 40 cancer types and IDs only in 32 of the 40 cancer types. In contrast to the aggregated analysis, examining somatic mutations attributed to different mutational signatures revealed distinct replication time dependencies.

At least six mutational signatures were predominately enriched in early replicating regions: SBS6 attributed to mismatched repair (enriched in early replicating regions in 2 out of 3 cancer types; 2/3), SBS11 attributed to temozolomide therapy (1/1), SBS15 due to DNA mismatch repair deficiency (1/1), SBS16 (2/2) and ID11 (3/5) both associated with alcohol consumption, and SBS84 (1/1) due to aberrant activities of activation-induced (AID) cytidine deaminases.

ID17 signature, most probably due to *TOP2A* mutations (Boot et al., 2022), was also enriched in early replicating regions in oesophageal squamous cell carcinoma.

All mutational signatures that were previously associated with age of diagnosis in at least one cancer type had a predominately increased normalized mutational density from early to late replicating regions: SBS1 (23/36), SBS5 (36/37), SBS40 (31/31), ID1 (24/24), ID2 (21/21), and ID5 (19/20). SBS3 (14/14) and ID6 (12/12), both attributed to defective homologous recombination, as well as mutational signatures attributed to defective polymerase epsilon (SBS10a, SBS10b, SBS14, and DBS3) were also enriched in late replicating regions in all examined cancer types (**Figure 2**). Similarly, signatures attributed to reactive oxygen species (SBS18 and SBS36) were enriched in late replicating regions in all examined cancer types. Additionally, most mutational signatures due to environmental and chemotherapeutic exposures were enriched in late replicating regions in all examined cancer types including signatures of tobacco smoking (SBS4, DBS2, ID3), tobacco chewing (SBS29), ultraviolet light (SBS7a/c/d, SBS38, DBS1, and ID13), aristolochic acid (SBS22), aflatoxin (SBS24), prior treatment (SBS31, SBS32, SBS35, and DBS5), and non-canonical AID activity (SBS85). Many of the mutational signatures with unknown aetiologies were also enriched in late replicating regions: SBS8, SBS12, SBS17a/b, SBS19, SBS28, SBS33, SBS34, SBS37, and SBS41. Two of the signatures attributed to the APOBEC3 family of deaminases, SBS2 (15/17) and DBS11 (2/3), had an increased normalized mutation density from early to late replicating regions.

Importantly, SBS13, attributed to the APOBEC3 family of deaminases, showed no dependence with replication timing in 7 of the 17 examined cancer types (*viz.*, bladder, breast, uterus, cervix,

298 ovary, and thyroid, acute lymphocytic leukaemia). This is consistent with prior reports where
299 SBS13 was attributed to uracil excision of deaminated cytosine followed by processing by DNA
300 translesion polymerases in breast cancer (Helleday et al., 2014; Petljak and Maciejowski, 2020).
301 Surprisingly, in 10/17 cancer types, SBS13 was enriched in late replicating regions indicating
302 that other mechanisms may also give rise to this mutational signature. Interestingly, signature
303 SBS7b, attributed to ultraviolet light, showed no dependencies with replication timing in
304 melanoma in contrast to all other signatures attributed to ultraviolet light (*viz.*, SBS7a/c/d,
305 SBS38, DBS1, and ID13); this indicates that processing of UV-induced DNA damage by base
306 excision repair and DNA translesion polymerases may give rise to signature SBS7b. SBS30, a
307 signature of deficient base excision repair due to mutations in the bifunctional DNA glycosylase
308 *NTHL1*, showed flat replication timing in 3 of the 3 examined cancer types. Mutations due to
309 signatures SBS20 (2/4), SBS21 (2/3), and SBS44 (3/5), all attributed to failure of mismatch
310 repair, as well as mutations due to signatures SBS39 (1/1), ID4 (5/8), and ID8 (7/18), all with
311 unknown aetiologies, were also generally unaffected by replication timing (**Figure 2**).
312
313 **The Effect of Nucleosome Occupancy on Mutational Signatures**
314 Nucleosomes are the basic packing units of chromatin with each nucleosome consisting of ~147
315 base-pair (bp) DNA wrapped around a histone octamer with 60 to 80 bp linker DNA between
316 consecutive nucleosomes (Davey et al., 2002; Richmond and Davey, 2003). Previous analyses
317 have revealed dependencies between mutational signatures operative in breast cancer and
318 nucleosome occupancy (Morganella et al., 2016) as well as a pan-cancer periodicity of mutation
319 rates within nucleosomes due to multiple substitution signatures (Pich et al., 2018). However,

beyond breast cancer, there has been no cancer-specific examination of the effect of nucleosome occupancy on the accumulation of somatic mutations due to different mutational signatures.

All types of somatic mutations and most mutational signatures were depleted near nucleosomes compared to simulated data mimicking the mutational landscapes of the examined cancer genomes (**Figure 3**). Similar to simulated data, aggregated small insertions and deletions exhibited a consistent behaviour across most tumour types with aggregated indels predominantly located at linker DNA in 31 of the 32 examined cancer types (**Figure 3**). Conversely, aggregated single base and doublet-base substitutions had dissimilar behaviour across cancer types with only a subset of cancer types exhibiting similar behaviour (25/40 for substitutions; 13/18 for doublet-base substitutions; **Figure 3**). Remarkably, the majority of SBS, DBS, and ID mutational signatures were similarly affected by nucleosome occupancy across most cancer types. Further, the effect of nucleosome occupancy tended to be consistent for many signatures with a shared aetiology. Different types of periodicities of mutation rates around the nucleosome structure were observed for most signatures associated with tobacco smoking, ultraviolet light, aristolochic acid, reactive oxygen species, and defective mismatch repair (**Figure 3**).

Mutational signatures attributed to tobacco smoking (SBS4, DBS2, and ID3) exhibited similar patterns around nucleosome positions across multiple cancer types (**Figure 3**; **Supplementary Figure S2**). ID3 was the only indel mutational signatures with expected and observed enrichment of mutations near nucleosomes (**Figure 3**). Similarly, the substitution mutational signatures of alkylating agents (SBS11), aflatoxin (SBS24), tobacco chewing (SBS29), platinum therapies (SBS31, SBS35), and azathioprine treatment (SBS32) were preferentially found in

nucleosome positions in all cancer types in which they were detected. Most mutational signatures due to direct or indirect damage from ultraviolet light (SBS7b/d, SBS38, and DBS1) had a higher number of mutations at nucleosome sites. In contrast, signatures SBS7c and ID13, also attributed to UV-light exposure, were depleted at nucleosomes and enriched at linker DNA sequences.

Mutations due to SBS9, associated with polymerase eta driven replication errors, and signatures SBS10a/b/c, SBS28 and DBS3, attributed to exonuclease mutations in *POLE* and/or *POLD1*, strikingly appeared at linker DNA. Some mutational signatures associated with microsatellite instability (*viz.*, SBS21, SBS26 and DBS7) were preferably located at linker DNA. In contrast, other microsatellite instability associated signatures, namely, SBS6, SBS14, SBS15, SBS20, and SBS44, as well as SBS30, due to deficiency in base excision repair, were found to match simulated data with expected high number of mutations at nucleosome occupied regions (**Figure 3**).

Signatures SBS16 (alcohol) and SBS22 (aristolochic acid) were depleted at nucleosome positions and enriched at linker DNA sequences in all cancer types in which these signatures were detected. Similar behaviour was observed for multiple signatures with unknown aetiologies, including: SBS12, SBS34, SBS37, and SBS41. In contrast, consistent propensity for elevated mutation burden at nucleosome regions was shown by multiple other mutational signatures with unknown aetiology, including: SBS19, SBS39, DBS4, DBS6, and DBS9.

Only one of the clock-like signatures, SBS1, behaved consistently with higher number of mutations at nucleosomes in 36/36 cancer types. In contrast, signature SBS5 behaved similarly in

366 only 18/37 examined cancer types. The behaviour of SBS40 was also inconsistent across most
367 cancer types (**Figure 3**). Signature SBS3, attributed to defective homologous recombination, was
368 slightly elevated in 11/14 cancer types in which it was found. The inconsistent behaviour of these
369 flat signatures (SBS3, SBS5, and SBS40) may reflect tissue-specific differences but it could also
370 be due to technical issues as, in some cases, there is a high-level ambiguity in assigning flat
371 signatures to individual samples (Alexandrov et al., 2020; Li et al., 2020; Maura et al., 2019).
372 Interestingly, nucleosome occupancy had a similar effect on the APOBEC3 mutational
373 signatures (SBS2, SBS13, and DBS11) in most cancer types with little effect on the
374 accumulation of somatic mutations from these signatures (**Figure 3**).
375
376 Signatures SBS18 and SBS36, attributed to mutations due to DNA damage induced by reactive
377 oxygen species, exhibited higher number of mutations at nucleosome regions and strong
378 periodicity of approximately ~192 bp (**Figure 3**). Interestingly, signatures SBS17a/b showed
379 similar behaviour providing further circumstantial evidence for the hypothesis that SBS17a/b
380 may also be due to ROS damage of the deoxyribonucleoside triphosphate (dNTP) pools
381 (Christensen et al., 2019; Dvorak et al., 2007; Focaccetti et al., 2015; Hidaka et al., 2008; Inoue
382 et al., 1998; Tomkova et al., 2018). Except for ID3, attributed to tobacco smoking, and ID17,
383 associated with *TOP2A* mutations, all other signatures of small insertions and deletions were
384 preferentially located at linker DNA sequences and depleted at nucleosome positions (**Figure 3**).
385
386 Lastly, prior analyses have revealed pan-cancer periodicity of mutation rates within nucleosomes
387 due to signatures SBS4, SBS7, SBS8, SBS9, SBS10, SBS14, SBS16, SBS17, and SBS18 (Pich
388 et al., 2018). Here, we both confirm and elaborate on the cancer- and signature-specific

389 periodicity of these signatures. For example, SBS4 shows strong periodicity in cancers of the
390 lung and head and neck but not in cancers of the liver or cancers of the oesophagus
391 (**Supplementary Figure S2*A*)**. Similarly, signatures SBS10a and SBS10b behaved differently
392 with SBS10a exhibiting no periodicity and SBS10b exhibiting strong periodicity of mutation
393 rates within nucleosomes across all tissue types (**Figure 3**). Analogously, the UV-light
394 associated signatures SBS7a/b/c/d showed different level of nucleosome periodicity. In contrast,
395 signatures SBS17a/b behaved consistently in almost all examined cancer types (**Figure 3**).
396 Lastly, the current examination provides the first report for periodicities of mutation rates near
397 nucleosomes for another three signatures: SBS22, SBS36, and SBS38 (**Figure 3**).
398
399 **The Effect of CTCF Binding on Mutational Signatures**
400 CCCTC-binding factor (CTCF) is a multi-purpose sequence-specific DNA-binding protein with
401 an essential role in transcriptional regulation, somatic recombination, and chromatin architecture
402 (Ghirlando and Felsenfeld, 2016; Kentepozidou et al., 2020; Kim et al., 2015; Merkenschlager
403 and Odom, 2013; Ong and Corces, 2014). The human genome harbours many CTCF binding
404 sites with prior studies reporting that mutations due to ultraviolet light are enriched in CTCF
405 binding sites (Poulos et al., 2016).
406
407 Somatic mutations exhibited clear patterns of both enrichment and/or periodicity for multiple
408 mutational signatures and CTCF binding sites (**Figure 4**). While some signatures were
409 consistently depleted at CTCF biding sites across the majority of cancer types when compared to
410 simulated data (SBS1, SBS9, SBS10a/b, SBS15, SBS37, SBS84, and SBS85), others were

411 commonly enriched (SBS3, SBS5, SBS7a/b/d, SBS12, SBS17a/b, SBS18, SBS22, and SBS40;
412 DBS1; ID5, ID6, ID8, and ID9; **Figure 4*A***).
413
414 Aggregated single base substitutions exhibited an inconsistent behaviour across cancer types
415 with enrichment in some cancers (*e.g.,* liver cancers) and depletions in others (*e.g.*, lymphomas).
416 In contrast, indels were enriched at CTCF binding sites in the majority of cancer types (**Figure**
417 **4*A***). Remarkably, the effect of CTCF occupancy tended to be also consistent for many signatures
418 with similar aetiologies. Strong periodicities of mutation rates around CTCF binding sites were
419 observed for UV-associated signature SBS7a but not for UV-associated signatures DBS1 and
420 SBS7b/c/d (**Figure 4*B***).
421
422 Mutations due to SBS9, associated with defective polymerase eta driven replication errors, and
423 signatures SBS10a/b, found in samples with mutations in *POLE* and/or *POLD1*, were strikingly
424 depleted at CTCF binding sites. Signatures SBS15, associated microsatellite instability, was
425 strongly depleted at CTCF binding sites (**Figure 4*A***).
426
427 Only one of the clock-like signatures, SBS1, exhibited a depletion of mutations at CTCF binding
428 sites (**Figure 4*A***) while simulated data indicated that SBS1 should be enriched at these sites
429 (**Figure 4*B***). Signature SBS3, attributed to defective homologous recombination, was highly
430 elevated in CTCF binding sites for breast, ovarian, stomach, and oesophageal cancers. Signatures
431 SBS17a/b exhibited a striking enrichment at CTCF binding sites in all cancer types with
432 sufficient number of mutations from each signature (**Figure 4*A***). SBS17a showed enrichment in
433 stomach and oesophageal cancers, while SBS17b shows enrichment for stomach, oesophageal,

