bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

# Sequence dependencies and mutation rates of localized mutational processes in cancer

Gustav Alexander Poulsgaard[1,2], Simon Grund Sørensen[1,2], Randi Istrup Juul[1,2], Morten Muhlig Nielsen[1,2], and Jakob Skou Pedersen[1,2,3]

**Affiliations:**

1: Department of Clinical Medicine, Aarhus University, Palle Juul-Jensens Boulevard 82, 8200, Aarhus N, Denmark

2: Department of Molecular Medicine (MOMA), Aarhus University Hospital, Palle Juul-Jensens Boulevard 99, 8200, Aarhus N, Denmark

3: Bioinformatics Research Center (BiRC), Aarhus University, 8000, Aarhus C, Denmark

# Abstract

*Background*: Cancer mutations accumulate through replication errors and DNA damage coupled with incomplete repair. Individual mutational processes often show strong sequence and regional preferences. As a result, some sequence contexts mutate at much higher rates than others. Mutational hotspots, with recurrent mutations across cancer samples, represent genomic positions with elevated mutation rates, often caused by highly localized mutational processes.

*Results*: We analyze the mutation rates of all 11-mer genomic sequence contexts using the PCAWG set of 2,583 pan-cancer whole genomes. We further associate individual mutations and contexts to mutational signatures and estimate their relative mutation rates. We show that hotspots generally identify highly mutable sequence contexts. Using these, we show that some mutational signatures are enriched in hotspot sequence contexts, corresponding to well-defined sequence preferences for the underlying localized mutational processes. This includes signature 17b (of unknown etiology) and signatures 62 (POLE), 7a (UV), and 72 (linked to lymphomas). In some cases, the mutation rate increases further when focusing on certain genomic regions, such as signature 62 in poised promoters, where the mutation is increased several thousand folds over the overall data set average.

*Conclusion*: We summarize our findings in a catalog of localized mutational processes, their sequence preferences, and their estimated mutation rates.

*Keywords*: pan-cancer, mutational processes, hotspots, mutation rate

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

# Introduction

Mutational signatures representing mutational processes have been identified and cataloged through analysis of large cancer genomic data sets. Some mutational processes show strong preferences for certain sequence or regional contexts, not captured by traditional mutational signature analysis. They cause variation in the mutation rate along cancer genomes with some positions displaying dramatically elevated mutation rates. These positions may manifest as mutational hotspots, which are recurrently mutated across cancer patients. Here, we use mutational hotspots identified across 2,583 whole cancer genomes to discover and characterize localized mutational processes, including their mutation rate and sequence dependency.

Cancer arises through an evolutionary process within the body, where cells accumulate somatic mutations throughout life [1, 2]. Consequently, the cancer genome represents a record of the mutational processes that have shaped it since the formation of the zygote. While the majority of mutations are neutral passengers, which do not impact the cellular phenotype, some driver mutations are under recurrent positive selection across many patients and may lead to mutational hotspots [3, 4]. However, the far majority of driver hotspots reside in the protein-coding regions [5]. Therefore, we focus on non-coding regions in the PCAWG dataset [6], where few drivers are expected [7] and where we hypothesize most hotspots are explained by localized mutational processes.

Mutagenesis is a multi-step process starting with either replication error or DNA damage, imperfect DNA repair, and then manifests through replication as mutations in descendent cells [8, 9]. Lesions are frequently formed from endogenous processes, such as the spontaneous deamination of cytosine to uracil, and the majority are successfully repaired by the DNA damage response (DDR) system [10]. Similarly, for lesions from exogenous mutagens, such as those found in tobacco smoke, the vast majority is cleared [11, 12]. Excessive lesion formation may overwhelm the DDR system and result in an increased mutation rate [13, 14].

Mutational processes act with varying intensities across the genome [11, 15–23] and certain sequence motifs experience dramatically elevated mutation rates. This is for instance the case for mutations induced by UV radiation (UV), which preferentially fall in TTT$\underline{\text{C}}$ST (S=C|G) contexts as C>T mutations [21, 24–30], and certain members of the Apolipoprotein B mRNA Editing Catalytic Polypeptide-like (APOBEC) family of DNA-editing enzymes, which

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

induce high loads of C>T and C>G mutations in T$\underline{\text{C}}$W (W = A | T) contexts [18, 31–38]. In addition, the APOBECs specifically target single-stranded regions of DNA-level stem-loop structures to produce strand-coordinated clusters of localized hypermutation, as discovered from highly context-specific mutational hotspots [36, 38–40]. Likewise, we may study other localized mutation processes through systematic analysis of hypermutable sites and their contexts across cancer genomes.

Recent large whole-genome sequencing (WGS) datasets have powered landmark discoveries of mutational processes [6, 11, 41, 42]. Mutational signature analysis has been a key tool for disentangling the mutational processes shaping these genomes [11, 18, 22, 43]. It exploits that mutational processes are shared across patients, though with varying intensities. Using non-negative matrix factorization (NMF), recurring profiles of mutation types and contexts that represent individual mutational processes are identified and their exposure in each genome evaluated [11, 18, 43].

Given the high number of free parameters and limited data availability, mutational signature analysis was only recently expanded from considering trinucleotide (±1 base pair [bp] neighbors) contexts to pentanucleotide contexts (±2 bp) [22, 44]. Some mutational processes may further depend on regional properties such as chromatin-organization [45–47], transcriptional activity [11, 48–50], and replication asymmetry [51, 52]. As all mutations are weighted equally, traditional signature analysis has limited power to learn the extended sequence contexts and regional preferences of rare localized mutational processes, which are generally underexplored [53].

We here aim to characterize the sequence-dependency and mutation rate of localized mutational processes. We categorized all single base substitutions based on their extended sequence contexts, by considering their five bp up- and down-stream regions (11-mers). This allowed us to evaluate the mutation rate for different sequence contexts. We then associated context-based categories of mutations with mutational signatures and their associated mutational processes. By exploiting that hotspots pinpoint sequence contexts with elevated mutation rates, we identified localized mutational processes and characterized their sequence and genomic feature preferences. Based on this, we decompose the factors that increase the mutation rate in increasingly smaller parts of the genome and evaluate how these factors explain the elevation in mutation rate. We contribute a comprehensive pan-cancer catalog of localized mutational processes associated with mutational signatures.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

# Results

## Baseline mutation rate across families of 11-mers

To estimate mutation rates, we initially identified 343,923 coding and 41,318,716 non-coding single nucleotide variants (SNVs) from the PCAWG set of 2,583 whole cancer genomes [6] (**Fig. 1a)**. Our analyses focused on the non-coding SNVs, which occur at an overall mutation rate of 5.96 SNV/patient/Mb (baseline mutation rate) across the dataset.

To investigate the sequence dependency of mutations, we classified all genomic positions (n=2,684,570,106) by their 5 bp up- and downstream context, which we considered as 11-mer sequences (**Fig. 1b**). To achieve strand symmetry, base pairs were viewed from the strand that contains the pyrimidine. Hence, 11-mer sequences representing genomic positions with a purine on the plus strand were reverse complemented.

The human genome (hg19) contains 2,097,090 unique strand-symmetric 11-mer sequences. Each 11-mer represents a family of concrete instances along the genome, with some families much larger than others. Unless otherwise stated, we will refer to 11-mer families simply as 11-mers . For each family, we calculated the average mutation rate per patient across the dataset, for example the AAAACTTACGG family has a mutation rate of 65.8 SNV/patient/Mb, and constitutes 500 instances with 85 SNVs across the 2,583 patients (**Fig. 1c**).

We chose to base our analysis on k-mers of length 11 as they provided an extended mutational context while allowing for a sufficient number of expected mutations for each family of k-mer (19.7 SNVs per 11-mer) to achieve useful mutation rate estimates (Methods; **Suppl. Table 1**).

## Highly variable 11-mer mutation rate

We observed a mean mutation rate of 7.47 SNV/patient/Mb across all families of 11-mers, with a high degree of variation (sd 13.1). 14.3% of 11-mers (*n* = 300,837) harbor no mutations at all, while the rest (85.7%; *n* = 1,796,253) have mutation rates ranging from 0.12 to 774 SNV/patient/Mb, displaying a 6,492-fold difference. This high variation illustrates the inherent heterogeneity of the mutation rate of 11-mers across the genome **(Fig. 1d)**.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

When we weigh 11-mer mutation rates by their number of genomic instances, we recover the baseline mutation rate (5.96 SNV/patient/Mb; **Fig. 1e**). In the downstream analyses, we focus on these weighted mutation rates to allow comparison between different genomic subsets.

Some of the variation in mutation rates is a consequence of the sampling variation caused by differences in 11-mer family sizes (i.e. their genomic spans) (**Fig. 1e**). Given uniform sizes, each family would span 1,280 instances. However, the observed number of instances per family range from 1 to 4,674,610 (median 608). For instance, non-mutated 11-mers (14.3% of all) only span 1.2% (31.2 Mb) of the genome, as most are represented by a small number of instances (median 83). Similarly, there are 1.5% highly mutated 11-mers (≥50 SNV/patient/Mb; *n* = 32,080), which only span 0.6% (16.7 Mb) of the genome and thus also represent smaller than average 11-mer families (median 128), though to a less extreme degree. However, the variation in family size is much greater for the highly mutated group than the non-mutated group (sd 2,075.4 vs 101.4).

Although many of the highly mutated 11-mers are rare, some of them are not. Common 11-mers, with equal to or more instances than the median (≥608), make up 9.4% (*n* = 3,023) of all the highly mutated 11-mers (*n* = 32,080). Thus, the high degree of variation in mutation rates across 11-mers does not appear to be governed by family size alone. Consistent with prior findings [30, 54–56], we expect that some of the variability is explained by highly mutable extended contexts.

## Assignment of mutated 11-mers to mutational processes

We next sought to identify and group mutated 11-mers by their underlying mutational processes, to characterize their relative mutation rates and extended sequence preferences. As a proxy for mutational processes, we used the 60 mutational signatures from the PCAWG consortium, generated using the SignatureAnalyzer software [11, 18, 22, 43].

