bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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-NC-ND 4.0 International license.

Signatures 1 and 17 show increased propensity to create mutational hotspots in the human genome

Authors

Claudia Arnedo-Pac[1,2], Ferran Muiños[1,2], Abel Gonzalez-Perez[1,2,#] & Nuria Lopez-Bigas[1,2,3,#]

Affiliations

[1] Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain.
[2] Centro de Investigación Biomédica en Red en Cáncer, Instituto de Salud Carlos III, Madrid, Spain.
[3] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.
[#] To whom correspondence should be addressed.

Abstract

Recurrence of somatic mutations at the exact same position across patients (hotspots) are often identified as potential cancer drivers, assuming that they are unlikely to be generated by neutral mutagenesis. Recent studies have challenged this by identifying examples of mutational processes that generate passenger hotspots. However, no comprehensive study to identify and quantify the determinants of hotspots formation across tumours has been carried out to date. In this work, we conducted a systematic analysis of passenger hotspot events across more than 7,500 whole genome sequences from different malignancies. We found that mutational signatures 1 (SBS1) and 17 (SBS17a and SBS17b) have the highest propensity to form hotspots, generating 5-80 times more than other common somatic mutational processes. The trinucleotide mutational probabilities and genomic sequence composition partially explain the high SBS1 hotspot propensity. Strikingly, the vast majority of hotspots (46-96%) contributed by different signatures remain unexplained after correcting for their sequence context preferences and the large-scale mutation rate variability. This finding reveals the extension of our lack of knowledge about how mutations occur, and highlights the need of identifying and subsequently modelling additional sequence and chromatin features that influence mutation rate at base resolution. This is key to accurately modelling the mutation rate under neutrality, an essential step for identifying cancer drivers, and –given the known activity of SBS1 across other organisms and in germ cells– also for reconstructing evolutionary histories and studying genome evolution.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

Introduction

Accurately modelling mutation rate in human and other species is essential to study somatic and germline evolution. The landscape of somatic mutations in the human genome is shaped by the interplay between DNA damage and repair. These processes act differently across genomic regions influenced by multiple features of the genome organisation. At large scale, variables such as chromatin accessibility, transcriptional activity and replication timing are known to impact the distribution of mutations[1–4]. At small scale, nucleotide sequence composition is a key determinant of mutagenesis and has been exploited to characterise mutational processes through the identification of mutational signatures[5–7].

Moreover, genomic features at the local-scale level --ranging from less than 10 nucleotides to up to a few thousand base pairs-- can also influence the distribution of mutations contributed by specific mutational processes or signatures[8]. For example, clusters of somatic mutations at transcription factor binding sites in melanomas are contributed by single base substitution signature 7 (SBS7) caused by exposure to UV-light[9–12]. Similarly, mutations contributed by SBS17, a signature of unknown aetiology that has been identified in esophageal and stomach cancers, as well as in neoplasms previously exposed to capecitabine or 5-FU, has been shown to cluster at CTCF binding sites in colorectal cancers[13,14]. The formation of non-canonical DNA secondary structures, such as DNA hairpins at short inverted repeats, are localised targets of SBS2 and SBS13 caused by the APOBEC family of cytidine deaminases and microsatellite instability related signatures[15,16]. At a larger scale, APOBEC has been shown to contribute localised or scattered hypermutation clusters in processes known as *kataegis* and *omikli*, respectively[6,17,18]. The DNA wrapping around nucleosomes is also known to affect the distribution of mutations contributed by different signatures, including SBS7 and SBS17 [19].

Less is known about the determinants of the extreme case of mutation clustering: mutational hotspots, defined as recurrent mutations affecting the exact same genomic position across cancers. They have been analysed mostly in the context of the identification of signals of positive selection in cancer[20–22], and it is often assumed that they indicate the presence of driver mutations. Nevertheless, recent studies have challenged this assumption by showing a high frequency of passenger hotspots across cancers[23–25]. There is thus a need to characterise the mutational processes and genomic determinants involved in the formation of passenger hotspots, as a key to reliably identifying driver mutations. In this line, it has been shown that UV-light damage can preferentially target specific residues within the binding sites of ETS-family of transcription factors, generating hotspots[26,27]. It is also known that hotspots of APOBEC3A mutations are more likely to appear at particular positions in ssDNA loops within DNA hairpins[24,28,29]. Although these studies demonstrate that mutational processes --without the action of positive selection-- can create passenger hotspots, no large-scale analysis to broadly characterise and quantify hotspots, and to explore the processes underlying their formation has

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

been carried out to date. The knowledge of mutational processes is essential to correctly model the mutation rate under neutrality. This, in turn, is key to accurately identifying drivers of tumorigenesis across cancer samples, modelling evolutionary trajectories, assessing the impact of variants in the human germline, and understanding the evolution of genomic sequences. It is within this framework that a comprehensive characterization of the underlying determinants of hotspots mutagenesis is of interest.