434 breast, pancreatic cancers, and non-Hodgkin's lymphomas. In contrast, simulated data indicate
435 that CTCF binding should have no effect on the accumulation of mutations from signatures
436 SBS17a/b (**Figure 4*B***).
437
438 **The Effect of Histone Modifications on Mutational Signatures**
439 Each nucleosome consists of four pairs of core histones: H2A, H2B, H3, and H4. Post-
440 translational modifications of histone tails play a key role in regulating DNA replication, gene
441 transcription, and DNA damage response (Allis and Jenuwein, 2016; Kouzarides, 2007; Mendez-
442 Acuna et al., 2010; Sun et al., 2020). For example, histone acetylation generally enables DNA
443 accessible for transcription, replication, and repair (Bar-Ziv et al., 2016; Dhar et al., 2017; Gong
444 and Miller, 2013; Hunt et al., 2013; Ruan et al., 2015; Sterner and Berger, 2000; Struhl, 1998;
445 Unnikrishnan et al., 2010; Vogelauer et al., 2002), while histone methylation has diverse
446 functions associated with both transcription activation and repression (Allis and Jenuwein, 2016;
447 Hyun et al., 2017). To evaluate the effect of histone modifications on the accumulation of
448 mutations from different mutational signatures, we mapped the depletion or enrichment of
449 mutations compared to simulated data in the context of the tissue specific positions of 11 histone
450 modifications: *(i)* H2AFZ, a replication-independent member of the histone H2A family that
451 renders chromatin accessible at enhancers and promoters regulating transcriptional activation and
452 repression (Giaimo et al., 2019; Lamaa et al., 2020; Subramanian et al., 2015; Zhang et al.,
453 2017); *(ii)* H3K4me1, histone mark often associated with enhancer activity (Kang et al., 2021a;
454 Kang et al., 2021b); *(iii)* H3K4me2, a histone post-translational modification enriched in *cis*-
455 regulatory regions, including both enhancers and promoters (Bernstein et al., 2005; Koch et al.,
456 2007; Wang et al., 2014); *(iv)* H3K4me3, post-translational modification enriched in active

promoters near transcription start sites (Calo and Wysocka, 2013); *(v)* H3K9ac, associated with active gene promoters and active transcription (Gates et al., 2017); *(vi)* H3K9me3, silencer, typical mark of constitutive heterochromatin (Cai et al., 2021; Saksouk et al., 2015); *(vii)* H3K27ac, histone modification generally contained at nucleosomes flanking enhancers (Calo and Wysocka, 2013; Heinz et al., 2015); *(viii)* H3K27me3, repressive, associated with silent genes (Cai *et al.*, 2021; Nestorov et al., 2013; Saksouk *et al.*, 2015); *(ix)* H3K36me3, associated with transcribed regions and playing a role in regulating DNA damage repair (Sun et al., 2020); *(x)* H3K79me2, detected in the transcribed regions of active genes (Wang et al., 2008); and *(xi)* H4K20me1, found in gene promoters and associated with gene transcriptional elongation and transcription activation (Li et al., 2011; Wang et al., 2008).

Aggregated substitutions, dinucleotides, and indels exhibited dissimilar behaviour for different histone modifications across cancer types. Aggregated substitutions were predominately depleted around H2AFZ, H3K4me2, H3K4me3, and H3K27ac in approximately half of the examined cancer types with generally no effect observed in the other half of cancer types (**Figure 5*A-C***). Aggregated doublets and indels did not have any clear pan-cancer preference but showed cancer-type specific enrichments and depletions. In contrast, many mutational signatures had generally similar behaviour in vicinity of different histone modifications.

Clock-like signature SBS1 was consistently depleted across cancer-types for multiple histone marks. In contrast, clock-like signatures SBS5 and SBS40 were generally unaffected by histone marks except for cancers of the brain and lymphatic system. APOBEC3 signatures, SBS2 and SBS13, were both enriched at activator histone mark H3K27ac for majority of cancer types.

SBS2 and SBS13 mutations were also enriched at H2AFZ locations for breast cancer. DBS11, a doublet-base substitution signature attributed to APOBEC3, was enriched in the vicinity of multiple activator histone marks in non-Hodgkin's lymphoma samples, including: H3K4me1, H2K4me3, H3K9ac, H327ac, H3K36me3, and H4K20me1 (**Figure 5*B***).

Ultraviolet-light signatures SBS7a/b/c/d and SBS38 were consistently depleted at H3K4me3, H3K9ac, and H3K27me3 locations (**Figure 5*A***). Interestingly, the doublet-signature attributed to ultraviolet-light, DBS1, was depleted only at H3K9ac and H3K27me3 locations, while the indel signature attributed to ultraviolet-light, ID13, was not depleted near any histone marks but it was enriched near H3K27ac modifications.

Signature SBS9, attributed to the activity of polymerase eta and found exclusively in B-cell malignancies, was highly enriched by the presence of multiple activator histone marks, including: H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, and H4K20me1 (**Figure 5*A***). In contrast, signatures SBS10a/b and DBS3, attributed to the failed activity of *POLE*, were all depleted near H3K4me3 and H3K27ac locations. Signature SBS10b was also depleted near H3K9ac, and H3K27me3, while signature SBS10a and DBS3 were also depleted near H3K4me1. SBS84 and SBS85, due to aberrant activities of activation-induced cytidine deaminases, were significantly enriched in the vicinity of most histone modifications.

Signatures SBS18 (reactive oxygen species) showed depletions across most cancer types at multiple histone marks, including: H2AFZ, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K27me3, and H3K79me2 (**Figure 5*A***). Further, SBS18 was enriched near H3K9me3 and H4K20me1 in medulloblastomas. In contrast SBS36, attributed to reactive

oxygen species accumulation due to *MUTYH* deficiencies, was depleted only at H3K27ac locations. Interestingly, mutations due to signatures SBS17a/b (unknown aetiology) were significantly depleted at activator histone marks, H3K4me1, H3K4me3, H3K9ac and H3K27ac; and repressive histone marks H3K9me3 and H3K27me3 in multiple cancer types (**Figure 5*A***). Tobacco smoking signatures, SBS4, DBS2, and ID3 exhibited consistent depletions near histone modifications H3K4me3 and H3K27ac in cancers of the lung, liver, oesophageal, and head and neck. Similarly, tobacco chewing signature, SBS29, showed depletions of mutations near H3K9ac and H3K27ac histone marks in liver cancer (**Figure 5*A***). Moreover, signature SBS22 (aristolochic acid) was depleted near H3K4me3 and H3K27ac in liver, kidney, and biliary cancers. In contrast, histone marks had little effect on SBS24 (aflatoxin) except enrichment at H3K36me3 histone mark in liver cancer (**Figure 5*A***). Signatures SBS31, SBS35, and DBS5 (due to chemotherapy with platinum drugs) mutations were all depleted near H3K27ac modifications, while SBS31 and SBS35 were both depleted at H3K4me3 modification sites and SBS31 was also found depleted near H3K4me1. Other mutational signatures were either unaffected by histone modifications or exhibited minor changes in a cancer-specific manner. Lastly, while enrichments and depletions of somatic mutations in the vicinity of histone marks were commonly observed for different mutational signatures (**Figure 5*A-C***), there was no specific pattern of mutations within 1,000 base-pairs for any of the examined histone modifications (*e.g.*, there was no periodicity like the one observed for CTCF binding sites or for nucleosomes). Exemplars of typically observed patterns of enrichments, depletions, or no changes around different histone modifications are provided for signatures SBS7a and ID1 across several histone modifications (**Figure 5*D***).

DISCUSSION

Our analysis provides a comprehensive resource that maps the effects of topographical genomic features on the cancer-specific accumulation of somatic mutations from distinct mutational signatures. The reported results confirmed many of the prior observations for strand asymmetry, replication timing, and nucleosome periodicity for some of the original 30 COSMICv2 SBS signatures (Morganella *et al.*, 2016; Pich *et al.*, 2018; Tomkova *et al.*, 2018). The examined larger dataset provided us with a greater resolution to identify previously unobserved pan-cancer and cancer-specific dependencies for some of these 30 signatures as well as to reveal the effect of genome architecture on the accumulation of another 47 mutational signatures across human cancer. Importantly, this report also provides the first-ever examination of the tissue-specific effect of CTCF binding and 11 different histone modifications on the accumulation of somatic mutations from different mutational signatures. In addition to the global view, in this discussion, we zoom into two specific case studies to further illustrate the power of examining topography of mutational signatures.

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

548 replication regions, lacked depletion of mutations at nucleosomes or CTCF binding sites, had
549 weak replication strand bias on lagging strand, and did not exhibit much of a strand-coordinated
550 mutagenesis (**Figure 6**). Based on these topographical differences, we have now split SBS28 into
551 two distinct signatures: *(i)* SBS28a due to *POLE* deficiency found in ultra-hypermutate
552 colorectal and uterine cancers; and *(ii)* SBS28b with unknown aetiology found in lung and
553 stomach cancers.
554
555 Second, our analyses revealed striking difference in topographical features of clustered and non-
556 clustered somatic mutations in 288 whole-genome sequenced B-cell malignancies (Alexandrov
557 *et al.*, 2020). In particular, the topographical behaviours of single base substitutions were
558 examined after separating them into non-clustered mutations, diffuse hypermutation of
559 substitutions termed *omikli* (Mas-Ponte and Supek, 2020), and longer clusters of strand-
560 coordinated substitutions termed *kataegis* (Bergstrom et al., 2022a; Nik-Zainal *et al.*, 2012;
561 Roberts *et al.*, 2012). In contrast to most cancer types, where *omikli* and *kataegis* are
562 predominately generated by APOBEC3 deaminases (Bergstrom et al., 2022b), in B-cell
563 malignancies, these clustered events are almost exclusively imprinted by the activity of AID
564 (Bergstrom *et al.*, 2022b). Further, the overall pattern of non-clustered mutations was very
565 different than the ones of *omikli* or *kataegis*. A representative example is provided using a single
566 malignant B-cell lymphoma (**Figure 7*A***) where non-clustered and clustered mutations have very
567 different trinucleotide patterns (**Figure 7*B-D***). Non-clustered mutations exhibited different
568 topographical features when compared to *omikli* or *kataegis*. Specifically, while non-clustered
569 mutations had some minor periodicity in regard to nucleosome occupancy, such periodicity was
570 not observed for any type of clustered events (**Figure 7*E***). Similarly, non-clustered mutations

were slightly depleted around CTCF binding sites while *omikli* and *kataegis* were very highly depleted (**Figure 7*F&H***). Further, non-clustered and *omikli* events were clearly enriched in late replication regions while *kataegis* was highly enriched in early replication regions (**Figure 7*G***). Distinct patterns of enrichments were also observed for both *omikli* and *kataegis* mutations in the vicinity of promoter and enhancer sites delineated by histone marks of H3K4me3, H3K9ac, H3K27ac, H3K36me3, and H4K20me1(**Figure 7*H***). Only very minor differences were observed for transcription or replication strand asymmetries between clustered and non-clustered somatic mutations across the 288 whole-genome sequenced B-cell malignancies (**Supplementary Figure S3**).

In summary, in this report we have performed a comprehensive topography analysis of mutational signatures encompassing 82,890,857 somatic mutations in 5,120 whole-genome sequenced tumours integrated with 516 tissue-matched topographical features from the ENCODE project. Our evaluation encompassed examining the effects of nucleosome occupancy, histone modifications, CTCF binding sites, replication timing, transcription strand asymmetry, and replication strand asymmetry on the accumulation of somatic mutations from more than 70 distinct mutational signatures. The results from these analyses have been provided as an online resource as a part of COSMIC signatures database, https://cancer.sanger.ac.uk/signatures/, where researchers can explore each mutational signature as well as each topographical feature in a cancer-specific manner.

FIGURE LEGENDS

Figure 1. Strand asymmetries and strand-coordinated mutagenesis of mutational signatures. ***(A)*** Transcription strand asymmetries of signatures of single base substitutions (SBSs). Rows represent the signatures, where *n* reflects the number of cancer types in which each signature was found. Columns display the six substitution subtypes based on the mutated pyrimidine base: C>A, C>G, C>T, T>A, T>C, and T>G. SBS signatures with transcription strand asymmetries on the transcribed and/or untranscribed strands with q-value ≤ 0.05 are shown in circles with blue and green colours, respectively. The colour intensity reflects the odds ratio between the ratio of real mutations and the ratio of simulated mutations, where each ratio is calculated using the number of mutations on the transcribed strand and the number of mutations on the untranscribed strand. Only odds ratios above 1.10 are shown. Circle sizes reflect the proportion of cancer types exhibiting a signature with specific transcription strand asymmetry. ***(B)*** Replication strand asymmetries of SBS signatures. Rows represent the signatures, where *n* reflects the number of cancer types in which each signature was found. Columns display the six substitution subtypes based on the mutated pyrimidine base: C>A, C>G, C>T, T>A, T>C, and T>G. SBS signatures with replicational strand asymmetries on the lagging strand or leading strand with q-value ≤ 0.05 are shown in circles with red and yellow colours, respectively. The colour intensity reflects the odds ratio between the ratio of real mutations and the ratio of simulated mutations, where each ratio is calculated using the number of mutations on the lagging strand and the number of mutations on the leading strand. Circle sizes reflect the proportion of cancer types exhibiting a signature with specific replication strand asymmetry. ***(C)*** Transcription strand asymmetries of signatures of doublet-base substitutions (DBSs) and of small insertions/deletions (IDs). Data are presented in a format similar to the one in panel *(A)*. ***(D)***

615 Replication strand asymmetries of DBS and ID mutational signatures. Data are presented in a
616 format similar to the one in panel *(B)*. ***(E)*** Strand-coordinated mutagenesis in SBS signatures.
617 Rows represent SBS signatures and columns reflect the lengths, in numbers of consecutive
618 mutations, of strand-coordinated mutagenesis groups. SBS signatures with statistically
619 significant strand-coordinated mutagenesis (q-value ≤ 0.05) are shown as circles under the
620 respective group length with a minimum length of 5 consecutive mutations. The size of each
621 circle reflects the number of consecutive mutation groups for the specified group length
622 normalized for each signature. The colour of each circle reflects the statistical significance of the
623 number of subsequent mutation groups for each group length with respect to the simulated
624 mutations using $-\log_{10}$(q-value).
625
626 **Figure 2. Interplay between replication timing and mutational signatures.** Top three panels
627 reflect results for all single base substitutions (SBSs), all dinucleotide substitutions (DBSs), and
628 all small insertions/deletions (IDs) across all examined cancer types with each cancer type
629 examined separately. Bottom panels reflect all somatic mutations attributed to a particular
630 signature across all cancer types. Replication time data are separated into deciles, with each
631 segment containing exactly 10% of the observed replication time signal (x-axes). Normalized
632 mutation densities per decile (y-axes) are presented for early (left) to late (right) replication
633 domains. Real data for SBS signatures are shown as blue bars, for DBS signatures as red bars,
634 and for ID signatures as green bars. In all cases, simulated somatic mutations are shown as
635 dashed lines. The total number of evaluated cancer types for a particular mutational signature is
636 shown on top of each plot (*e.g.,* 36 cancer types were evaluated for SBS1). For each signature,
637 the number of cancer types where the mutation density increases with replication timing is