Cancer genomes were grouped into cohorts with shared signature exposure (≥5% exposure; Methods), allowing us to study 11-mers across genomes with potential for shared mutational processes. We obtained 57 signature-exposed cohorts (**Fig. 2a**) each representing between 1 and 2,049 genomes inferred to share a mutational process either pan-cancer or cancer type-specific (**Fig. 2b**). As the mutation burden of a cancer genome is typically explained by multiple signatures, the signature-exposed cohorts also overlap in their ascribed genomes. Consequently, some genomes are members of several signature-exposed cohorts.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Some processes were exclusive to distinct tissues, such as UV exposure to the skin (signature 7a; 89 melanoma genomes), while other widely active processes of unknown etiologies, such as signature 17b, possibly related to gastrointestinal cancer or 5-fluorouracil exposure, were found across many cancer types (240 genomes, 13 cancer types). The intrinsic clock-like process of 5-methylcytosine deamination (signature 1) was active in the far majority (70.7%) of all genomes (1,825 genomes, 37 cancer types).

From the 11-mers in each signature-exposed cohort (**Fig. 2c**), we computed the cohort-wise mutation rates (**Fig. 2d**). As expected, we observed that some of these signature-exposed cohorts had much elevated mutation rates compared to the pan-cancer baseline mutation rate, including cohorts defined by signatures associated with mismatch repair (MMR; 63.5 ± 13.2 SNV/patient/Mb; 10.7x), POLE (579.5 ± 183.9; 97.4x), and UV (79.2 ± 66.6; 13.3x) (**Fig. 2d**; **Suppl. Fig. 4**).

For each signature-exposed cohort, we next identified the subset of 11-mers that can be explained primarily by the defining signature. We use the probabilities that individual signatures generated the observed mutations to assign 11-mers to their explanatory mutational process (**Fig. 2c**; Methods).

We characterized the mutation rates of these signature-assigned 11-mers, and found that the rates of a number of signatures were much higher than both the baseline (**Fig. 2e; Suppl. Fig. 4**) and previous analysis step (**Fig. 2f**), most notably signatures related to UV (7a), APOBEC (13), MMR deficiency (74), and POLE deficiency (10a). The 11-mers ascribed to signatures of age, MMR, POLE, and APOBEC generally spanned low fractions of the genome (2-8%). While the 11-mers assigned to tobacco, UV, and signature 17b, spanned large fractions of the genome (42%, 37%, and 26%, respectively; **Fig. 2g**).

We evaluated sequence preferences as logo plots relative to the genomic base composition (**Fig. 2h**) and relative to the composition dictated by the mutational signature (Methods; **Suppl. Fig. 4, Suppl. Fig. 5**). We observed that the base composition in the signature-assigned 11-mer sets mostly recapitulated the composition expected from the signature.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

## Hotspots identify 11-mers with high mutation rates

We consider hotspots as proxies for highly mutable positions in the genome. We hypothesize they may be targeted by highly localized and hence context specific mutational processes, which we aim to characterize. From recurrently mutated positions (**Fig. 3a**), we identified 2,842,934 SNVs across 1,339,497 hotspots in the non-coding part of the genome and 17,856 SNVs across 8,173 hotspots in protein-coding regions (**Fig. 3b**) [5, 7].

Highly recurrent hotspots, where ≥25 genomes share the mutation, are mainly found in protein-coding regions (62% [8 out of 13]; **Fig. 3b**). These include drivers in known cancer genes such as *KRAS*, *BRAF*, and *TP53* [57] and they are the results of recurrent positive selection [7]. We omit these from our analysis, as they are primarily a result of recurrent selection rather than shared localized mutation processes [5, 7].

We next asked whether any mutational signatures were enriched at hotspots, which would suggest they captured localized mutational processes with strong context preferences. For this, we evaluated the contribution of each mutational signature to the mutations of each hotspot. We then divided the hotspots into recurrence classes, where recurrence class one represents SNVs outside of hotspots, so-called singletons. We found that several mutational signatures of both known and unknown etiologies were enriched among hotspots and that the enrichment often increased with recurrence (**Fig. 3c**). Specifically, we found that the signature 17b signal in highly recurrent (5, 6, 7+) SNVs was 6.4-fold enriched from singletons. We also found hotspot-enriched signatures related to UV (signatures 7a, 67, 75, 7b), POLE (62, 10a), POLI (9), and linked to lymphoma (72) as well as several of unknown etiologies (17b, 17a, 19, 68, 28, 30).

Using the full dataset, we compared mutation rates across nested 11-mer subsets with increasing recurrence (**Fig. 4a**): a set of 11-mers that harbor at least one (1+) singleton (n = 1,796,253 11-mers), a set of 11-mers with mutations in two or more (2+) genomes (n = 351,996 11-mers), and a set of 11-mers mutated in five or more (5+) genomes (n = 3,817 11-mers). The genomic span of these 11-mer sets were 712- (2+; 954 Mb/1.3 Mb) and 3,813-times (5+; 23 Mb/6.2 Kb) higher than the hotspot positions used to define them. The mutation rate of the hotspot set (2+; 10.02 SNV/patient/Mb) was 1.7x increased over the full 11-mer set (1+; 5.96 SNV/patient/Mb), while the highly recurrent hotspots (5+; 25.53 SNV/patient/Mb) set had 4.3x increased mutation rates.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

When we held out the hotspot mutations used to identify the included 11-mers, the mutation rates were still elevated by 1.6x for the 2+ set and by 4.5x for the 5+ set (**Suppl Fig. 1**), which shows that the high mutation rates of these 11-mers are not simply the result of ascertainment bias and that the higher rates are also driven by singletons. Thus, hotspots enable us to capture highly mutable 11-mer families.

## Characterization of mutational signatures enriched at hotspots

We applied the recurrence-stratification on signature-assigned 11-mer sets. For signature 17b-assigned 11-mers with high recurrence levels (5+), we found a 9.3-fold enrichment in mutation rate and strong enrichment of adenines in the three 5'-positions offset (fourth, third, and second neighbor) from the mutated base (AAACTT; **Fig. 4b**). When we accounted for the nucleotide composition bias from the mutational signature profile (**Suppl. Fig. 5**; Methods), the 5'-A-tract remained highly enriched (**Suppl. Fig. 4**). A subset (AACTT) of this motif has also been reported by Stobbe *et al.* (2019) [21], while Alexandrov *et al.* (2020) [22] showed high mutation type probabilities in ACTTA when fitting to pentanucleotide signatures. The wide range of cancer types affected by this signature in the PCAWG dataset includes adenocarcinomas of the digestive system (esophagus, stomach, colorectum, pancreas, and biliary bladder; n=170), breast (n=4), and lung (n=3), as well as B-cell non-Hodgkin lymphoma (BNHL; n=38), bone osteosarcoma (n=13), head and neck squamous cell carcinoma (n=4), hepatocellular carcinoma (n=4), skin melanoma (n=3), and chromophobe renal cell carcinoma (n=1).

We also found that the UV-associated signature 7a was enriched in hotspots (**Fig. 3c**), and the mutation rate of signature 7a-assigned 11-mers with 5+ hotspots was enriched 178-fold compared to the baseline mutation rate (**Fig. 4c**). The nucleotide composition of this 11-mer subset displayed trends towards the TCS (S=C|G) center trinucleotide flanked by additional up- and downstream thymines (TTTCST). This motif has previously been reported [21, 29, 30]. While the emergence of this motif is driven by highly mutated 11-mers with mutation rates above the mean (164 SNV/patient/Mb), we observed a different nucleotide composition in the lowly mutated contexts (WSYT; W=A|T, Y=C|T; **Suppl. Fig. 2**).

In genomes from adenocarcinoma of the colorectum and ovary (n=8), 11-mers with 5+ hotspots assigned to the mutational signature 62 of POLE deficiency displays specificity toward the TTCG motif at mutation rates 396-fold higher than the baseline (**Fig. 4d**). From a pentanucleotide signature model, Alexandrov *et al.* (2020) [22] showed that signature 62 has moderate preference towards C>T substitutions in a TTCG context, however they found that

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

C>A substitutions in TTCTT were much more likely for this signature. The TTCG context has also been reported by others [22, 58, 59]. Our findings suggest that POLE-associated signature 62 displays highly localized mutagenesis in TTCG contexts. We also found highly increased mutation rates and strong sequence specificities towards the TTTCTTT hepta-nucleotide motif for POLE-signatures 10a (265-fold) and 61 (184-fold; **Suppl. Fig. 4**). This is an extension of the highly mutable TTCTT motif modeled by the POLE-associated pentanucleotide signatures 10a, 61, 62, 63, and 66 from Alexandrov *et al*. (2020) [22].

For signature 72, which is associated with B-cell lymphomas (BNHL and chronic lymphocytic leukemia), we observed 4.3-fold increased mutation rates in the 5+ set over the baseline (**Fig. 4e**). The nucleotide context showed a strong trend toward the WGCT motif. Though signature 72 has no clear etiology, this motif highly resembles a known hotspot motif (AGCT) of AID activity [60, 61], known to be involved in lymphomagenesis [62].

The AID hotspot motif also emerged from the 5+ set assigned to signature 19, and the mutation rates increased 1.4-fold over the baseline (**Fig. 4f**). Signature 19 is active in BNHL genomes, but no etiology has been proposed for this signature. Though the mutational profile of signature 19 is very different from signature 72 (cosine similarity = 0.24), the similar sequence contexts of these signature-assigned 11-mers with hotspots support a relatedness to AID-mutagenesis.

## Localized mutational processes are operative in distinct genomic elements

To evaluate whether the hotspot-associated mutational processes show preference for specific genomic regions, we examined the mutation rate of signature-assigned 11-mers found within functional genomic elements from ENCODE [63] and compared them to the equivalent subsets of genome-wide 11-mers. We expected the mean mutation rate of 11-mers in each genomic region to be equal to that of the genome-wide subset when genomic regions do not affect mutagenicity. Contrarily, we found that certain genomic regions contain 11-mers with higher mutation rates compared to the corresponding genome-wide subset (**Fig. 5**).