In the present work, we leveraged somatic mutations of more than 7,500 whole genome sequences of tumours from 49 cancer types, and we systematically detected and quantified the mutational processes creating passenger hotspots. We discovered that mutational processes active across tumours exhibit very different propensities to form hotspots, with signatures 1 (SBS1) and 17 (SBS17a and SBS17b) generating 5-80 times more hotspots than other common mutational processes considered in our study (SBS2, SBS3, SBS4, SBS5, SBS7a-b, SBS8, SBS13, SBS18, SBS40, and SBS93). A part of the higher propensities of these two signatures to form hotspots are explained by the specificity of their activity across trinucleotides, the trinucleotide composition of the human genome and the megabase-scale variability of the mutation rate. However, a large proportion (74-96%) of the hotspots attributed to SBS1, SBS17a and SBS17b remain unexplained after correcting for these features. Other small scale sequence and chromatin features may thus play an important role in hotspot formation. We explore CTCF binding sites in the case of SBS17, and CpG islands for SBS1, and find that although those regions accumulate significantly more hotspots than their neighbouring areas, they account for only a tiny proportion of the unexplained hotspots along the genome. This leaves other, mostly unknown, DNA sequence and chromatin features as contributors to the high propensity of these two signatures to form hotspots across tissues. Altogether, our results show that mutational processes, particularly SBS1 and SBS17, have an unexpected propensity to create passenger hotspots. This finding highlights our current difficulty to accurately estimate the mutation rate –specifically contributed by these signatures– under neutrality at nucleotide resolution, and stresses the need for identifying and subsequently modelling additional small scale features that influence it.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

Results

Mutational hotspots across cancers

We collected and filtered publicly available whole genome sequencing (WGS) data from 7,507 tumours comprising 83,410,018 somatic mutations (SNVs and short indels) from primary and metastatic cancers (Fig. 1a-c; Supplementary Table S1; Methods). Samples were classified into 49 different cancer types, 8 of them meta-groups, following the Memorial Sloan Kettering Cancer Center (MSKCC) OncoTree hierarchy[30] (Fig. 1a; Supplementary Table S1 and S2). Since our objective was the study of the distribution of mutational hotspots along the genome under neutrality, we removed mutations overlapping the coding sequence of known cancer driver genes[31,32] and their surrounding non-coding regions (Fig. 1e; Supplementary Table S3; Methods). Next, we identified hotspots of somatic mutations across the samples of each cancer type using a new method named HotspotFinder. Briefly, HotspotFinder identifies and annotates unique genomic positions that are recurrently mutated (two or more times) to the same alternate (e.g., two C>T transitions) across tumours by independently analysing single nucleotide variants (SNVs), multi-nucleotide variants (MNVs), small insertions and deletions (Fig. 1d; Methods; Supplementary Note 1). HotspotFinder may also be used to identify hotspots of mutations with different alternates (e.g., C>T and C>G; Supplementary Note 1).