638 shown next to ➚ (*e.g.,* 23 cancer types for SBS1). Similarly, the number of cancer types where
639 the mutation density decreases with replication timing is shown next to ➘ (*e.g.,* 0 cancer types
640 for SBS1). Lastly, the number of cancer types where the mutation density is not affected by
641 replication timing is shown next to ➞ (*e.g.,* 13 cancer types for SBS1).
642
643 **Figure 3. Relationship between mutational signatures and nucleosome occupancy.** Top three
644 panels reflect results for all single base substitutions (SBSs), all doublet-base substitutions
645 (DBSs), and all small insertions/deletions (IDs) across all examined cancer types with each
646 cancer type examined separately. Bottom panels reflect all somatic mutations attributed to a
647 particular signature across all cancer types. In all cases, solid lines correspond to real somatic
648 mutations with blue solid lines reflecting SBSs, red solid lines DBSs, and green solid lines
649 reflecting IDs. Solid lines and dashed lines display the average nucleosome signal (y-axes) along
650 a 2 kilobase window (x-axes) centred at the mutation start site for real and simulated mutations,
651 respectively. The mutation start site is annotated in the middle of each plot and denoted as 0. The
652 2 kilobase window encompasses 1,000 base-pairs 5' adjacent to each mutation as well as 1,000
653 base-pairs 3' adjacent to each mutation. For each mutational signatures, the total number of
654 similar and considered cancer types using an *X/Y* format, with *X* being the number of cancer
655 types where a signature has similar nucleosome behaviour (Pearson correlation ≥ 0.5 and q-value
656 ≤ 0.05) and *Y* representing the total number of examined cancer types for that signature. For
657 example, signature SBS3 annotated with 11/14 reflects a total of 14 examined cancer types with
658 similar nucleosome behaviour observed in 11 of these 14 cancer types.
659

Figure 4. Relationship between mutational signatures and CTCF binding sites. ***(A)*** Enrichments and depletions of somatic mutations within CTCF binding sites. Heatmaps display only mutational signatures and cancer types that have at least one statistically significant enrichment or depletion of somatic mutations attributed to signatures of either single base substitutions (SBSs), doublet-base substitutions (DBSs), or small insertions/deletions (IDs). Red colours correspond to enrichments of real somatic mutations when compared to simulated data. Blue colours correspond to depletions of real somatic mutations when compared to simulated data. The intensities of red and blue colours reflect the degree of enrichments or depletions based on the fold change. White colours correspond to lack of data for performing statistical comparisons *(e.g.*, signature not being detected in a cancer type). Statistically significant enrichments and depletions are annotated with * (q-value ≤ 0.05). ***(B)*** The top three panels reflect average CTCF occupancy signal for all SBSs, DBS, and IDs across all examined cancer types. Bottom panels reflect all somatic mutations attributed for several exemplar mutational signatures across all cancer types. In all cases, solid lines correspond to real somatic mutations with blue solid lines reflecting SBSs, red solid lines reflecting DBSs, and green solid lines reflecting IDs. Solid lines and dashed lines display the average CTCF binding signal (y-axes) along a 2 kilobase window (x-axes) centred at the mutation start site for real and simulated mutations, respectively. The mutation start site is annotated in the middle of each plot and denoted as 0. The 2 kilobase window encompasses 1,000 base-pairs 5' adjacent to each mutation as well as 1,000 base-pairs 3' adjacent to each mutation.

Figure 5. Relationships between mutational signatures and histone modifications. ***(A-C)*** Relationships between 11 histone modifications and signatures of single base substitutions

(SBSs) in panel *(A)*, doublet-base substitutions (DBSs) in panel *(B)*, and small insertions/deletions (IDs) in panel *(C)*. The examined histone modifications encompass H2AFZ, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, H3K27ac, H3K27me3, H3K36me3, H3K79me2, and H4K20me1. Rows and columns reflect the mutational signatures and histone modifications, respectively. The circle in each cell is separated in red, blue, and grey segments proportionate to the cancer types in which the signature has a specific behaviour. A red segment in a circle reflects the signature being enriched in the vicinity of a histone modification (q-value ≤ 0.05 and at least 5% enrichment). A blue segment in a circle reflects the signature being depleted in the vicinity of a histone modification (q-value ≤ 0.05 and at least 5% depletion). A grey segment in a circle corresponds to neither depletion nor enrichment of the signature in the vicinity of a histone modification. Cells without a circle correspond to insufficient data to perform any statistical comparisons. ***(D)*** Exemplars of enrichment, depletions, or no effect for several histone modifications and signatures SBS7a and ID1. Solid lines and dashed lines display the average signal for a particular histone modification (y-axes) along a 2 kilobase window (x-axes) centred at the mutation start site for real and simulated mutations, respectively. The mutation start site is annotated in the middle of each plot and denoted as 0. The 2 kilobase window encompasses 1,000 base-pairs 5' adjacent to each mutation as well as 1,000 base-pairs 3' adjacent to each mutation.

Figure 6. Topography of signature SBS28 in *POLE* deficient (*POLE*$^{-}$) and *POLE* proficient (*POLE*$^{+}$) samples. ***(A)*** Mutational patterns of signature SBS28 in *POLE*$^{-}$ and *POLE*$^{+}$ samples displayed using the conventional 96 mutational classification schema for single base substitutions. ***(B)*** Nucleosome occupancy of SBS28 in *POLE*$^{-}$ and *POLE*$^{+}$ samples. Blue solid

lines and grey dashed lines display the average nucleosome signal (y-axes) along a 2 kilobase window (x-axes) centred at the mutation start site for real and simulated mutations, respectively. The mutation start site is annotated in the middle of each plot and denoted as 0. The 2 kilobase window encompasses 1,000 base-pairs 5' adjacent to each mutation as well as 1,000 base-pairs 3' adjacent to each mutation. ***(C)*** CTCF occupancy of SBS28 in *POLE*$^{-}$ and *POLE*$^{+}$ samples. Blue solid lines and grey dashed lines display the average CTCF binding signal (y-axes) along a 2 kilobase window (x-axes) centred at the mutation start site for real and simulated mutations, respectively. The mutation start site is annotated in the middle of each plot and denoted as 0. The 2 kilobase window encompasses 1,000 base-pairs 5' adjacent to each mutation as well as 1,000 base-pairs 3' adjacent to each mutation. ***(D)*** Replication timing of SBS28 mutations in *POLE*$^{-}$ and *POLE*$^{+}$ samples. Replication time data are separated into deciles, with each segment containing exactly 10% of the observed replication time signal (x-axes). Normalized mutation densities per decile (y-axes) are presented for early (left) to late (right) replication domains. Normalized mutation densities of real somatic mutations and simulated somatic mutations from early to late replicating regions are shown as blue bars and dashed lines, respectively. ***(E)*** Replication strand asymmetry of SBS28 mutations in *POLE*$^{-}$ and *POLE*$^{+}$ samples. Bar plots display the number of mutations accumulated on the lagging strand and leading strand for six substitution subtypes based on the mutated pyrimidine base: C>A, C>G, C>T, T>A, T>C, and T>G in red and yellow colours, respectively. Simulated mutations on lagging and leading strands are displayed in hatched bar plots. Statistically significant strand asymmetries are shown with stars: * q-value $\leq$ 0.05; ** q-value $\leq$ 0.01; *** q-value $\leq$ 0.001. ***(F)*** Enrichments and depletions of SBS28 somatic mutations in *POLE*$^{-}$ and *POLE*$^{+}$ samples within CTCF binding sites, histone modifications, and nucleosome occupied regions. Red colours correspond to enrichments of real

729 somatic mutations when compared to simulated data. Blue colours correspond to depletions of
730 real somatic mutations when compared to simulated data. The intensities of red and blue colours
731 reflect the degree of enrichments or depletions based on the fold change. White colours
732 correspond to lack of data for performing statistical comparisons. Statistically significant
733 enrichments and depletions are annotated with * (q-value ≤ 0.05). ***(G)*** Strand-coordinated
734 mutagenesis of SBS28 mutations in *POLE*$^{-}$ and *POLE*$^{+}$ samples. Rows represent SBS28
735 signature in *POLE*$^{-}$ and *POLE*$^{+}$ samples across all cancer types and columns reflect the lengths,
736 in numbers of consecutive mutations, of strand-coordinated mutagenesis groups. Statistically
737 significant strand-coordinated mutagenesis (q-value ≤ 0.05) are shown as circles under the
738 respective group length with a length starting from 2 to 11 consecutive mutations. The size of
739 each circle reflects the number of consecutive mutation groups for the specified group length
740 normalized for each SBS28 signature in *POLE*$^{-}$ and *POLE*$^{+}$ samples. The colour of each circle
741 reflects the statistical significance of the number of subsequent mutation groups for each group
742 length with respect to the simulated mutations using $-\log_{10}$(q-value).
743
744 **Figure 7. Topography of non-clustered, *omikli*, and *kataegis* substitutions across 288 whole-**
745 **genome sequenced B-cell malignancies**. ***(A)*** A rainfall plot of an example B-cell malignancy
746 sample, MALY-DE_SP116612, depicting the intra-mutational distance (IMD) distributions of
747 substitutions across genomic coordinates. Each dot represents the minimum distance between
748 two adjacent mutations. Dots are coloured based on their corresponding classifications.
749 Specifically, non-clustered mutations are shown in grey, doublet-base substitutions (DBSs) in
750 red, multi-base substitutions (MBSs) in black, *omikli* events in green, *kataegis* events in orange,
751 and all other clustered events in blue. The red line depicts the sample-dependent IMD threshold

for each sample. Specific clustered mutations may be above this threshold due to corrections for regional mutation density. ***(B-D)*** The trinucleotide mutational spectra for the different catalogues of non-clustered, *omikli,* and *kataegis* mutations for the exemplar sample (DBSs and MBSs are not shown). ***(E)*** Nucleosome occupancy of non-clustered, *omikli*, and *kataegis* mutations of B-cell malignancies. Blue solid lines and grey dashed lines display the average nucleosome signal (y-axes) along a 2 kilobase window (x-axes) centred at the mutation start site for real and simulated mutations, respectively. The mutation start site is annotated in the middle of each plot and denoted as 0. The 2 kilobase window encompasses 1,000 base-pairs 5' adjacent to each mutation as well as 1,000 base-pairs 3' adjacent to each mutation. ***(F)*** CTCF occupancy of non-clustered, *omikli*, and *kataegis* mutations of B-cell malignancies. Blue solid lines and grey dashed lines display the average CTCF signal (y-axes) along a 2 kilobase window (x-axes) centred at the mutation start site for real and simulated mutations, respectively. The mutation start site is annotated in the middle of each plot and denoted as 0. The 2 kilobase window encompasses 1,000 base-pairs 5' adjacent to each mutation as well as 1,000 base-pairs 3' adjacent to each mutation. ***(G)*** Replication timing of non-clustered, *omikli*, and *kataegis* mutations of B-cell malignancies. Replication time data are separated into deciles, with each segment containing exactly 10% of the observed replication time signal (x-axes). Normalized mutation densities per decile (y-axes) are presented for early (left) to late (right) replication domains. Normalized mutation densities of real somatic mutations and simulated somatic mutations from early to late replicating regions are shown as blue bars and dashed lines, respectively. ***(H)*** Enrichments and depletions of non-clustered, *omikli*, and *kataegis* mutations of B-cell malignancies within CTCF binding sites and histone modifications. Red colours correspond to enrichments of real somatic mutations when compared to simulated data. Blue

775 colours correspond to depletions of real somatic mutations when compared to simulated data.
776 The intensities of red and blue colours reflect the degree of enrichments or depletions based on
777 the fold change. White colours correspond to lack of data for performing statistical comparisons.
778 Statistically significant enrichments and depletions are annotated with * (q-value ≤ 0.05).
779

SUPPLEMENTARY FIGURE LEGENDS

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

803 with the underlaying data reflecting fold enrichments after inflating the number of somatic
804 mutations in genic regions to remove any transcription strand asymmetry.
805
806 **Figure S2. Topography of signature SBS4 across all cancer types.** ***(A)*** Nucleosome
807 occupancy of SBS4 across all cancer types and within individual cancer type. In all cases, blue
808 solid lines and grey dashed lines display the average nucleosome signal (y-axes) along a 2
809 kilobase window (x-axes) centred at the mutation start site for real and simulated mutations of
810 SBS4, respectively. The mutation start site is annotated in the middle of each plot and denoted as
811 0. The 2 kilobase window encompasses 1,000 base-pairs 5' adjacent to each mutation as well as
812 1,000 base-pairs 3' adjacent to each mutation. ***(B)*** CTCF occupancy of SBS4 across all cancer
813 types and for each cancer type. In all cases, blue solid lines and grey dashed lines display the
814 average CTCF binding signal (y-axes) along a 2 kilobase window (x-axes) centred at the
815 mutation start site for real and simulated mutations, respectively. The mutation start site is
816 annotated in the middle of each plot and denoted as 0. The 2 kilobase window encompasses
817 1,000 base-pairs 5' adjacent to each mutation as well as 1,000 base-pairs 3' adjacent to each
818 mutation. ***(C)*** Replication timing of SBS4 across all cancer types and for each cancer type.
819 Replication time data were separated into deciles, with each segment containing exactly 10% of
820 the observed replication time signal (x-axes). Normalized mutation densities per decile (y-axes)
821 are presented for early (left) to late (right) replication domains. In all cases, blue bars and dashed
822 lines show the normalized mutation densities of real and simulated somatic mutations,
823 respectively. ***(D)*** Enrichments and depletions of SBS4 somatic mutations within CTCF binding
824 sites and histone modifications. Only histone modifications H3K4me3, H3K9me3, and H3K27ac
825 are shown as they were the only ones with statistically significant results (q-value ≤ 0.05).