For signature 17b, the mutation rate of 11-mers drastically increased in enhancers (10.6-fold), insulators (19.6-fold), heterochromatin (12.1-fold), and repetitive regions (104-fold) (**Fig. 5a**). We found that enhancers and heterochromatin displayed weak 5’-A-tracts, while the repetitive regions were strongly enriched for an extended motif

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(GAAACTTCTTT; **Fig. 5a**) beyond what is captured by hotspots (AAACTT; **Fig. 4b**). Interestingly, the same 11-bp sequence context in repetitive regions also showed high mutation rates for POLE signatures 78 (183.7-fold) and 63 (1,120-fold) (**Suppl. Fig. 4**).

To further evaluate GAAACTTCTTT mutability in repetive elements, we annotated 11-mer instances with repeat-classes from RepeatMasker [64] (Methods; **Suppl. Fig. 3**). We found that this 11-mer is indeed highly mutable (72.8-fold) in repetitive regions pan-cancer. Additionally, we observed that the mutated instances almost exclusively (82.8%; 1,200 out of 1,450) occured in alpha satellite repeats, characteristic of the centromeres.

For the UV signature 7a, 11-mer mutation rates increased in heterochromatin (215-fold), enhancers (222-fold), promoters (384-fold), and repetitive regions (552-fold) (**Fig. 5b**). The 11-mer subsets within insulators, enhancers, heterochromatin and repetitive regions had strong sequence tendencies towards the TTTCSTT (S=C|G) motif, consistent with previous reports of T-tracts in UV hotspot motifs [26, 65, 66]. This motif was far less pronounced in promoters, even though they have previously been coupled to increased UV-mutability [27, 29].

The POLE-associated (signature 62) subsets displayed strong sequence preferences for the POLE-motif (TTCG) and dramatically increased mutation rates in promoters (1,771-fold), enhancers (1,392-fold), and repetitive elements (3,603-fold) (**Fig. 5c**). The latter showed not only the highest mutation rate, but also strong sequence preference (ATTCGA) for an adenine flanking each end of the POLE-motif.

Last, we found increased mutation rate for the B-cell lymphoma signature 72 in active (96-fold), weak (54-fold), and poised promoters (195-fold), which were further enriched for the motif (AGCT) seen in genome-wide hotspots (**Fig. 5d**). Similarly, signature 19 with the same hotspot-motif, displayed strong sequence dependency and increased mutation rate in poised promoters (202-fold) (**Suppl. Fig. 4**).

## Several signatures exhibit strongly localized behavior

In combination, we identified sets of positions in specific genomic regions that are targeted by localized mutational processes and subject to much elevated mutation rates (**Fig. 6**). We can decompose the increase in mutation rate into explanatory factors. Together, these factors each define increasingly smaller parts of the genome where the underlying

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

processes are increasingly active. This allows us to identify the sequence characteristics of highly mutable contexts and the relative rate increase they contribute.

For instance, for signature 17b (**Fig. 6a**), the exposure-cohort has a modestly increased mutation rate over the baseline (1.5x rate increase; 2,300 Mb genomic span), which is further increased for the large context set where it dominates (1.4x; 699 Mb). Recurrently mutated contexts (4.4x; 12.5 Mb) and repetitive regions (11.3x; 7.6 Kb) further restrict the set of positions to a well-defined 11-bp context (GAAACTTCTTT) with a dramatically elevated mutation rate (104-fold). This mutational signature has been associated with gastrointestinal cancers and exposure to the genotoxic chemotherapy 5-fluorouracil, though no explanation exists for increased mutability in this highly defined nucleotide sequence [67]. Where available (136 out of 240 patients), the clinical data showed that no patients were exposed to neoadjuvant chemotherapy, thus these tumors are treatment naive and we can rule out 5-fluorouracil as the explanatory process for them.

Samples exposed to the main UV-signature (7a) generally have high mutation rates (10x; 2,130 Mb). When further restricted to contexts where the signature dominates (1.9x; 1,000 Mb), contexts with mutational recurrence (9.3x; 9.1 Mb), and finally repetitive regions (3.1x; 23.8 Kb) the mutation rate increases at scales similar to signature 17b (**Fig. 6b**). Despite their differences in exposed tissues, the processes underlying signatures 7a (UV) and 17b (unknown) both prefer sequence motifs with A/T-tracts 5' to the mutated nucleotide at similar rates.

Generally, patients exposed to POLE-signature 62 had very high mutation rates (77.7x; 2,120 Mb) with high fractions (median exposure 17.9%) of mutations explained by this signature (**Fig. 6c**). Consequently, signature 62-contexts increased only slightly in mutation rate over the exposed cohort (1.3x; 206 Mb). Extending on signature contribution, both high mutational recurrence (4.0x; 3.3 Mb) and location in repetitive regions (9.1x; 8.0 Kb) contributed large mutation rate increases. Compared to signature 62, mutational recurrence contributed slightly less to the mutation rate in POLE-associated signatures 10a (2.0x; 3 Mb) and 61 (1.6x; 15 Mb). However, for the highly mutables contexts, POLE-signatures 10a and 61 showed preference for a different core motif (TTCT) than for signature 62 (TTCG) (**Suppl. Fig. 4**). This may reflect that POLE deficiency can lead to distinct mechanistic processes.

While the mutational signature 72 by itself did not result in dramatic mutation rate changes, mutational recurrence provided an increased rate (5.0x; 0.89 Mb) similar to the effect seen in

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the above examples (4-9x). Signature 72 and 19 of unknown etiologies shared the preference for a common motif (AGCT) known as the AID-hotspot motif [60, 61].

In the four cases above, hotspots contributed with a 4-9x increased mutation rate over mutational signatures, which is consistent with our signature-agnostic hotspot-characterization (4.3x; **Fig. 4a**).

# Discussion

In this study, we exploited mutational hotspots to define subsets of the genome that are targeted by localized mutational processes and systematically catalog their mutation rates and sequence preferences. We found that mutation rates increase by 4-400 fold compared to the average pan-cancer mutation rate (baseline) in sequence contexts subject to localized mutational processes associated with UV (signature 7a), POLE (signature 62), lymphomas (signature 72), and an unknown etiology (signature 17b). This is 5-18 times higher than what can be explained by cancer type and mutational signatures alone. Additionally, we found that mutation rates are further elevated (104-3,604 fold) in distinct sequence motifs within genomic regions related to repetitive DNA (signatures 17b, 7a, 62) and promoters (signature 72). We provide a comprehensive catalog of localized mutational processes, their sequence motifs, and their observed mutation rates (**Suppl. Fig. 4**).

Consistent with literature, we found that UV-associated mutagenesis (signature 7a) targets TTTCST-sequences (S=C|G), which are highly mutated across multiple genomic regions [21, 26, 30]. However, the highly mutated contexts are more ambiguous in promoters, and thus we did not observe a clear motif for these regions. Interestingly, melanoma genomes frequently harbor hotspot mutations in promoter elements explained by ETS-mediated sensitization of DNA to UV-induced cyclobutane pyrimidine dimer formation [27, 28, 68, 69]. The binding of DNA by ETS-transcription factors is estimated to contribute a 16-170-fold elevated mutation rate at ETS-binding sites (CTTCCGG and YYTTCC) [28, 69]. We did not observe this ETS-motif in our analyses. However, for UV-assigned 11-mers with high recurrence, we found a bimodal distribution of mutation rates associated with different sequence preferences (TTTCST [high] and WSYT [low]), thus potentially capturing multiple mechanisms by which UV may induce mutations. This shows that our k-mer-centric and rate-based analysis approach can aid in the generation of mechanistic hypotheses for mutational processes. Similar approaches will gain increased power in future large whole-genome cancer datasets.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We observed that two signatures of unknown etiology (signatures 19 and 72) are associated with a hotspot motif (WGCT), which highly resembles the known AID hotspot motif (AGCT) [60, 61]. Additionally, these processes have increased mutability in promoters, which is in line with reported AID off-target effects [70]. Thus, the potential of capturing AID-mutagenesis through signatures 19 and 72 may be further explored.

We found that the rate of signature 17b-mutations is elevated (9-fold) in a genome-wide hotspot motif (AAACTT) (**Fig. 4b**), which adds more context to the previously identified signature 17-motifs (ACTTA and AACTT) [21, 22, 71].

Consistent with signature 17 mutations being enriched in cohesin/CTCF-binding sites [72–74], we found a 20-fold mutation rate increase in certain contexts within insulator elements (**Fig. 5a**). However, in these regions, we did not observe the signature 17b-characteristic 5'-A-tract before the CTT core nucleotides. Thus, the mutational mechanism acting in these elements may be distinct from those causing AAACTT-hotspot mutations in the rest of the genome.

Unexpectedly, we also found a highly enriched 11-mer (GAAACTTCTTT) in the alpha satellite repeats of centromeric regions, which was associated with both signature 17b and the POLE signatures (63 and 78). This 11-mer contains the reported 5'-A-tract, however it also contains some intrinsic repeat structure that may be broken down into triplicates of the repeat unit, S(W)$_{2\text{-}3}$ (S=C|G; W=A|T). Such repeats may adopt secondary DNA structures that facilitate mutagenesis by certain processes, such as APOBECs targeting single-stranded DNA in stem-loops [36, 38, 40] or MMR deficiency leading to increased mutability of AT-rich short inverted repeats [39]. As alpha satellite repeats are replicated in the late S-phase [75], the mutational processes shaping this part of the genome are likely linked to late replication. Mutagenesis from POLE deficiency and the signature 17 process are both associated with late replication [36, 52]. Taken together, this is consistent with GAAACTTCTTT being associated with these processes in our analyses.

Just like the other motifs subject to tissue-specific localized mutational processes, the AAACTT motif possesses properties that either increase susceptibility to DNA damage, avoidance of repair, or both. Replication-timing and strand-asymmetry profiles of signature 17-mutations have been shown to be similar to those found for signatures of tobacco and UV exposure. Thus, they may share the property of being linked to environmental DNA-damage mechanisms [52]. Specifically, oxidative damage to the dGTP pool has been proposed as a

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

possible explanation for signature 17-mutations, resulting from exposure to gastric acid in gastrointestinal tumors or exposure to the genotoxic chemotherapeutic 5-fluorouracil in treated tumors [19, 52, 71, 76]. However, these hypotheses do not explain the characteristic motif of signature 17-mutagenesis and the mechanisms involved remain largely unexplained [67].