A total of 1,562,007 alternate specific hotspots of four different types of mutations were identified across individual cancer types (3,106,182 across the pan-cancer cohort): 1,361,633 corresponded to SNVs (87.2%), 125,658 to deletions (8.0%), 72,892 to insertions (4.7%), and 1,824 to MNVs (0.1%) (Fig. 1f-g; Supplementary Table S4). Hotspots covered approximately 0.06% of the mappable hg38 reference genome (approximately 2,531 Mbps; see Methods) and the vast majority (99.44%) were located in non-coding regions. The largest number of hotspots (n=1,095,121) were observed across skin melanomas, followed by colorectal (n=165,473 hotspots) and esophageal cancers (n=106,025 hotspots) (Fig. 1f). In all cancer types except retinoblastomas at least one hotspot was observed. The majority of hotspots were small, comprising 2 or 3 mutated samples (Fig. 1h, Supp. Fig. 1), although we observed a few exceptions (Fig. 1h, Supp. Fig. 1-2). Hotspots of insertions and deletions were particularly abundant (at similar or greater rates than SNVs hotspots) across cancer types with active indel mutational processes (i.e., 19.3% and 30.1% insertion and deletion hotspot frequency in colorectal tumours) (Fig. 1f-g, Supp. Fig. 3). High rates of hotspots of insertions and deletions across other tumour types with few samples may be due to other mutational processes, or sequencing and/or calling errors and biases across cohorts as shown in other studies[33]. Given that the vast majority of hotspots identified are composed of SNVs, we decided to focus on the study of their formation. Henceforth, we use hotspot as synonymous with SNV hotspot.

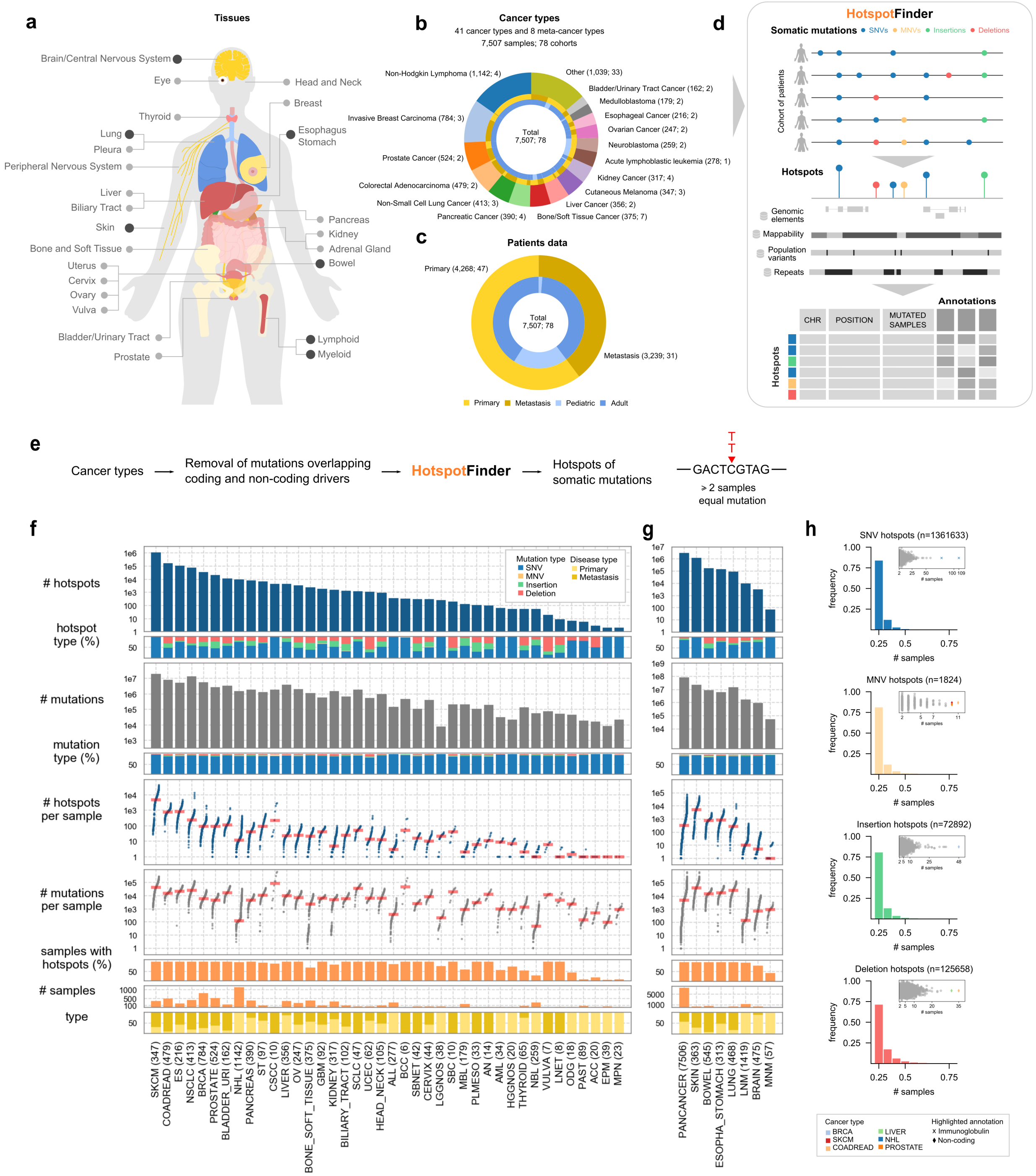
As expected, the number of hotspots per cancer type increased with sample size and mutation burden (Spearman's R=0.93, p=9e-18 and R=0.72, p=2e-7, respectively; Supp Fig. 4). However, we noticed that hotspots formed at different rates across tumour types (Fig. 2a; Supp.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