Heatmap displays SBS4 signature for each cancer type in which SBS4 is found. Red colours correspond to enrichments of real somatic mutations when compared to simulated data. Blue colours correspond to depletions of real somatic mutations when compared to simulated data. The intensities of red and blue colours reflect the degree of enrichments or depletions based on the fold change. Statistically significant enrichments and depletions are annotated with * (q-value ≤ 0.05). ***(E)*** Transcription strand asymmetries of SBS4 across cancer types. Rows represent SBS4 combined across all cancer types as well as SBS4 within each individual cancer type in which SBS4 has been detected. Columns display the six substitution subtypes based on the mutated pyrimidine base: C>A, C>G, C>T, T>A, T>C, and T>G. SBS4 signature with transcription strand asymmetries on the transcribed and/or untranscribed strands with q-value ≤ 0.05 are shown in circles with blue and green colours, respectively. The colour intensity reflects the odds ratio between the ratio of real mutations and the ratio of simulated mutation, where each ratio is calculated using the number of mutations on the transcribed strand and the number of mutations on the untranscribed strand. Only odds ratios above 1.10 are shown. Circle sizes reflect the proportion of cancer types exhibiting SBS4 with specific transcription strand asymmetry. ***(F)*** Somatic mutations in genic and intergenic regions for SBS4 across cancer types. Rows represent SBS4 combined across all cancer types as well as SBS4 within each individual cancer type in which SBS4 has been detected. Columns display the six substitution subtypes based on the mutated pyrimidine base: C>A, C>G, C>T, T>A, T>C, and T>G. SBS4 signature with genic and intergenic regions asymmetries with q-value ≤ 0.05 are shown in circles with cyan and grey colours, respectively. The colour intensity reflects the odds ratio between the ratio of real mutations and the ratio of simulated mutations, where each ratio is calculated using the number of mutations in the genic regions and the number of mutations in the intergenic regions.

849 Only odds ratios above 1.10 are shown. Circle sizes reflect the proportion of cancer types
850 exhibiting SBS4 with specific genic versus intergenic regions asymmetry.
851
852 **Figure S3. Strand asymmetries of non-clustered, *omikli*, and *kataegis* substitutions across**
853 **288 whole-genome sequenced B-cell malignancies.** Transcription strand asymmetries are
854 shown in the left panels where bars display the six substitution subtypes based on the mutated
855 pyrimidine base: C>A, C>G, C>T, T>A, T>C, and T>G (depicted on the x-axes). Y-axes
856 correspond to the numbers of single base substitutions. Blue bars reflect real transcribed
857 substitutions, while shaded blue bars correspond to simulated transcribed substitutions.
858 Similarly, green bars reflect real untranscribed mutations, whereas shaded green bars correspond
859 to simulated untranscribed substitutions. Replication strand asymmetries are shown in the middle
860 panels where bars display the six substitution subtypes based on the mutated pyrimidine base:
861 C>A, C>G, C>T, T>A, T>C, and T>G (depicted on the x-axes). Y-axes correspond to the
862 numbers of single base substitutions. Red bars reflect real substitutions on the lagging strand,
863 while shaded red bars correspond to simulated substitutions on the lagging strand. Similarly,
864 yellow bars reflect real substitutions on the leading strand, whereas shaded yellow bars
865 correspond to simulated substitutions on the leading strand. Comparisons of genic and intergenic
866 regions are shown in the right panels where bars display the six substitution subtypes based on
867 the mutated pyrimidine base: C>A, C>G, C>T, T>A, T>C, and T>G (depicted on the x-axes). Y-
868 axes correspond to the numbers of single base substitutions. Cyan bars reflect real substitutions
869 in genic regions, while shaded cyan bars correspond to simulated substitutions in genic regions.
870 Similarly, grey bars reflect real substitutions in intergenic regions, whereas shaded grey bars
871 correspond to simulated substitutions in intergenic regions. Results for non-clustered mutations

872 are shown in panel ***(A)***, *omikli* mutations in panel ***(B)***, and *kataegis* mutations in panel ***(C)***.

873 Statistically significant strand asymmetries are shown with stars: * q-value ≤ 0.05; ** q-value ≤

874 0.01; *** q-value ≤ 0.001.

875

876 **SUPPLEMENTARY TABLES**

877 **Table S1.** ENCODE datasets utilized for locations of CTCF binding sites, nucleosome occupancy,

878 histone modification sites, and replication timing.

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 R01ES030993-01A1 and R01ES032547. 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. 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 provisional patent applications for “Clustered mutations for the treatment of cancer” (U.S. provisional application serial number 63/289,601) and “Artificial intelligence architecture for predicting cancer biomarker” (serial number 63/269,033). All other authors declare no competing interests.

AUTHOR CONTRIBUTIONS

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

REFERENCES

Akdemir, K.C., Le, V.T., Kim, J.M., Killcoyne, S., King, D.A., Lin, Y.P., Tian, Y., Inoue, A., Amin, S.B., Robinson, F.S., et al. (2020). Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. Nat Genet *52*, 1178-1188. 10.1038/s41588-020-0708-0.

Alexandrov, L.B., Ju, Y.S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T., et al. (2016). Mutational signatures associated with tobacco smoking in human cancer. Science *354*, 618-622. 10.1126/science.aag0299.

Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N., et al. (2020). The repertoire of mutational signatures in human cancer. Nature *578*, 94-101. 10.1038/s41586-020-1943-3.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L., et al. (2013). Signatures of mutational processes in human cancer. Nature *500*, 415-421. 10.1038/nature12477.

Allis, C.D., and Jenuwein, T. (2016). The molecular hallmarks of epigenetic control. Nat Rev Genet *17*, 487-500. 10.1038/nrg.2016.59.

Bar-Ziv, R., Voichek, Y., and Barkai, N. (2016). Chromatin dynamics during DNA replication. Genome Res *26*, 1245-1256. 10.1101/gr.201244.115.

Bergstrom, E.N., Kundu, M., Tbeileh, N., and Alexandrov, L.B. (2022a). Examining clustered somatic mutations with SigProfilerClusters. Bioinformatics. 10.1093/bioinformatics/btac335.

Bergstrom, E.N., Luebeck, J., Petljak, M., Khandekar, A., Barnes, M., Zhang, T., Steele, C.D., Pillay, N., Landi, M.T., Bafna, V., et al. (2022b). Mapping clustered mutations in cancer reveals APOBEC3 mutagenesis of ecDNA. Nature *602*, 510-517. 10.1038/s41586-022-04398-6.

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas, E.J., 3rd, Gingeras, T.R., et al. (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. Cell *120*, 169-181. 10.1016/j.cell.2005.01.001.

Boot, A., Liu, M., Stantial, N., Shah, V., Yu, W., Nitiss, K.C., Nitiss, J.L., Jinks-Robertson, S., and Rozen, S.G. (2022). Recurrent mutations in topoisomerase IIalpha cause a previously undescribed mutator phenotype in human cancers. Proc Natl Acad Sci U S A *119*. 10.1073/pnas.2114024119.

Cai, Y., Zhang, Y., Loh, Y.P., Tng, J.Q., Lim, M.C., Cao, Z., Raju, A., Lieberman Aiden, E., Li, S., Manikandan, L., et al. (2021). H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. Nat Commun *12*, 719. 10.1038/s41467-021-20940-y.

Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? Mol Cell *49*, 825-837. 10.1016/j.molcel.2013.01.038.

Chang, J., Tan, W., Ling, Z., Xi, R., Shao, M., Chen, M., Luo, Y., Zhao, Y., Liu, Y., Huang, X., et al. (2017). Genomic analysis of oesophageal squamous-cell carcinoma identifies alcohol drinking-related mutation signature and genomic alterations. Nat Commun *8*, 15290. 10.1038/ncomms15290.

Chen, C.L., Rappailles, A., Duquenne, L., Huvet, M., Guilbaud, G., Farinelli, L., Audit, B., d'Aubenton-Carafa, Y., Arneodo, A., Hyrien, O., and Thermes, C. (2010). Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. Genome Res *20*, 447-457. 10.1101/gr.098947.109.

Christensen, S., Van der Roest, B., Besselink, N., Janssen, R., Boymans, S., Martens, J.W.M., Yaspo, M.L., Priestley, P., Kuijk, E., Cuppen, E., and Van Hoeck, A. (2019). 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. Nat Commun *10*, 4571. 10.1038/s41467-019-12594-8.

Consortium, E.P., Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799-816. 10.1038/nature05874.

Daigaku, Y., Keszthelyi, A., Muller, C.A., Miyabe, I., Brooks, T., Retkute, R., Hubank, M., Nieduszynski, C.A., and Carr, A.M. (2015). A global profile of replicative polymerase usage. Nat Struct Mol Biol *22*, 192-198. 10.1038/nsmb.2962.

Davey, C.A., Sargent, D.F., Luger, K., Maeder, A.W., and Richmond, T.J. (2002). Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution. J Mol Biol *319*, 1097-1113. 10.1016/S0022-2836(02)00386-8.

Dhar, S., Gursoy-Yuzugullu, O., Parasuram, R., and Price, B.D. (2017). The tale of a tail: histone H4 acetylation and the repair of DNA breaks. Philos Trans R Soc Lond B Biol Sci *372*. 10.1098/rstb.2016.0284.

Dvorak, K., Payne, C.M., Chavarria, M., Ramsey, L., Dvorakova, B., Bernstein, H., Holubec, H., Sampliner, R.E., Guy, N., Condon, A., et al. (2007). Bile acids in combination with low pH induce oxidative stress and oxidative DNA damage: relevance to the pathogenesis of Barrett's oesophagus. Gut *56*, 763-771. 10.1136/gut.2006.103697.

Focaccetti, C., Bruno, A., Magnani, E., Bartolini, D., Principi, E., Dallaglio, K., Bucci, E.O., Finzi, G., Sessa, F., Noonan, D.M., and Albini, A. (2015). Effects of 5-fluorouracil on morphology, cell cycle, proliferation, apoptosis, autophagy and ROS production in endothelial cells and cardiomyocytes. PLoS One *10*, e0115686. 10.1371/journal.pone.0115686.

Gates, L.A., Shi, J., Rohira, A.D., Feng, Q., Zhu, B., Bedford, M.T., Sagum, C.A., Jung, S.Y., Qin, J., Tsai, M.J., et al. (2017). Acetylation on histone H3 lysine 9 mediates a switch from transcription initiation to elongation. J Biol Chem *292*, 14456-14472. 10.1074/jbc.M117.802074.

Ghirlando, R., and Felsenfeld, G. (2016). CTCF: making the right connections. Genes Dev *30*, 881-891. 10.1101/gad.277863.116.

Giaimo, B.D., Ferrante, F., Herchenrother, A., Hake, S.B., and Borggrefe, T. (2019). The histone variant H2A.Z in gene regulation. Epigenetics Chromatin *12*, 37. 10.1186/s13072-019-0274-9.

Gong, F., and Miller, K.M. (2013). Mammalian DNA repair: HATs and HDACs make their mark through histone acetylation. Mutat Res *750*, 23-30. 10.1016/j.mrfmmm.2013.07.002.

Haradhvala, N.J., Polak, P., Stojanov, P., Covington, K.R., Shinbrot, E., Hess, J.M., Rheinbay, E., Kim, J., Maruvka, Y.E., Braunstein, L.Z., et al. (2016). Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. Cell *164*, 538-549. 10.1016/j.cell.2015.12.050.

Heinz, S., Romanoski, C.E., Benner, C., and Glass, C.K. (2015). The selection and function of cell type-specific enhancers. Nat Rev Mol Cell Biol *16*, 144-154. 10.1038/nrm3949.

Helleday, T., Eshtad, S., and Nik-Zainal, S. (2014). Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet *15*, 585-598. 10.1038/nrg3729.

Hidaka, K., Yamada, M., Kamiya, H., Masutani, C., Harashima, H., Hanaoka, F., and Nohmi, T. (2008). Specificity of mutations induced by incorporation of oxidized dNTPs into DNA by human DNA polymerase eta. DNA Repair (Amst) *7*, 497-506. 10.1016/j.dnarep.2007.12.005.