The signature 17-mutational process has been shown to correlate with the helical periodicity of DNA wound around the nucleosome core [77]. The highest mutation rates are found in the nucleosome-facing minor grooves, likely explained by hindered base excision repair in these sites [77]. While the rigid structure of long A-tracts may constrain DNA winding around the nucleosome [78], short A-tracts likely affect nucleosomal DNA flexibility and thus direct their positioning within the nucleosome with respect to the dyad [79, 80]. Such intra-nucleosomal forces may in turn hinder DNA repair at nucleosome-facing minor groove CTT lesions, thus in part explaining the A-tract motif associated with these mutations. At least, it is possible that lesions in proximity of A-tracts are repaired at different rates than the rest of the genome [81].

In agreement with existing literature [21, 22, 58, 59], we found POLE-mutagenesis to be associated with two highly mutated motifs (TTCG and TTTCTTT) and that their mutation rates dramatically increased over the baseline (184-396-fold). Mutations localized to the TTCG motif seem to be more pronounced for signature 62 than any other POLE signature, though this signature also encompasses TCT mutations. Fang *et al.* (2020) [59] suggest that mutations acquired in distinct domains of the POLE gene may give rise to distinct mutational patterns depending on the mutant-POLE DNA-affinity. Thus, it is possible that there exists even more examples of single mutagenic mechanisms generating different mutation types dependent on their specific loss- or gain-of-function mutants.

# Conclusion

Our findings provide higher resolution of the sequences targeted by localized mutational processes and contribute mutation rate estimates of these. Our comprehensive catalog (**Suppl. Fig. 4**) of mutational processes may aid the construction of more accurate models of the mutational processes in cancer, which capture the mutation rate variation. Such models are important for accurate statistical driver identification among the landscape of passenger hotspot mutations caused by localized processes [82]. In addition, the models may also

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

contribute to deeper understanding of cancer risk, somatic evolution, cancer development, and tumor biology.

The mutational patterns of localized processes active across cancers may serve as future biomarkers for detection of such processes and their associated etiologies in cancer samples. In samples with weak mutation signals, catalogs of localized mutational processes may power detection of active processes through targeted sequencing of their possible genomic targets. For cancer-associated mutational processes, this may translate to new opportunities for liquid biopsies to enable early cancer detection and surveillance of cancer evolution in the patient.

# Methods

## Whole cancer genome data set

The analysis was based on the full set of SNVs calls of 2,583 cancer genomes calls generated by the The Pan-Cancer Analysis of Whole-Genomes (PCAWG) consortium [6]. The GRCh37/hg19 reference genome was used throughout.We focused on SNVs in the non-protein-coding part of the autosomal chromosomes. We excluded protein-coding regions to reduce potential signals of positive selection.The sex chromosomes were excluded as they include a higher rate of false SNVs calls [6].

## Counting k-mer occurrences

First, we counted the number k-mer instances in chromosome 1-22 using the oligonucleotideFrequency function from the Biostrings (version 2.50.2) package in R (version 3.5.1). We obtained the chromosome sequences through the R package BSgenome.Hsapiens.UCSC.hg19 (version 1.4.0). Second, we summed the counts of identical k-mers across the chromosomes. Third, to achieve strand symmetry, we collapsed reverse complementary pairs of k-mers and represented them by the sequence with a center pyrimidine (C or T) together with the total pair sum. For example, for k=11, the AAAGAAGTTTC ($n_{purine}$ = 5,250) and GAAACTTCTTT ($n_{pyrimidine}$ = 5,495) pair was represented by GAAACTTCTTT ($n_{total}$ = 10,745).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

# Mutational signature annotation

## Genome-wide mutational signature annotation

Signature posterior probabilities for the 96 different trinucleotide mutation types in each genome were calculated with SignatureAnalyzer and provided by the PCAWG consortium [22]. We downloaded 60 mutational signature annotations of all 2,583 whitelisted PCAWG genomes (www.synapse.org/#!Synapse:syn11761189.6), which describe the exposure to signature X in genome Y. We classified a genome as exposed to a given mutational signature, when the signature load was equal to or above 5% of the genome's mutation burden.

## SNV-level mutational signature annotation

We assigned the signature posterior probabilities to each mutation, which were annotated with the most likely signature as in [7].

## 11-mer-wide mutational signature annotation

To further focus our analyses on sequence contexts explained by the mutational signature used to define the signature-exposure cohort, we assign 11-mers to the signature that primarily explains all 11-mer instances with SNVs.

The principle of signature assignment of 11-mers relies on three steps:

1) We calculated the mean posterior signature probabilities across identical 11-mers within an exposure cohort. We averaged posterior signature probabilities of SNVs in hotspots for each hotspot position to yield a position-wise mean, which we then used to calculate the mean across instances of an 11-mer family. This captures the average predicted probability that a given mutational signature generated the mutations.
2) We annotated each 11-mer with the signature that had the highest mean posterior probability and referred to this signature as most likely to explain this set of SNVs.
3) We identified the 11-mer families annotated with the signature in question for the given signature-exposed subset.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

## Definition of hotspots and recurrently mutated 11-mers

We used SNV recurrence to identify 11-mers with high expected mutability. A recurrence count for hotspots was defined as the number of pan-cancer genomes with a shared position-specific SNV. We annotated 11-mers with the highest recurrence count observed across its instances. This annotation was used to further subset 11-mers into two groups: (1) 11-mers where at least one instance had a hotspot, i.e. recurrence of two or more, (2+ k-mer set) (2) 11-mers where at least one instance had a hotspot of recurrence five or more (5+ k-mer set).

## Genomic regions

We annotated the mutated 11-mer instances with which genomic region they occurred in and stratified 11-mers according to 15 different regions defined by ENCODE [83, 84]. Further characterization of repeat elements was performed using RepeatMasker (http://www.repeatmasker.org/) [64].

## Mutation rate calculation

For each 11-mer its mutation rate (SNV / patient / Mb) was calculated as follows

$$11mer\ mutation\ rate\ = \frac{n_{SNV}}{genomic\ span \times n_{patients}}.$$

For a set of 11-mers, the mean mutation rate was calculated as follows

$$mean\ 11mer\ mutation\ rate\ = \frac{\sum_{k-mers\ in\ set} n_{SNV}}{(\sum_{k-mers\ in\ set} instances) \times n_{patients}}\ .$$

## Sequence context

Sequence information in logo plots was calculated as the Kullback-Leibler divergence between the observed and expected frequency of nucleotides at each position. The expected distribution was derived as the genome-wide autosomal distribution of nucleotides, i.e. A = 29.5%, T = 29.5%, C = 20.5%, and G = 20.5%.

The surprise of observing a nucleotide, $a$, at a given position, $i$, is estimated as the Kullback-Leibler divergence:

$$D_{KL}\ (p_i, q_i)\ =\ \sum_{a\ \in\ (A,C,G,T)} p_{a,i} \times log_2 \frac{p_{a,i}}{q_{a,i}}\ ,$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

where $p_{a,i}$ is the observed frequency and $q_{a,i}$ is the expected frequency of nucleotide $a$ in position $i$. The divergence is visualized using a logo plot with letter $height_{a,i}$ proportional to letter frequency, $p_{a,i}$, and divergence, $D_{KL}(p_i, q_i)$, in that position:

$$height_{a,i} = p_{a,i} \times D_{KL}(p_i, q_i).$$

## Null model of 11-mer nucleotide composition mutational signatures

A null model of 11-mer nucleotide distribution was derived by weighting the genomic 11-mer distribution by the mutation context probability of each signature.

We thereby derived an expected nucleotide distribution in 11-mers under a null model for each signature (**Suppl. Fig. 5**). To achieve this, we more specifically (1) created position frequency matrices for each 11-mer family, (2) collapsed and summed base counts across families sharing trinucleotide context, and (3) weighted each trinucleotide group by the mutation type probability from a given signature.

# Author contributions

Gustav Alexander Poulsgaard (GAP), Simon Grund Sørensen (SGS), Randi Istrup Juul (RIJ), Morten Muhlig Nielsen (MMN), and Jakob Skou Pedersen (JSP).

JSP conceived the project. GAP performed data analysis with contributions from SGS, RIJ, and MMN. GAP drafted the manuscript and all figures. JSP supervised the project. All authors discussed the results and contributed to the final version of this manuscript.

# Acknowledgement

We thank the Independent Research Fund Denmark | Medical Sciences (8021-00419B), Aarhus University, Aarhus University Research Foundation (AUFF-E-2020-6-14), the Danish Cancer Society (R307-A17932), and the Health Research Foundation of Central Denmark Region (R56-A2972-B1845) for funding the project.

We thank the Pan-Cancer Analysis of Whole Genomes (PCAWG) project under the International Cancer Genome Consortium (ICGC) for data access, including variant calls and mutational signatures.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

All computing was performed at the GenomeDK high-performance computing (HPC) facility. We thank GenomeDK and Aarhus University for providing HPC resources and support.

# Ethics declarations

Not applicable, as all data comes from the published and fully consented PCAWG study [6].

# Competing interests

The authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

# References

1. Kinzler KW, Vogelstein B. Lessons from hereditary colorectal cancer. Cell. 1996;87:159–70.

2. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100:57–70.

3. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009;458:719–24.

4. Bozic I, Antal T, Ohtsuki H, Carter H, Kim D, Chen S, et al. Accumulation of driver and passenger mutations during tumor progression. Proc Natl Acad Sci U S A. 2010;107:18545–50.

5. Juul RI, Nielsen MM, Juul M, Feuerbach L, Pedersen JS. The landscape and driver potential of site-specific hotspots across cancer genomes. NPJ Genom Med. 2021;6:33.

6. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020;578:82–93.

7. Rheinbay E, Nielsen MM, Abascal F, Wala JA, Shapira O, Tiao G, et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature. 2020;578:102–11.

8. Pagès V, Fuchs RPP. How DNA lesions are turned into mutations within cells? Oncogene. 2002;21:8957–66.

9. Fedeles BI, Essigmann JM. Impact of DNA lesion repair, replication and formation on the mutational spectra of environmental carcinogens: Aflatoxin B1 as a case study. DNA Repair . 2018;71:12–22.

10. Lindahl T. Instability and decay of the primary structure of DNA. Nature. 1993;362:709–15.

11. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500:415–21.

12. Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S, et al. Mutational signatures associated with tobacco smoking in human cancer. Science. 2016;354:618–22.

13. Ghosal G, Chen J. DNA damage tolerance: a double-edged sword guarding the genome. Transl Cancer Res. 2013;2:107–29.

14. Tubbs A, Nussenzweig A. Endogenous DNA Damage as a Source of Genomic Instability in Cancer. Cell. 2017;168:644–56.

15. Benzer S. ON THE TOPOGRAPHY OF THE GENETIC FINE STRUCTURE. Proc Natl Acad Sci U S A. 1961;47:403–15.

16. Rubin AF, Green P. Mutation patterns in cancer genomes. Proc Natl Acad Sci U S A. 2009;106:21766–70.

17. Martincorena I, Seshasayee ASN, Luscombe NM. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. Nature. 2012;485:95–8.

18. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012;149:979–93.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

19. Dulak AM, Stojanov P, Peng S, Lawrence MS, Fox C, Stewart C, et al. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nat Genet. 2013;45:478–86.

20. Bertl J, Guo Q, Juul M, Besenbacher S, Nielsen MM, Hornshøj H, et al. A site specific model and analysis of the neutral somatic mutation rate in whole-genome cancer data. BMC Bioinformatics. 2018;19:147.

21. Stobbe MD, Thun GA, Diéguez-Docampo A, Oliva M, Whalley JP, Raineri E, et al. Recurrent somatic mutations reveal new insights into consequences of mutagenic processes in cancer. PLoS Comput Biol. 2019;15:e1007496.

22. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578:94–101.

23. Lee CA, Abd-Rabbo D, Reimand J. Functional and genetic determinants of mutation rate variability in regulatory elements of cancer genomes. Genome Biol. 2021;22:133.

24. Cheung MK, Bockrath RC. On the specificity of UV mutagenesis in E. coli. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 1970;10:521–3. doi:10.1016/0027-5107(70)90015-1.

25. Ikehata H, Ono T. The mechanisms of UV mutagenesis. J Radiat Res. 2011;52:115–25.

26. Krauthammer M, Kong Y, Ha BH, Evans P, Bacchiocchi A, McCusker JP, et al. Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. Nat Genet. 2012;44:1006–14.

27. Elliott K, Boström M, Filges S, Lindberg M, Van den Eynden J, Ståhlberg A, et al. Elevated pyrimidine dimer formation at distinct genomic bases underlies promoter mutation hotspots in UV-exposed cancers. PLoS Genet. 2018;14:e1007849.

28. Mao P, Brown AJ, Esaki S, Lockwood S, Poon GMK, Smerdon MJ, et al. ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma. Nat Commun. 2018;9:2626.

29. Lindberg M, Boström M, Elliott K, Larsson E. Intragenomic variability and extended sequence patterns in the mutational signature of ultraviolet light. Proc Natl Acad Sci U S A. 2019;116:20411–7.

30. Zhang Y, Xiao Y, Yang M, Ma J. Cancer mutational signatures representation by large-scale context embedding. Bioinformatics. 2020;36:i309–16.

31. Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat Genet. 2013;45:970–6.

32. Burns MB, Lackey L, Carpenter MA, Rathore A, Land AM, Leonard B, et al. APOBEC3B is an enzymatic source of mutation in breast cancer. Nature. 2013;494:366–70.

33. Chan K, Roberts SA, Klimczak LJ, Sterling JF, Saini N, Malc EP, et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. Nat Genet. 2015;47:1067–72.

34. Supek F, Lehner B. Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. Cell. 2017;170:534–47.e23.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

35. Nordentoft I, Lamy P, Birkenkamp-Demtröder K, Shumansky K, Vang S, Hornshøj H, et al. Mutational context and diverse clonal development in early and late bladder cancer. Cell Rep. 2014;7:1649–63.

36. Buisson R, Langenbucher A, Bowen D, Kwan EE, Benes CH, Zou L, et al. Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. Science. 2019;364. doi:10.1126/science.aaw2872.

37. Petljak M, Alexandrov LB, Brammeld JS, Price S, Wedge DC, Grossmann S, et al. Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. Cell. 2019;176:1282–94.e20.

38. Langenbucher A, Bowen D, Sakhtemani R, Bournique E, Wise JF, Zou L, et al. An extended APOBEC3A mutation signature in cancer. Nat Commun. 2021;12:1602.

39. Zou X, Morganella S, Glodzik D, Davies H, Li Y, Stratton MR, et al. Short inverted repeats contribute to localized mutability in human somatic cells. Nucleic Acids Res. 2017;45:11213–21.

40. McDaniel YZ, Wang D, Love RP, Adolph MB, Mohammadzadeh N, Chelico L, et al. Deamination hotspots among APOBEC3 family members are defined by both target site sequence context and ssDNA secondary structure. Nucleic Acids Res. 2020;48:1353–71.

41. Downing JR, Wilson RK, Zhang J, Mardis ER, Pui C-H, Ding L, et al. The Pediatric Cancer Genome Project. Nat Genet. 2012;44:619–22.

42. Priestley P, Baber J, Lolkema MP, Steeghs N, de Bruijn E, Shale C, et al. Pan-cancer whole-genome analyses of metastatic solid tumours. Nature. 2019;575:210–6.

43. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. Cell Rep. 2013;3:246–59.

44. Bayati M, Rabiee HR, Mehrbod M, Vafaee F, Ebrahimi D, Forrest ARR, et al. CANCERSIGN: a user-friendly and robust tool for identification and classification of mutational signatures and patterns in cancer genomes. Sci Rep. 2020;10:1286.

45. Schuster-Böckler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature. 2012;488:504–7.

46. Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence M, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature. 2015;518:360–4.

47. García-Nieto PE, Schwartz EK, King DA. Carcinogen susceptibility is regulated by genome architecture and predicts cancer mutagenesis. EMBO J. 2017. https://www.embopress.org/doi/abs/10.15252/embj.201796717.

48. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 2010;463:191–6.

49. Haradhvala NJ, Polak P, Stojanov P, Covington KR, Shinbrot E, Hess JM, et al. Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. Cell. 2016;164:538–49.

50. Perera D, Poulos RC, Shah A, Beck D, Pimanda JE, Wong JWH. Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. Nature. 2016;532:259–63.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

51. Woo YH, Li W-H. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. Nat Commun. 2012;3:1004.

52. Tomkova M, Tomek J, Kriaucionis S, Schuster-Böckler B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. Genome Biol. 2018;19:129.

53. Nesta AV, Tafur D, Beck CR. Hotspots of Human Mutation. Trends Genet. 2021;37:717–29.

54. Hodgkinson A, Ladoukakis E, Eyre-Walker A. Cryptic variation in the human mutation rate. PLoS Biol. 2009;7:e1000027.

55. Aggarwala V, Voight BF. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. Nat Genet. 2016;48:349–55.

56. Carlson J, Locke AE, Flickinger M, Zawistowski M, Levy S, Myers RM, et al. Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. Nat Commun. 2018;9:3753.

57. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer. 2018;18:696–705.

58. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell. 2018;173:1823.

59. Fang H, Barbour JA, Poulos RC, Katainen R, Aaltonen LA, Wong JWH. Mutational processes of distinct POLE exonuclease domain mutants drive an enrichment of a specific TP53 mutation in colorectal cancer. PLoS Genet. 2020;16:e1008572.

60. Tang C, Bagnara D, Chiorazzi N, Scharff MD, MacCarthy T. AID Overlapping and Polη Hotspots Are Key Features of Evolutionary Variation Within the Human Antibody Heavy Chain (IGHV) Genes. Front Immunol. 2020;11:788.

61. Yeap L-S, Hwang JK, Du Z, Meyers RM, Meng F-L, Jakubauskaitė A, et al. Sequence-Intrinsic Mechanisms that Target AID Mutational Outcomes on Antibody Genes. Cell. 2015;163:1124–37.

62. Lenz G, Staudt LM. Aggressive lymphomas. N Engl J Med. 2010;362:1417–29.

63. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004;306:636–40.

64. Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. 1996-2010. 2014.

65. Brash DE, Haseltine WA. UV-induced mutation hotspots occur at DNA damage hotspots. Nature. 1982;298:189–92.

66. Wang CI, Taylor JS. In vitro evidence that UV-induced frameshift and substitution mutations at T tracts are the result of misalignment-mediated replication past a specific thymine dimer. Biochemistry. 1992;31:3671–81.

67. Koh G, Degasperi A, Zou X, Momen S, Nik-Zainal S. Mutational signatures: emerging concepts, caveats and clinical applications. Nat Rev Cancer. 2021;:1–19.

68. Fredriksson NJ, Elliott K, Filges S, Van den Eynden J, Ståhlberg A, Larsson E. Recurrent

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

promoter mutations in melanoma are defined by an extended context-specific mutational signature. PLoS Genet. 2017;13:e1006773.

69. Premi S, Han L, Mehta S, Knight J, Zhao D, Palmatier MA, et al. Genomic sites hypersensitive to ultraviolet radiation. Proc Natl Acad Sci U S A. 2019;116:24196–205.

70. Qian J, Wang Q, Dose M, Pruett N, Kieffer-Kwon K-R, Resch W, et al. B cell super-enhancers and regulatory clusters recruit AID tumorigenic activity. Cell. 2014;159:1524–37.

71. Christensen S, Van der Roest B, Besselink N, Janssen R, Boymans S, Martens JWM, et al. 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. Nat Commun. 2019;10:4571.

72. Katainen R, Dave K, Pitkänen E, Palin K, Kivioja T, Välimäki N, et al. CTCF/cohesin-binding sites are frequently mutated in cancer. Nat Genet. 2015;47:818–21.

73. Kaiser VB, Taylor MS, Semple CA. Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types. PLoS Genet. 2016;12:e1006207.

74. Hornshøj H, Nielsen MM, Sinnott-Armstrong NA, Świtnicki MP, Juul M, Madsen T, et al. Pan-cancer screen for mutations in non-coding elements with conservation and cancer specificity reveals correlations with expression and survival. NPJ Genom Med. 2018;3:1.