1001 Huang, M.N., Yu, W., Teoh, W.W., Ardin, M., Jusakul, A., Ng, A.W.T., Boot, A., Abedi-
1002 Ardekani, B., Villar, S., Myint, S.S., et al. (2017). Genome-scale mutational signatures of
1003 aflatoxin in cells, mice, and human tumors. Genome Res *27*, 1475-1486. 10.1101/gr.220038.116.
1004 Hunt, C.R., Ramnarain, D., Horikoshi, N., Iyengar, P., Pandita, R.K., Shay, J.W., and Pandita,
1005 T.K. (2013). Histone modifications and DNA double-strand break repair after exposure to
1006 ionizing radiations. Radiat Res *179*, 383-392. 10.1667/RR3308.2.
1007 Hyun, K., Jeon, J., Park, K., and Kim, J. (2017). Writing, erasing and reading histone lysine
1008 methylations. Exp Mol Med *49*, e324. 10.1038/emm.2017.11.
1009 Inoue, M., Kamiya, H., Fujikawa, K., Ootsuyama, Y., Murata-Kamiya, N., Osaki, T., Yasumoto,
1010 K., and Kasai, H. (1998). Induction of chromosomal gene mutations in Escherichia coli by direct
1011 incorporation of oxidatively damaged nucleotides. New evaluation method for mutagenesis by
1012 damaged DNA precursors in vivo. J Biol Chem *273*, 11069-11074. 10.1074/jbc.273.18.11069.
1013 Islam, S.M.A., Díaz-Gay, M., Wu, Y., Barnes, M., Vangara, R., Bergstrom, E.N., He, Y., Vella,
1014 M., Wang, J., Teague, J.W., et al. (2022). Uncovering novel mutational signatures by *de novo*
1015 extraction with SigProfilerExtractor. bioRxiv, 2020.2012.2013.422570.
1016 10.1101/2020.12.13.422570.
1017 Jager, M., Blokzijl, F., Kuijk, E., Bertl, J., Vougioukalaki, M., Janssen, R., Besselink, N.,
1018 Boymans, S., de Ligt, J., Pedersen, J.S., et al. (2019). Deficiency of nucleotide excision repair is
1019 associated with mutational signature observed in cancer. Genome Res *29*, 1067-1077.
1020 10.1101/gr.246223.118.
1021 Kang, Y., Kang, J., and Kim, A. (2021a). Histone H3K4me1 strongly activates the DNase I
1022 hypersensitive sites in super-enhancers than those in typical enhancers. Biosci Rep *41*.
1023 10.1042/BSR20210691.
1024 Kang, Y., Kim, Y.W., Kang, J., and Kim, A. (2021b). Histone H3K4me1 and H3K27ac play
1025 roles in nucleosome depletion and eRNA transcription, respectively, at enhancers. bioRxiv,
1026 2021.2001.2005.425373. 10.1101/2021.01.05.425373.
1027 Kentepozidou, E., Aitken, S.J., Feig, C., Stefflova, K., Ibarra-Soria, X., Odom, D.T., Roller, M.,
1028 and Flicek, P. (2020). Clustered CTCF binding is an evolutionary mechanism to maintain
1029 topologically associating domains. Genome Biol *21*, 5. 10.1186/s13059-019-1894-x.
1030 Kim, S., Yu, N.K., and Kaang, B.K. (2015). CTCF as a multifunctional protein in genome
1031 regulation and gene expression. Exp Mol Med *47*, e166. 10.1038/emm.2015.33.
1032 Koch, C.M., Andrews, R.M., Flicek, P., Dillon, S.C., Karaoz, U., Clelland, G.K., Wilcox, S.,
1033 Beare, D.M., Fowler, J.C., Couttet, P., et al. (2007). The landscape of histone modifications
1034 across 1% of the human genome in five human cell lines. Genome Res *17*, 691-707.
1035 10.1101/gr.5704207.
1036 Koren, A., Polak, P., Nemesh, J., Michaelson, J.J., Sebat, J., Sunyaev, S.R., and McCarroll, S.A.
1037 (2012). Differential relationship of DNA replication timing to different forms of human mutation
1038 and variation. Am J Hum Genet *91*, 1033-1040. 10.1016/j.ajhg.2012.10.018.
1039 Kouzarides, T. (2007). Chromatin modifications and their function. Cell *128*, 693-705.
1040 10.1016/j.cell.2007.02.005.
1041 Kucab, J.E., Zou, X., Morganella, S., Joel, M., Nanda, A.S., Nagy, E., Gomez, C., Degasperi, A.,
1042 Harris, R., Jackson, S.P., et al. (2019). A Compendium of Mutational Signatures of
1043 Environmental Agents. Cell *177*, 821-836 e816. 10.1016/j.cell.2019.03.001.
1044 Lamaa, A., Humbert, J., Aguirrebengoa, M., Cheng, X., Nicolas, E., Cote, J., and Trouche, D.
1045 (2020). Integrated analysis of H2A.Z isoforms function reveals a complex interplay in gene
1046 regulation. Elife *9*. 10.7554/eLife.53375.

1047 Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter,
1048 S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer
1049 and the search for new cancer-associated genes. Nature *499*, 214-218. 10.1038/nature12213.
1050 Letouze, E., Shinde, J., Renault, V., Couchy, G., Blanc, J.F., Tubacher, E., Bayard, Q., Bacq, D.,
1051 Meyer, V., Semhoun, J., et al. (2017). Mutational signatures reveal the dynamic interplay of risk
1052 factors and cellular processes during liver tumorigenesis. Nat Commun *8*, 1315. 10.1038/s41467-
1053 017-01358-x.
1054 Li, S., Crawford, F.W., and Gerstein, M.B. (2020). Using sigLASSO to optimize cancer mutation
1055 signatures jointly with sampling likelihood. Nat Commun *11*, 3575. 10.1038/s41467-020-17388-
1056 x.
1057 Li, X.C., Wang, M.Y., Yang, M., Dai, H.J., Zhang, B.F., Wang, W., Chu, X.L., Wang, X.,
1058 Zheng, H., Niu, R.F., et al. (2018). A mutational signature associated with alcohol consumption
1059 and prognostically significantly mutated driver genes in esophageal squamous cell carcinoma.
1060 Ann Oncol *29*, 938-944. 10.1093/annonc/mdy011.
1061 Li, Z., Nie, F., Wang, S., and Li, L. (2011). Histone H4 Lys 20 monomethylation by histone
1062 methylase SET8 mediates Wnt target gene activation. Proc Natl Acad Sci U S A *108*, 3116-3123.
1063 10.1073/pnas.1009353108.
1064 Martincorena, I., and Campbell, P.J. (2015). Somatic mutation in cancer and normal cells.
1065 Science *349*, 1483-1489. 10.1126/science.aab4082.
1066 Mas-Ponte, D., and Supek, F. (2020). DNA mismatch repair promotes APOBEC3-mediated
1067 diffuse hypermutation in human cancers. Nat Genet *52*, 958-968. 10.1038/s41588-020-0674-6.
1068 Maura, F., Degasperi, A., Nadeu, F., Leongamornlert, D., Davies, H., Moore, L., Royo, R.,
1069 Ziccheddu, B., Puente, X.S., Avet-Loiseau, H., et al. (2019). A practical guide for mutational
1070 signature analysis in hematological malignancies. Nat Commun *10*, 2969. 10.1038/s41467-019-
1071 11037-8.
1072 Meier, B., Volkova, N.V., Hong, Y., Schofield, P., Campbell, P.J., Gerstung, M., and Gartner, A.
1073 (2018). Mutational signatures of DNA mismatch repair deficiency in C. elegans and human
1074 cancers. Genome Res *28*, 666-675. 10.1101/gr.226845.117.
1075 Mendez-Acuna, L., Di Tomaso, M.V., Palitti, F., and Martinez-Lopez, W. (2010). Histone post-
1076 translational modifications in DNA damage response. Cytogenet Genome Res *128*, 28-36.
1077 10.1159/000296275.
1078 Merkenschlager, M., and Odom, D.T. (2013). CTCF and cohesin: linking gene regulatory
1079 elements with their targets. Cell *152*, 1285-1297. 10.1016/j.cell.2013.02.029.
1080 Moody, S., Senkin, S., Islam, S.M.A., Wang, J., Nasrollahzadeh, D., Penha, R.C.C., Fitzgerald,
1081 S., Bergstrom, E.N., Atkins, J., He, Y., et al. (2021). Mutational signatures in esophageal
1082 squamous cell carcinoma from eight countries of varying incidence. medRxiv,
1083 2021.2004.2029.21255920. 10.1101/2021.04.29.21255920.
1084 Morganella, S., Alexandrov, L.B., Glodzik, D., Zou, X., Davies, H., Staaf, J., Sieuwerts, A.M.,
1085 Brinkman, A.B., Martin, S., Ramakrishna, M., et al. (2016). The topography of mutational
1086 processes in breast cancer genomes. Nat Commun *7*, 11383. 10.1038/ncomms11383.
1087 Nestorov, P., Tardat, M., and Peters, A.H. (2013). H3K9/HP1 and Polycomb: two key epigenetic
1088 silencing pathways for gene regulation and embryo development. Curr Top Dev Biol *104*, 243-
1089 291. 10.1016/B978-0-12-416027-9.00008-5.
1090 Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones,
1091 D., Hinton, J., Marshall, J., Stebbings, L.A., et al. (2012). Mutational processes molding the
1092 genomes of 21 breast cancers. Cell *149*, 979-993. 10.1016/j.cell.2012.04.024.

1093 Ong, C.T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology
1094 and function. Nat Rev Genet *15*, 234-246. 10.1038/nrg3663.
1095 Petljak, M., Alexandrov, L.B., Brammeld, J.S., Price, S., Wedge, D.C., Grossmann, S., Dawson,
1096 K.J., Ju, Y.S., Iorio, F., Tubio, J.M.C., et al. (2019). Characterizing Mutational Signatures in
1097 Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. Cell *176*, 1282-1294 e1220.
1098 10.1016/j.cell.2019.02.012.
1099 Petljak, M., and Maciejowski, J. (2020). Molecular origins of APOBEC-associated mutations in
1100 cancer. DNA Repair (Amst) *94*, 102905. 10.1016/j.dnarep.2020.102905.
1101 Phillips, D.H. (2018). Mutational spectra and mutational signatures: Insights into cancer
1102 aetiology and mechanisms of DNA damage and repair. DNA Repair (Amst) *71*, 6-11.
1103 10.1016/j.dnarep.2018.08.003.
1104 Pich, O., Muinos, F., Sabarinathan, R., Reyes-Salazar, I., Gonzalez-Perez, A., and Lopez-Bigas,
1105 N. (2018). Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA
1106 Minor Groove around Nucleosomes. Cell *175*, 1074-1087 e1018. 10.1016/j.cell.2018.10.004.
1107 Polak, P., Karlic, R., Koren, A., Thurman, R., Sandstrom, R., Lawrence, M., Reynolds, A.,
1108 Rynes, E., Vlahovicek, K., Stamatoyannopoulos, J.A., and Sunyaev, S.R. (2015). Cell-of-origin
1109 chromatin organization shapes the mutational landscape of cancer. Nature *518*, 360-364.
1110 10.1038/nature14221.
1111 Poulos, R.C., Thoms, J.A.I., Guan, Y.F., Unnikrishnan, A., Pimanda, J.E., and Wong, J.W.H.
1112 (2016). Functional Mutations Form at CTCF-Cohesin Binding Sites in Melanoma Due to
1113 Uneven Nucleotide Excision Repair across the Motif. Cell Rep *17*, 2865-2872.
1114 10.1016/j.celrep.2016.11.055.
1115 Pursell, Z.F., Isoz, I., Lundstrom, E.B., Johansson, E., and Kunkel, T.A. (2007). Yeast DNA
1116 polymerase epsilon participates in leading-strand DNA replication. Science *317*, 127-130.
1117 10.1126/science.1144067.
1118 Richmond, T.J., and Davey, C.A. (2003). The structure of DNA in the nucleosome core. Nature
1119 *423*, 145-150. 10.1038/nature01595.
1120 Roberts, S.A., Sterling, J., Thompson, C., Harris, S., Mav, D., Shah, R., Klimczak, L.J.,
1121 Kryukov, G.V., Malc, E., Mieczkowski, P.A., et al. (2012). Clustered mutations in yeast and in
1122 human cancers can arise from damaged long single-strand DNA regions. Mol Cell *46*, 424-435.
1123 10.1016/j.molcel.2012.03.030.
1124 Ruan, K., Yamamoto, T.G., Asakawa, H., Chikashige, Y., Kimura, H., Masukata, H., Haraguchi,
1125 T., and Hiraoka, Y. (2015). Histone H4 acetylation required for chromatin decompaction during
1126 DNA replication. Sci Rep *5*, 12720. 10.1038/srep12720.
1127 Saini, N., and Gordenin, D.A. (2020). Hypermutation in single-stranded DNA. DNA Repair
1128 (Amst) *91-92*, 102868. 10.1016/j.dnarep.2020.102868.
1129 Saksouk, N., Simboeck, E., and Dejardin, J. (2015). Constitutive heterochromatin formation and
1130 transcription in mammals. Epigenetics Chromatin *8*, 3. 10.1186/1756-8935-8-3.
1131 Sancar, A. (2016). Mechanisms of DNA Repair by Photolyase and Excision Nuclease (Nobel
1132 Lecture). Angew Chem Int Ed Engl *55*, 8502-8527. 10.1002/anie.201601524.
1133 Schuster-Bockler, B., and Lehner, B. (2012). Chromatin organization is a major influence on
1134 regional mutation rates in human cancer cells. Nature *488*, 504-507. 10.1038/nature11273.
1135 Shinbrot, E., Henninger, E.E., Weinhold, N., Covington, K.R., Goksenin, A.Y., Schultz, N.,
1136 Chao, H., Doddapaneni, H., Muzny, D.M., Gibbs, R.A., et al. (2014). Exonuclease mutations in
1137 DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins
1138 of replication. Genome Res *24*, 1740-1750. 10.1101/gr.174789.114.

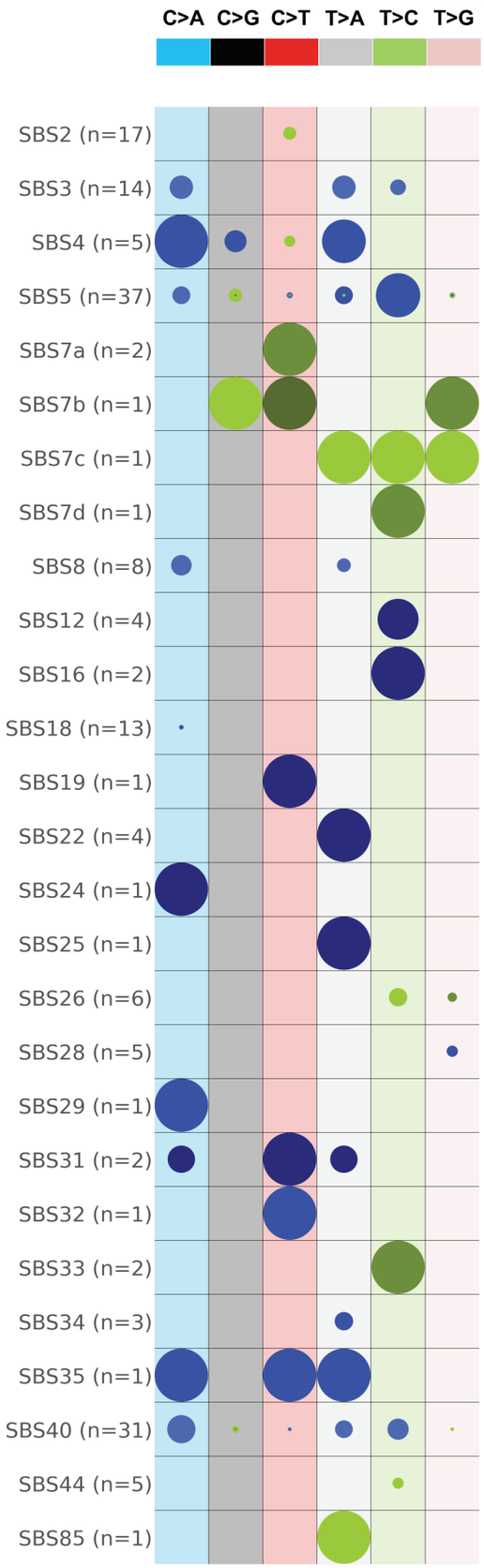
1139 Stamatoyannopoulos, J.A., Adzhubei, I., Thurman, R.E., Kryukov, G.V., Mirkin, S.M., and
1140 Sunyaev, S.R. (2009). Human mutation rate associated with DNA replication timing. Nat Genet
1141 *41*, 393-395. 10.1038/ng.363.
1142 Sterner, D.E., and Berger, S.L. (2000). Acetylation of histones and transcription-related factors.
1143 Microbiol Mol Biol Rev *64*, 435-459. 10.1128/MMBR.64.2.435-459.2000.
1144 Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719-
1145 724. 10.1038/nature07943.
1146 Struhl, K. (1998). Histone acetylation and transcriptional regulatory mechanisms. Genes Dev *12*,
1147 599-606. 10.1101/gad.12.5.599.
1148 Subramanian, V., Fields, P.A., and Boyer, L.A. (2015). H2A.Z: a molecular rheostat for
1149 transcriptional control. F1000Prime Rep *7*, 01. 10.12703/P7-01.
1150 Sun, Z., Zhang, Y., Jia, J., Fang, Y., Tang, Y., Wu, H., and Fang, D. (2020). H3K36me3,
1151 message from chromatin to DNA damage repair. Cell Biosci *10*, 9. 10.1186/s13578-020-0374-z.
1152 Supek, F., and Lehner, B. (2015). Differential DNA mismatch repair underlies mutation rate
1153 variation across the human genome. Nature *521*, 81-84. 10.1038/nature14173.
1154 Tomkova, M., Tomek, J., Kriaucionis, S., and Schuster-Bockler, B. (2018). Mutational signature
1155 distribution varies with DNA replication timing and strand asymmetry. Genome Biol *19*, 129.
1156 10.1186/s13059-018-1509-y.
1157 Unnikrishnan, A., Gafken, P.R., and Tsukiyama, T. (2010). Dynamic changes in histone
1158 acetylation regulate origins of DNA replication. Nat Struct Mol Biol *17*, 430-437.
1159 10.1038/nsmb.1780.
1160 Vogelauer, M., Rubbi, L., Lucas, I., Brewer, B.J., and Grunstein, M. (2002). Histone acetylation
1161 regulates the time of replication origin firing. Mol Cell *10*, 1223-1233. 10.1016/s1097-
1162 2765(02)00702-5.
1163 Vohringer, H., Hoeck, A.V., Cuppen, E., and Gerstung, M. (2021). Learning mutational
1164 signatures and their multidimensional genomic properties with TensorSignatures. Nat Commun
1165 *12*, 3628. 10.1038/s41467-021-23551-9.
1166 Wang, Y., Li, X., and Hu, H. (2014). H3K4me2 reliably defines transcription factor binding
1167 regions in different cells. Genomics *103*, 222-228. 10.1016/j.ygeno.2014.02.002.
1168 Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh,
1169 T.Y., Peng, W., Zhang, M.Q., and Zhao, K. (2008). Combinatorial patterns of histone
1170 acetylations and methylations in the human genome. Nat Genet *40*, 897-903. 10.1038/ng.154.
1171 Watanabe, Y., Fujiyama, A., Ichiba, Y., Hattori, M., Yada, T., Sakaki, Y., and Ikemura, T.
1172 (2002). Chromosome-wide assessment of replication timing for human chromosomes 11q and
1173 21q: disease-related genes in timing-switch regions. Hum Mol Genet *11*, 13-21.
1174 10.1093/hmg/11.1.13.
1175 Zhang, T., Joubert, P., Ansari-Pour, N., Zhao, W., Hoang, P.H., Lokanga, R., Moye, A.L.,
1176 Rosenbaum, J., Gonzalez-Perez, A., Martinez-Jimenez, F., et al. (2021). Genomic and
1177 evolutionary classification of lung cancer in never smokers. Nat Genet *53*, 1348-1359.
1178 10.1038/s41588-021-00920-0.
1179 Zhang, Y., Ku, W.L., Liu, S., Cui, K., Jin, W., Tang, Q., Lu, W., Ni, B., and Zhao, K. (2017).
1180 Genome-wide identification of histone H2A and histone variant H2A.Z-interacting proteins by
1181 bPPI-seq. Cell Res *27*, 1258-1274. 10.1038/cr.2017.112.
1182 Zhivagui, M., Ng, A.W.T., Ardin, M., Churchwell, M.I., Pandey, M., Renard, C., Villar, S.,
1183 Cahais, V., Robitaille, A., Bouaoun, L., et al. (2019). Experimental and pan-cancer genome

1184 analyses reveal widespread contribution of acrylamide exposure to carcinogenesis in humans.
1185 Genome Res *29*, 521-531. 10.1101/gr.242453.118.

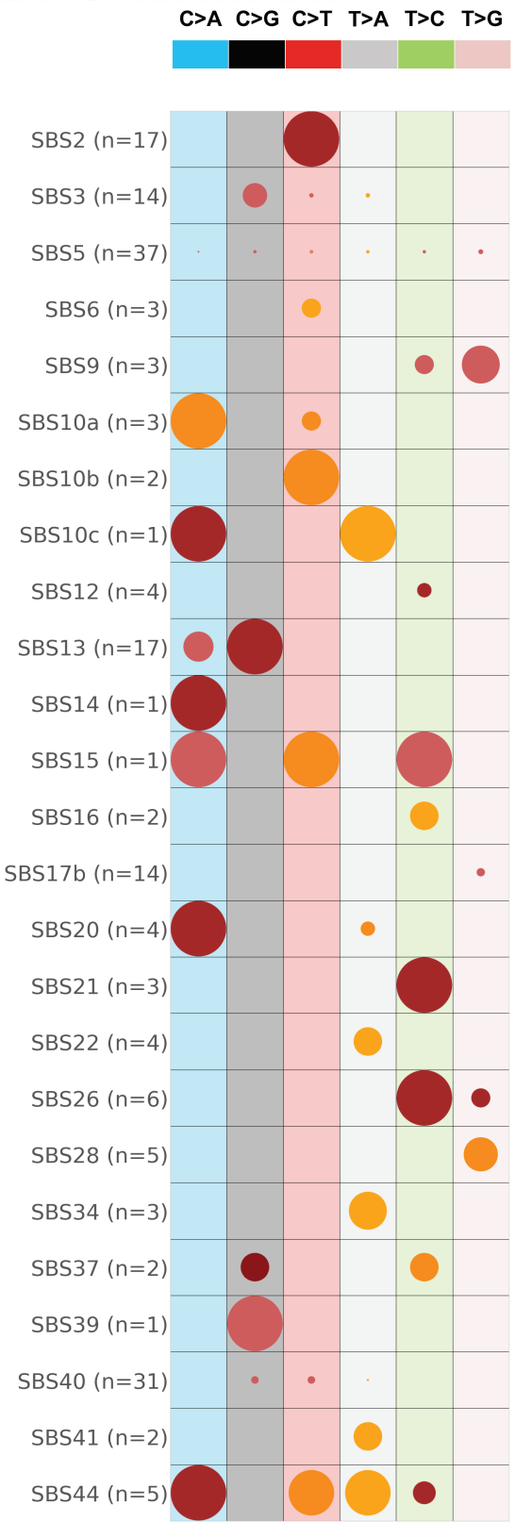
1186

Figure 1. Strand asymmetries and strand-coordinated mutagenesis of mutational signatures.

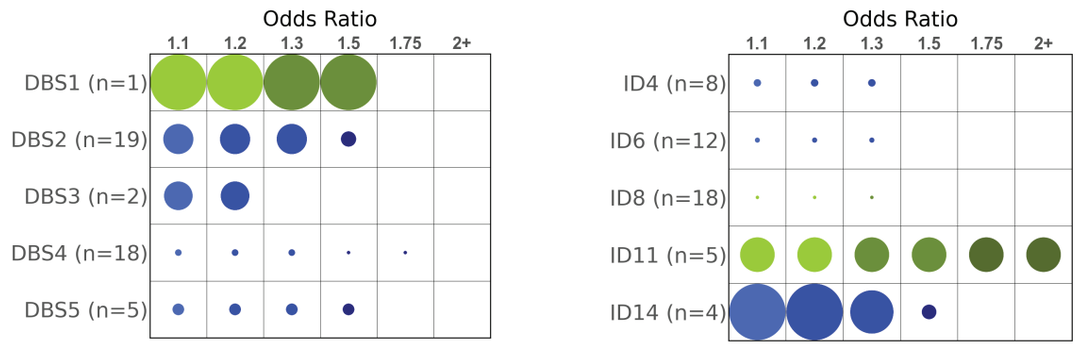
A



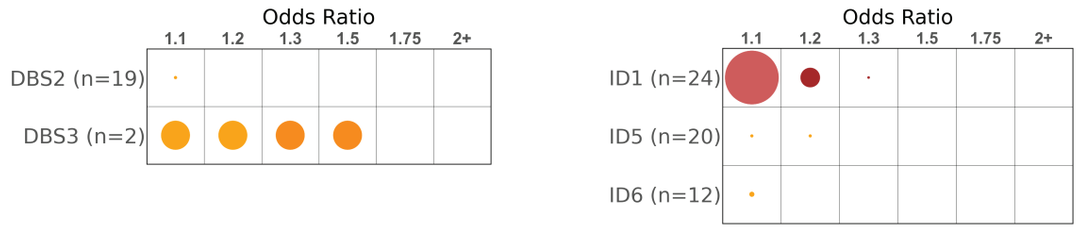
B



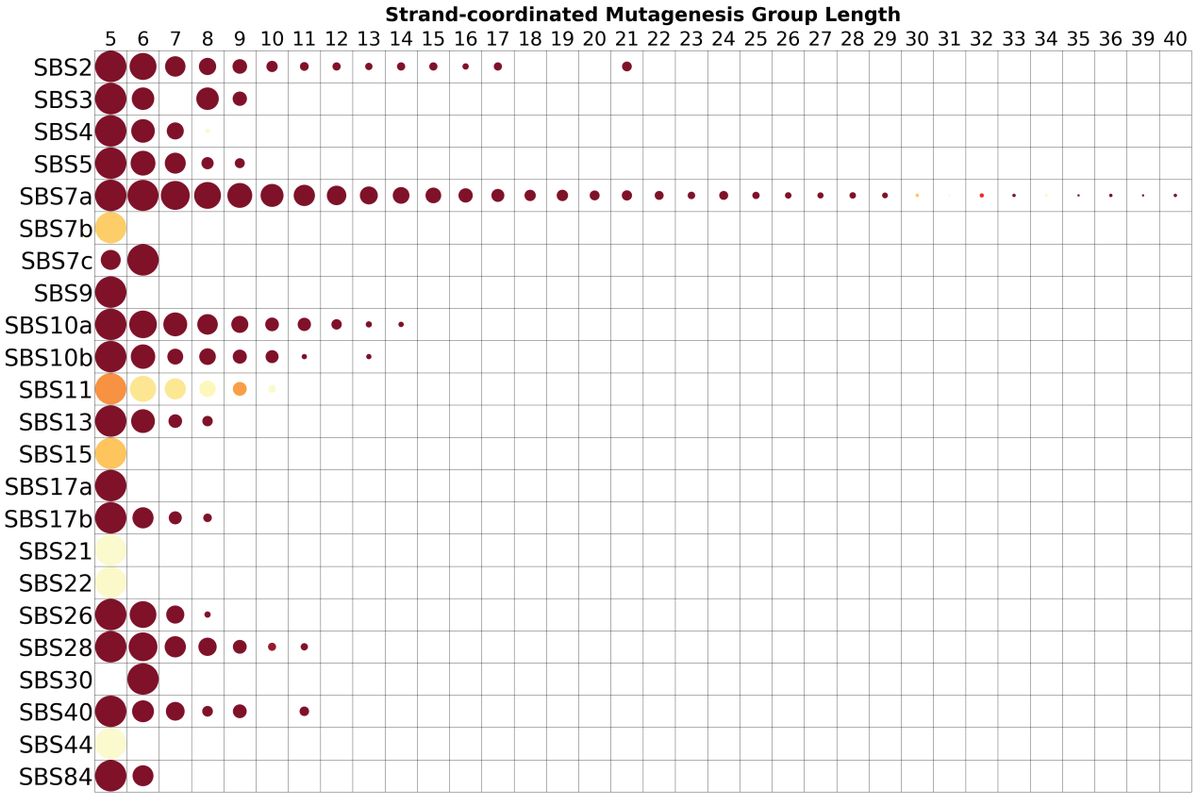
C

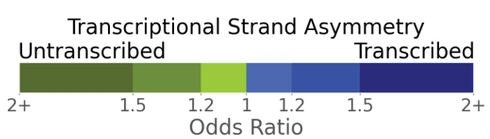


D



E





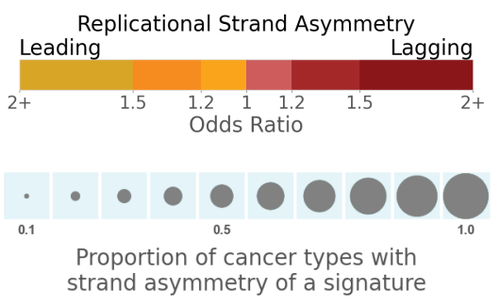


Figure 2. Interplay between replication timing and mutational signatures.