75. Ten Hagen KG, Gilbert DM, Willard HF, Cohen SN. Replication timing of DNA sequences associated with human centromeres and telomeres. Mol Cell Biol. 1990;10:6348–55.

76. Pich O, Muiños F, Lolkema MP, Steeghs N, Gonzalez-Perez A, Lopez-Bigas N. The mutational footprints of cancer therapies. Nat Genet. 2019;51:1732–40.

77. Pich O, Muiños F, Sabarinathan R, Reyes-Salazar I, Gonzalez-Perez A, Lopez-Bigas N. Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes. Cell. 2018;175:1074–87.e18.

78. Segal E, Widom J. Poly(dA:dT) tracts: major determinants of nucleosome organization. Curr Opin Struct Biol. 2009;19:65–71.

79. Moyle-Heyrman G, Zaichuk T, Xi L, Zhang Q, Uhlenbeck OC, Holmgren R, et al. Chemical map of Schizosaccharomyces pombe reveals species-specific features in nucleosome positioning. Proc Natl Acad Sci U S A. 2013;110:20158–63.

80. Dršata T, Špačková N, Jurečka P, Zgarbová M, Šponer J, Lankaš F. Mechanical properties of symmetric and asymmetric DNA A-tracts: implications for looping and nucleosome positioning. Nucleic Acids Res. 2014;42:7383–94.

81. Suter B, Schnappauf G, Thoma F. Poly(dA·dT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters in vivo. Nucleic Acids Res. 2000;28:4083–9.

82. Hess JM, Bernards A, Kim J, Miller M, Taylor-Weiner A, Haradhvala NJ, et al. Passenger Hotspot Mutations in Cancer. Cancer Cell. 2019;36:288–301.e14.

83. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol. 2010;28:817–25.

84. ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583:699–710.

Figure 1

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holde
preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the pre
perpetuity. It is made available under aCC-BY 4.0 International license.

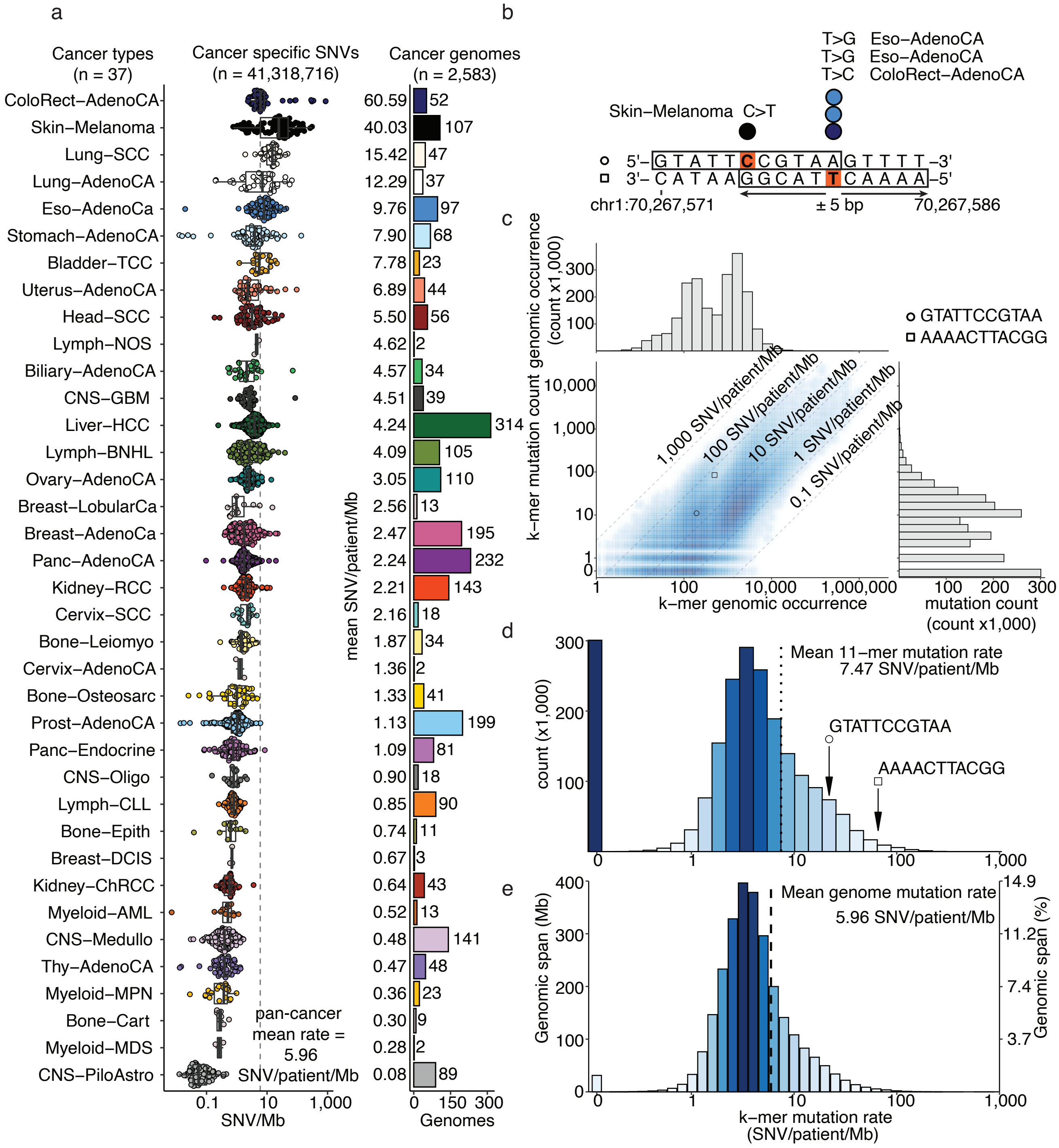


**Figure 1. Mutation data and differential mutability of 11-mers.**
(a) Non-coding mutation rate (left) and the number of cancer genomes (right) grouped and colored by cancer type. Fig. 1a may be regarded as the color legend for cancer types throughout all figures. (b) Illustration of singleton and hotspot single nucleotide variants (SNVs). Strand symmetry is assumed in the analysis and mutated base pairs are represented by their reference pyrimidines (orange). Mutations are annotated with the ±5 bp nucleotide context on the strand of the mutated pyrimidine and represented as 11-mers (framed) in the downstream analysis. (c) Distributions of 11-mer (n=2,097,090) reference (hg19) occurrence (x-axis) and pan-cancer mutation count (y-axis). (d) K-mer count distribution of 11-mer mutation rates. (e) Genomic occurrence (span) distribution of 11-mer mutation rates. K-mer span = 2,684,570,106 bp (100%)

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holde
preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the pre
perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 2.