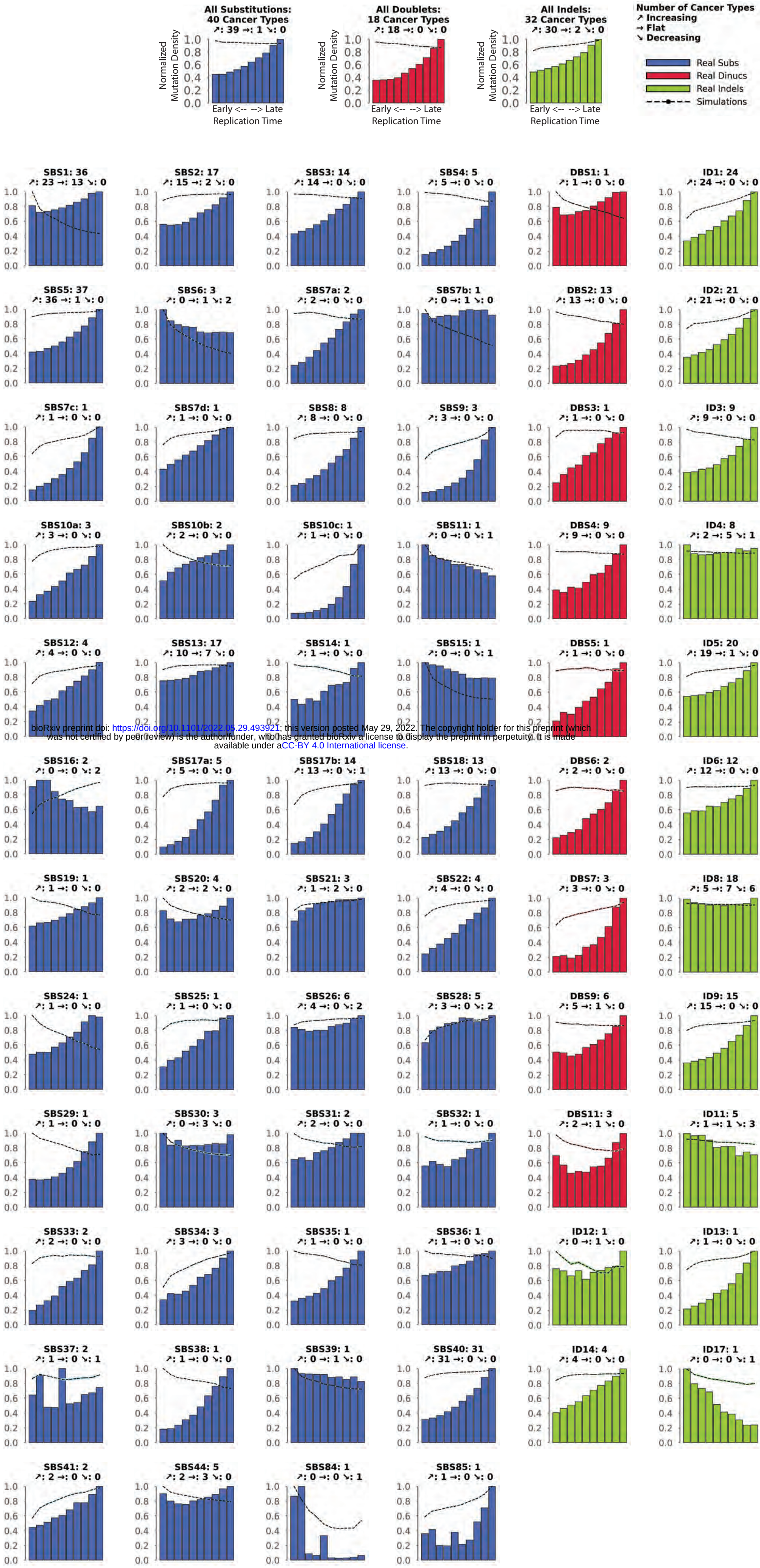
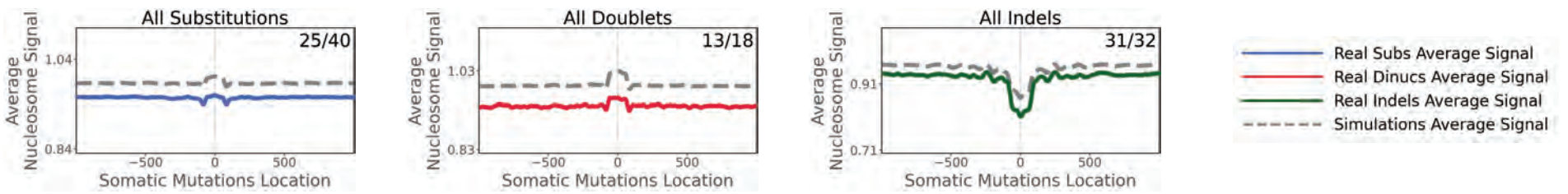
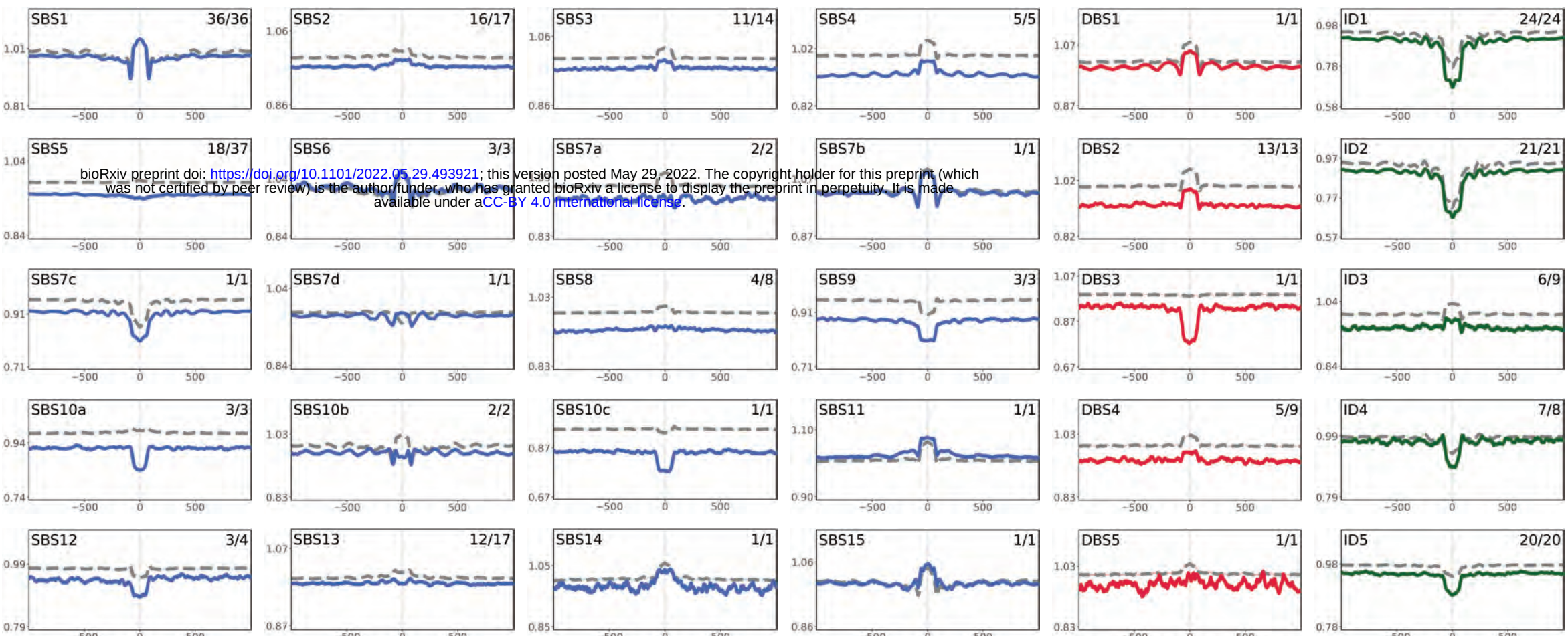


Figure 3. Relationship between mutational signatures and nucleosome occupancy.





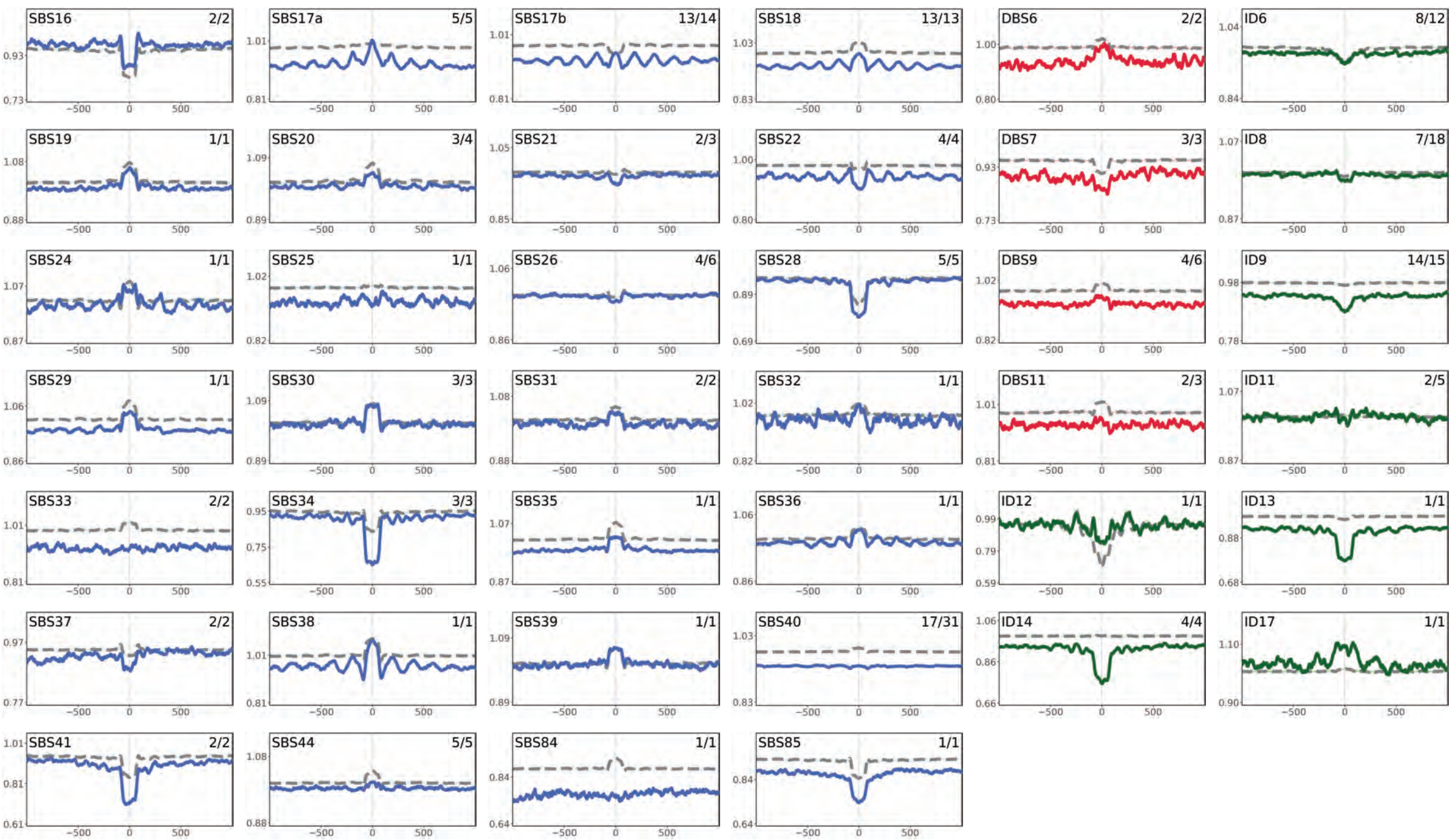
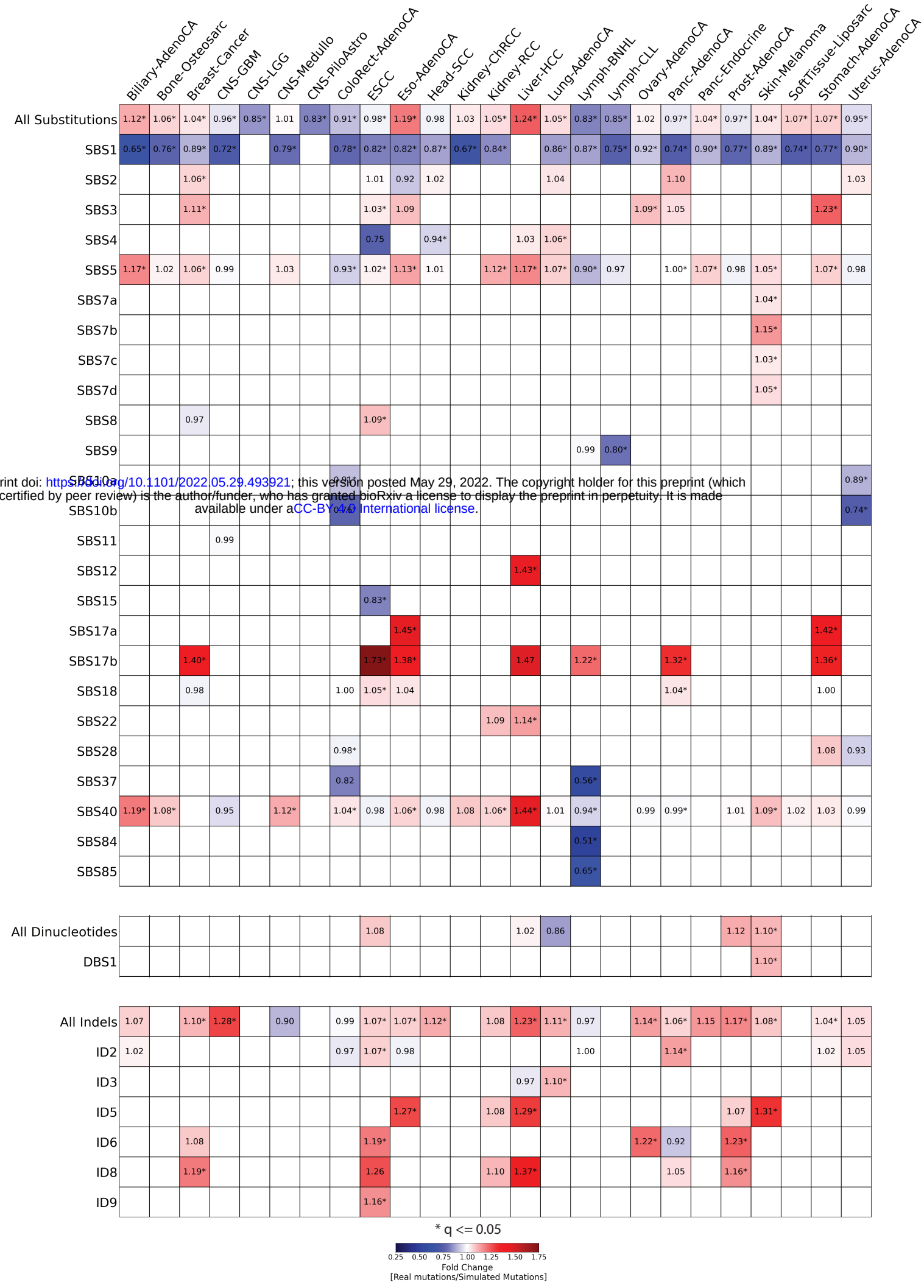


Figure 4. Relationship between mutational signatures and CTCF binding sites.

A



B

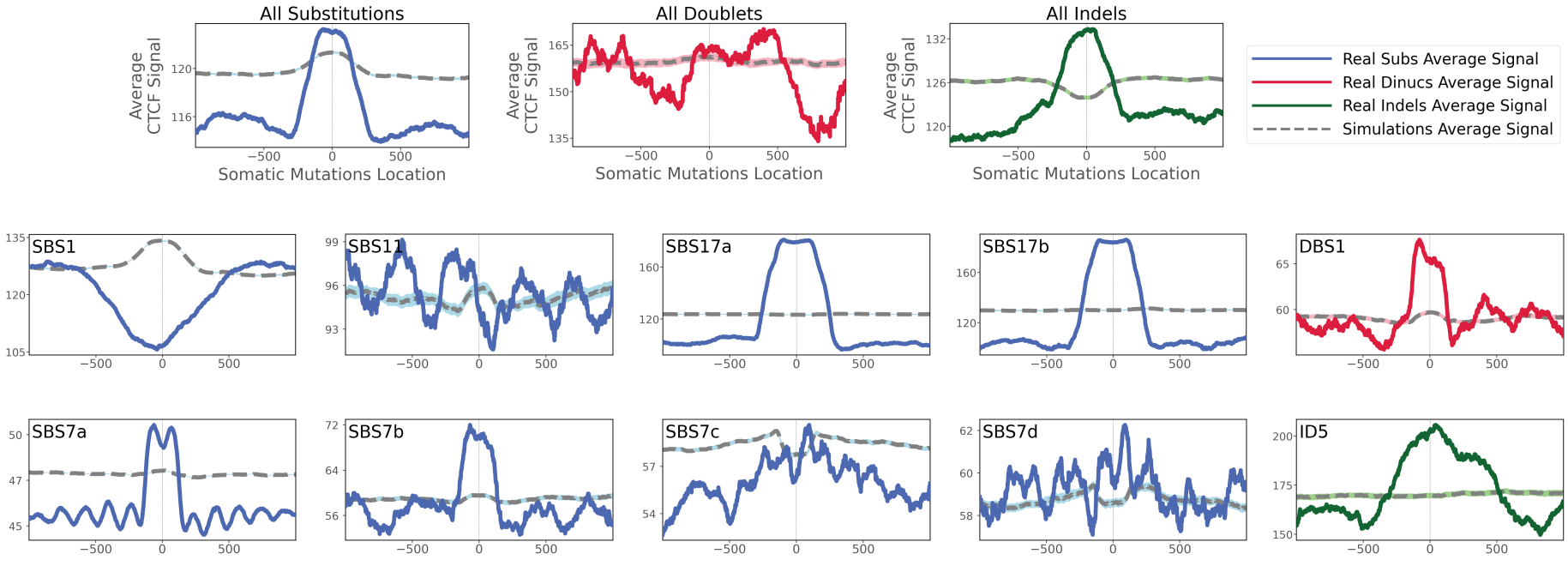
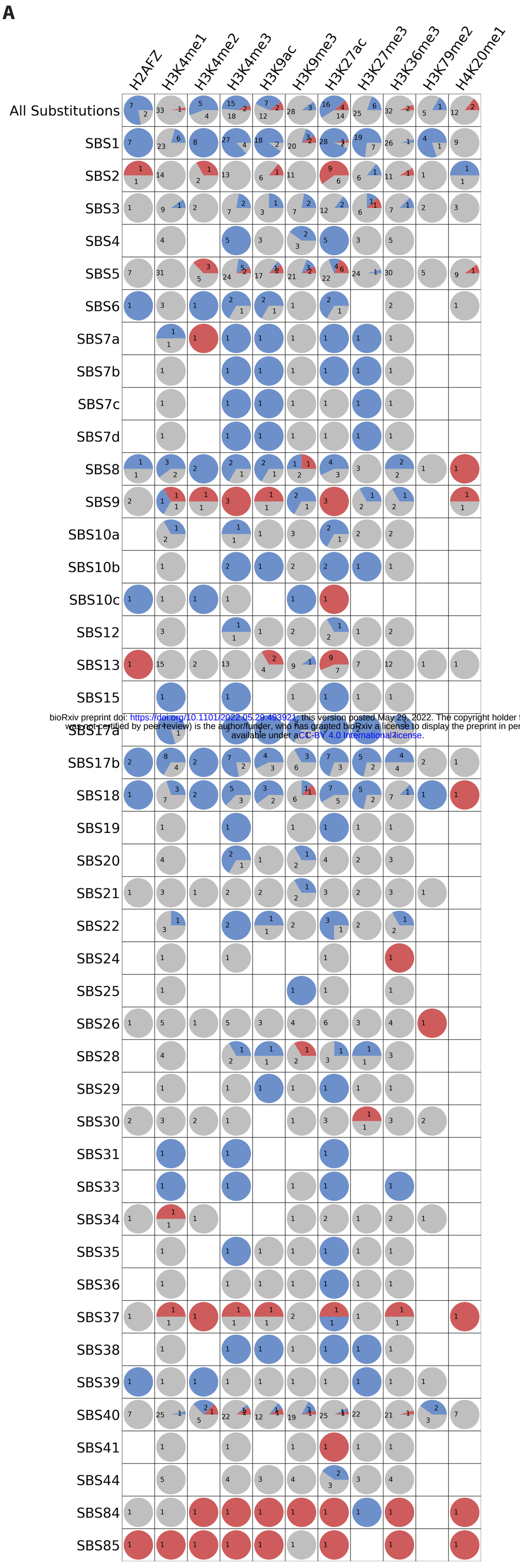


Figure 5. Relationships between mutational signatures and histone modifications.



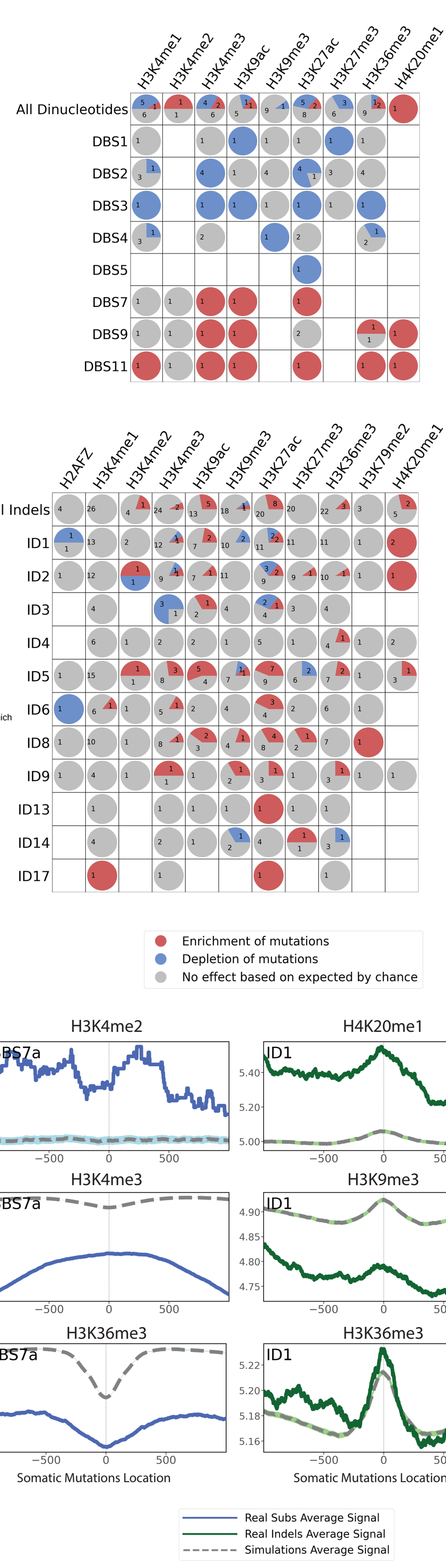


Figure 6. Topography of signature SBS28 in POLE deficient (POLE-) and POLE proficient (POLE+) samples.

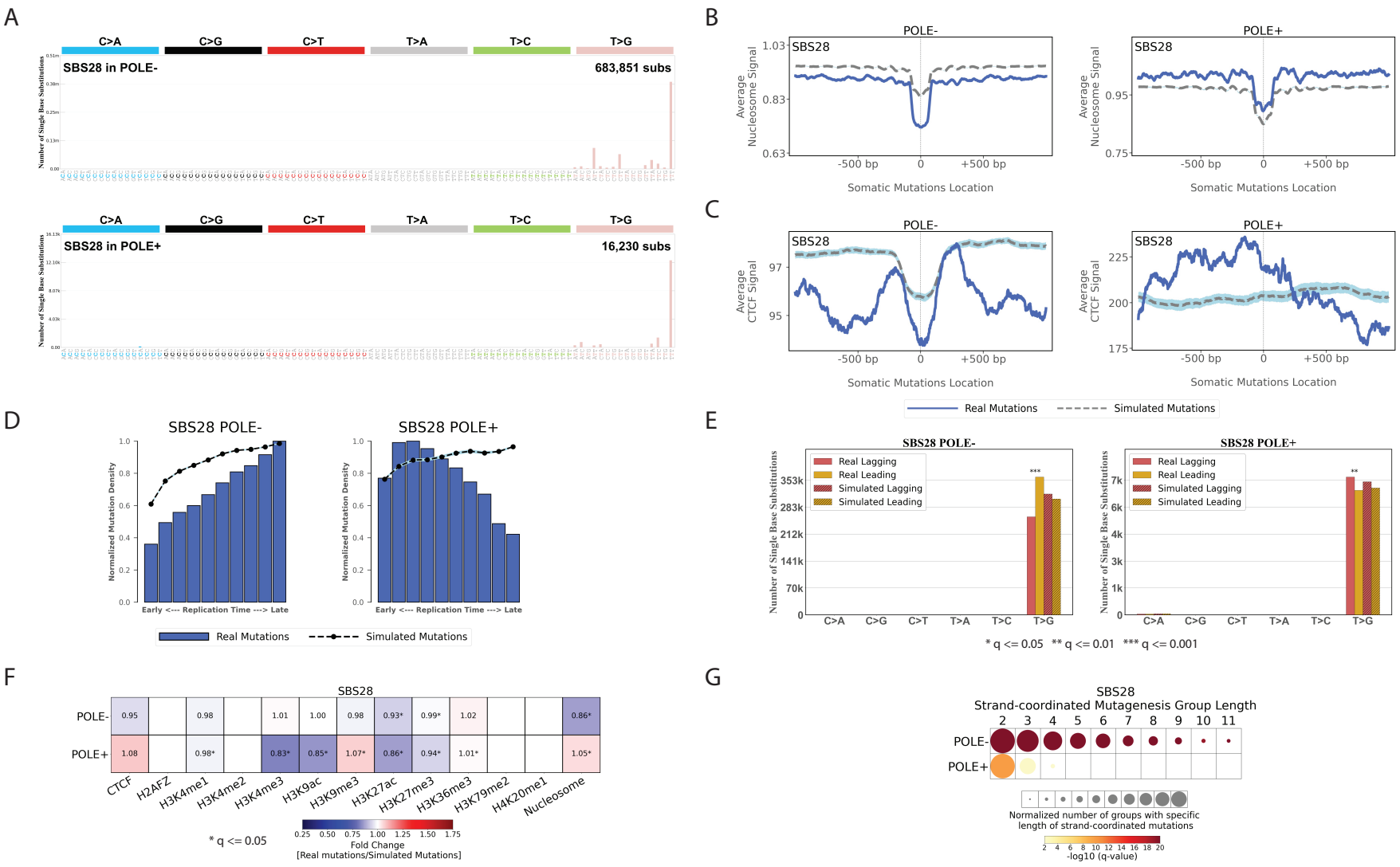


Figure 7. Topography of non-clustered, omikli, and kataegis substitutions across 288 whole-genome sequenced B-cell malignancies.