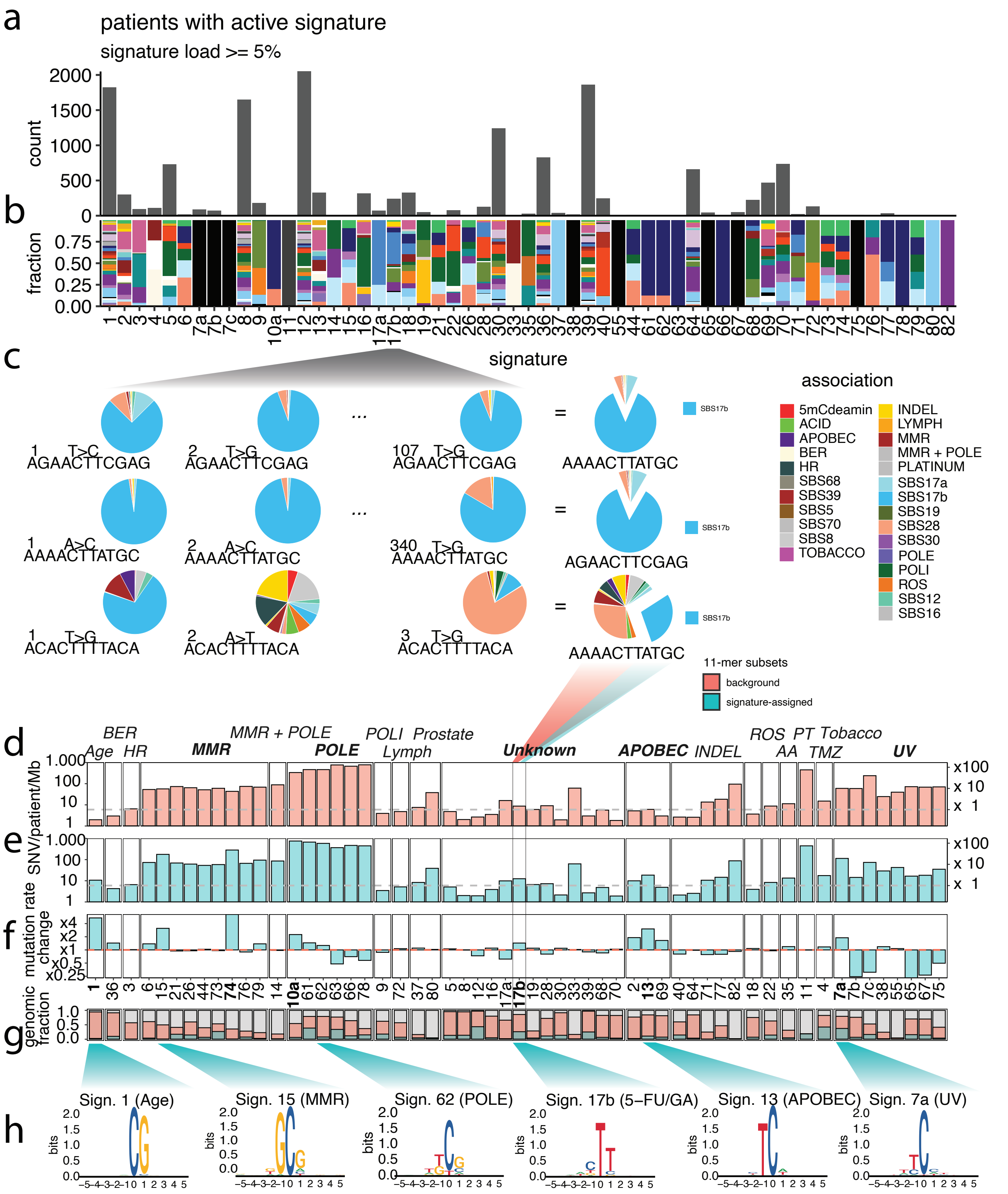


**Figure 2. Assignment of cohorts and 11-mers to mutational signatures.**
(a) Stratification of genomes based on mutational signature load into 60 so-called activity cohorts. Each activity cohort comprises a number from 0 to 2049 genomes (median 48). Cancer type color legend can be found in Figure 1a. (b) Fraction of cancer types in each activity cohort. (c) 11-mer counting across the genomes subjected to signature X. Each mutated position has a probability distribution of possible explanatory signatures. The average signature probabilities across identical 11-mers make up a probability distribution for each unique 11-mer. Hard assignment of an 11-mer to a signature is based on the max probability in the signature distribution. (d) Mean mutation rate of (red) all mutated 11-mers in an activity-cohort. The mean mutation rates (left y-axis) are compared to the global mutation rate (grey dashed line) and represented with a fold-change (right y-axis). Some of the signature-stratified subsets of 11-mers have mutation rates far exceeding the global rate, thus these signatures are extremely active in a subset of contexts. (e) Mean mutation rate of (blue) signature-assigned subset of 11-mers. The mean mutation rates (left y-axis) are compared to the global mutation rate (grey dashed line) and represented with a fold-change (right y-axis). (f) Fold-change in mutation rate from (red dashed line) all mutated 11-mers in an activity-cohort to (blue) signature-assigned subset of 11-mers. (g) Fraction of the genome spanned by 11-mers selected in each analysis step. (h) Sequence logos show emerging context from the subset of signature-assigned 11-mers, mostly capturing contexts evident from the signatures' profiles.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 3

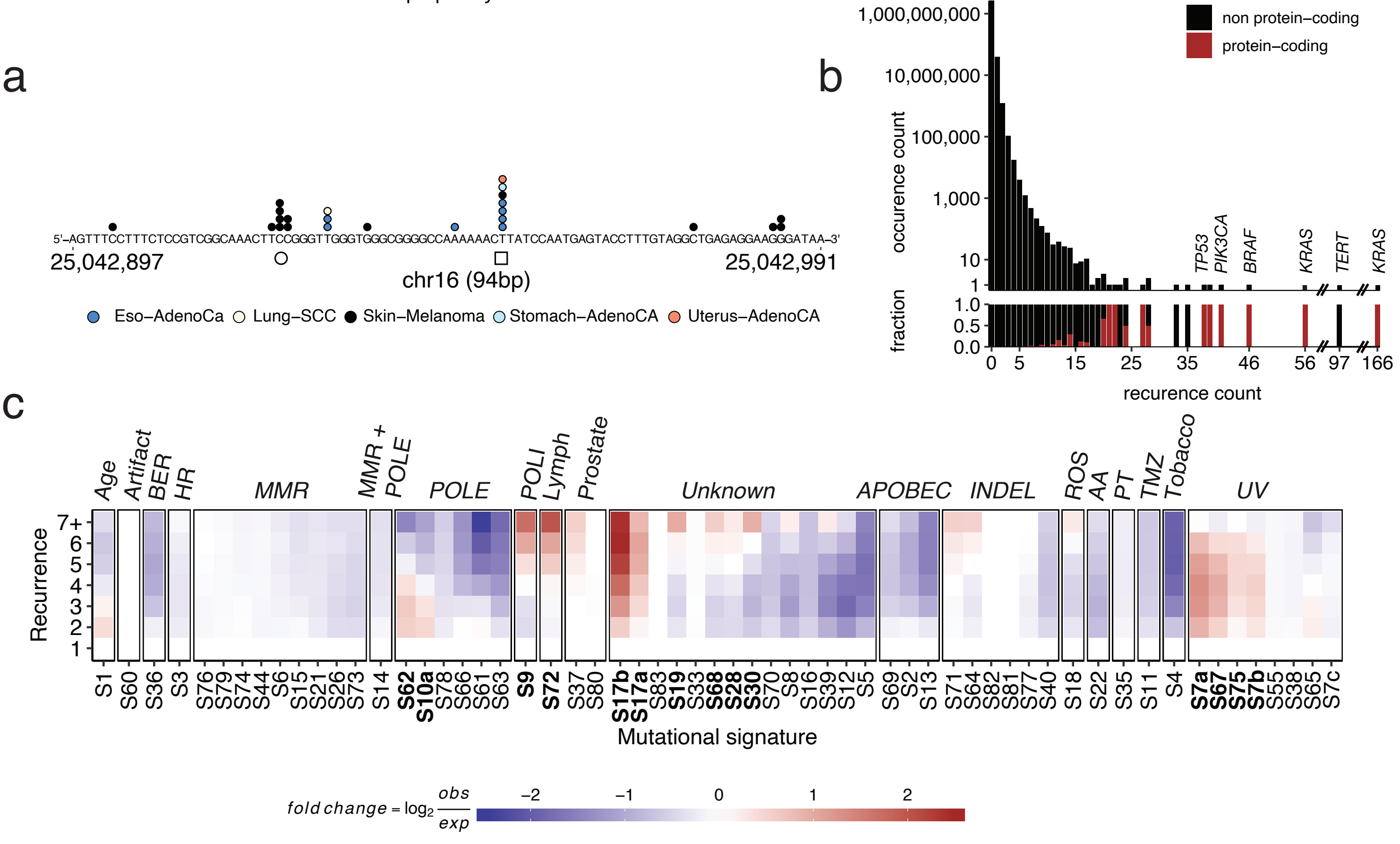
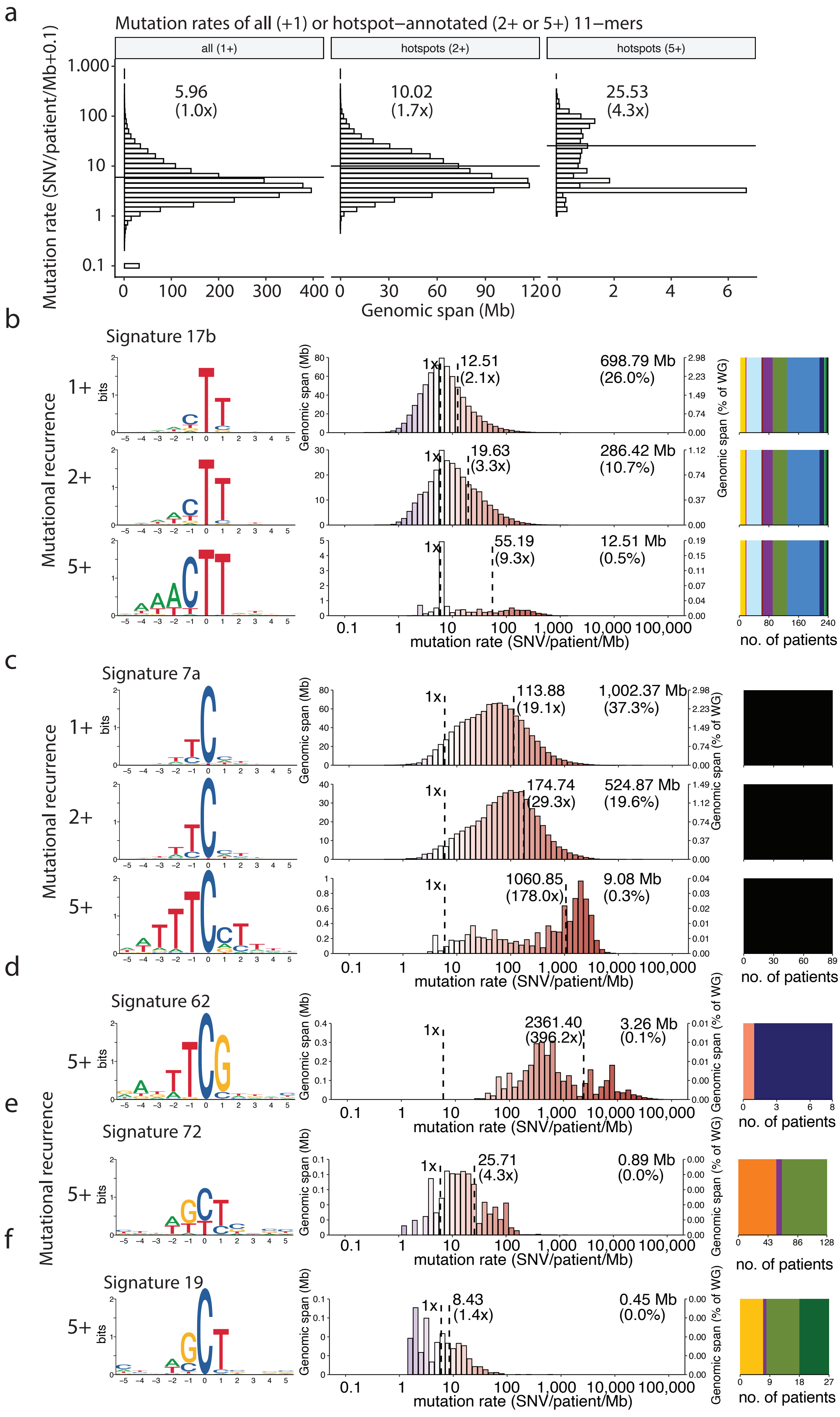


**Figure 3. Hotspot overview and identification of enriched localized mutational processes.**
(a) Pan-cancer recurrent and singleton SNVs in a 94-bp window on chromosome 16. Two cases of 11-mer instances containing hotspots. SNVs are colored by cancer type. (b) Hotspot recurrence counts (x-axis) and frequency in counts (y-axis; top) with the proportion (bottom) of positions in coding (red) or non-protein-coding regions (black). (c) All SNVs (n = 41,318,716) grouped by their pan-cancer recurrence count (1-7+). Heatmap showing the relative contribution of all mutational signatures (x-axis) to mutations of hotspots of increasing recurrence (y-axis). Colors represent log2-fold change in mean contribution relative to singleton SNVs (recurrence 1). Several mutational signatures are enriched (red) in highly recurrent hotspots (recurrence 5, 6, 7+).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

# Figure 4



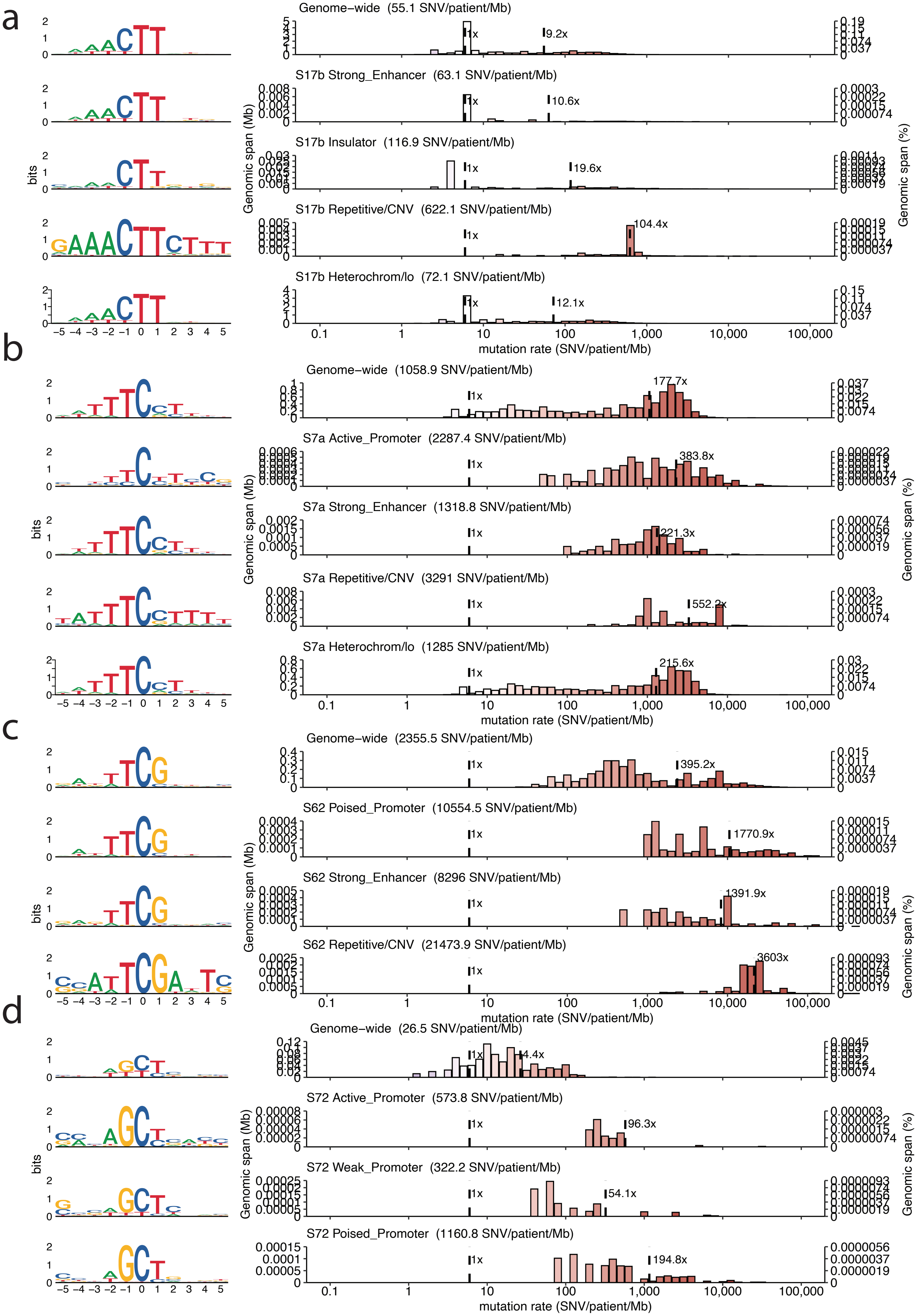
bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

**Figure 4. Mutational recurrence capture highly mutated 11-mers.**

(a) Mutation rates of all mutated 11-mers (1+; 98.8% [2,653 Mb] of the genome) and 11-mers with a hotspot in at least one of its instances for all hotspots (2+; 35.5% [954 Mb] of the genome), and highly recurrent hotspots (5+; 0.9% [23 Mb] of the genome). (b) Signature 17b-assigned 11-mers of all recurrences-levels (1+; top horizontal panels), 11-mers with a hotspot in at least one of its instances (2+; middle horizontal panels), and 11-mers with a highly recurrent hotspot in at least one of its instances (5+; bottom horizontal panels). Logo plots (left) visually highlight contexts different from the background base distribution and thus have the potential to capture sequence dependencies. Histograms (middle) show the distribution of mutation rates of the 11-mers subsets. Stacked bar plots (right) represent the cancer type distribution, colored as in Fig.1a. (c) Plots as in (b) with 11-mers assigned to UV-signature 7a. (d) POLE-signature 62-assigned 11-mers with a highly recurrent hotspot in at least one of its instances (5+). Plots as in (a). (e) Signature 72-assigned 11-mers with a highly recurrent hotspot in at least one of its instances (5+). Plots as in (a). (f) Signature 19-assigned 11-mers with a highly recurrent hotspot in at least one of its instances (5+). Plots as in (a).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

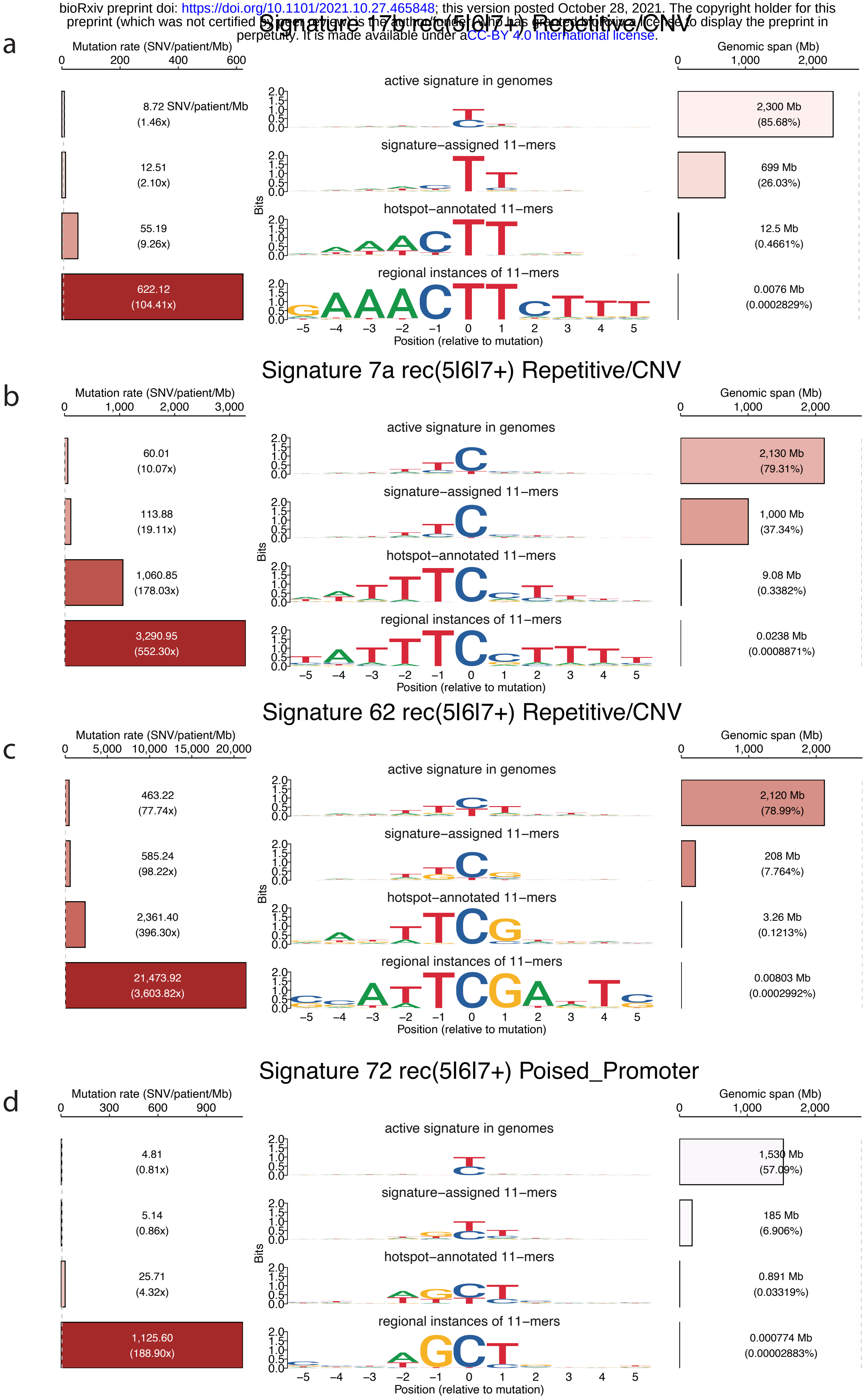
# Figure 5



bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

**Figure 5. Genomic regions with highly mutable 11-mers.**
(a) Sequence characteristics (left; logo plot) and mutation rate distribution (right; histogram) of signature 17b-assigned 11-mers of high recurrence (5+) across the whole genome (top) and most highly mutated genetic elements (below). (b) Signature 7a-assigned 11-mers of high recurrence (5+) characterized as in (a). (c) Signature 62-assigned 11-mers of high recurrence (5+) characterized as in (a). (d) Signature 72-assigned 11-mers of high recurrence (5+) characterized as in (a).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2021.10.27.465848; this version posted October 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

**Figure 6. Decomposition of factors increasing the mutation rate.**
(a) For signature 17b, each analysis step is represented by the mean 11-mer mutation rate (left; barplot), sequence characteristics (middle; logo plot), and genomic span (right; barplot). Mutation rate is quantified as SNV/patient/Mb and the increase from the baseline mutation rate (5.96 SNV/patient/Mb) is stated in parenthesis. For logo plots, letter height is measured in bits (uniform background). The genomic span is given in mega base pair (Mb) with percent of whole genome in parenthesis. (b) For signature 7a, as in (a). (c) For signature 62, as in (a). (d) For signature 72, as in (a).