Fig. 5). In order to quantify and compare hotspots formation across cancer types, we calculated the number of SNVs required to generate one hotspot (hotspot conversion rates) in each group (Methods). We found large variability in hotspot conversion rates across malignancies, ranging from 21 mutations in melanomas to 3,036 mutations in medulloblastomas (Fig. 2b). After melanomas, the lowest conversion rates were observed across esophageal, colorectal, bladder-urinary and non-small cell lung cancers (36, 208, 333, and 569 mutations, respectively) (Fig. 2b). Altogether, these findings indicate that the rate of hotspot formation differs more than 144 fold across cancer types.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

Figure 1



bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

Fig. 1. Identification of hotspots across cancers. a) Cancer types analysed depicting specific and meta-cancer types in light and dark grey, respectively. **b**) Number of patients and sequencing cohorts among specific cancer types. **c**) Detail of the number of primary, metastatic, adult and paediatric tumours analysed. **d**) Schematic overview of HotspotFinder, a new algorithm to identify hotspots of somatic mutations (Methods; Supplementary Note 1). **e**) Overview of the steps for hotspots identification. **f**) Summary of total hotspots identified across specific cancer types and **g**) meta-cancer types. **h**) Histograms of hotspot size (number of mutated samples per hotspot) considering only hotspots from specific cancer types. Embedded dotplots show hotspot sizes per individual hotspot, where the shape and the colour represents the overlapping genomic element and the cancer type where the hotspot was identified, respectively. Cancer types are listed as follows: Acute Lymphoblastic Leukemia (ALL), Acute Myeloid Leukemia (AML), Adrenocortical Carcinoma (ACC), Anal Cancer (AN), Basal Cell Carcinoma (BCC), Biliary Tract (BILIARY_TRACT), Bladder/Urinary Tract (BLADDER_URI), Bone/Soft Tissue (BONE_SOFT_TISSUE), Bowel (BOWEL), CNS/Brain (BRAIN), Cervix (CERVIX), Colorectal Adenocarcinoma (COADREAD), Cutaneous Melanoma (SKCM), Cutaneous Squamous Cell Carcinoma (CSCC), Endometrial Carcinoma (UCEC), Ependymoma (EPM), Esophageal cancer (ES), Esophagus/Stomach (ESOPHA_STOMACH), Glioblastoma Multiforme (GBM), Head and Neck (HEAD_NECK), High-Grade Glioma NOS (HGGNOS), Invasive Breast Carcinoma (BRCA), Kidney (KIDNEY), Liver (LIVER), Low-Grade Glioma NOS (LGGNOS), Lung (LUNG), Lung Neuroendocrine Tumor (LNET), Lymphoid Neoplasm (LNM), Medulloblastoma (MBL), Myeloid Neoplasm (MNM), Myeloproliferative Neoplasms (MPN), Neuroblastoma (NBL), Non-Hodgkin Lymphoma (NHL), Non-Small Cell Lung Cancer (NSCLC), Oligodendroglioma (ODG), Ovarian Cancer (OV), Pancreas (PANCREAS), Pilocytic Astrocytoma (PAST), Pleural Mesothelioma (PLMESO), Prostate (PROSTATE), Retinoblastoma (RBL), Skin (SKIN), Small Bowel Cancer (SBC), Small Bowel Neuroendocrine Tumor (SBNET), Small Cell Lung Cancer (SCLC), Stomach cancer (ST), Thyroid (THYROID), Vulva (VULVA).

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

Figure 2

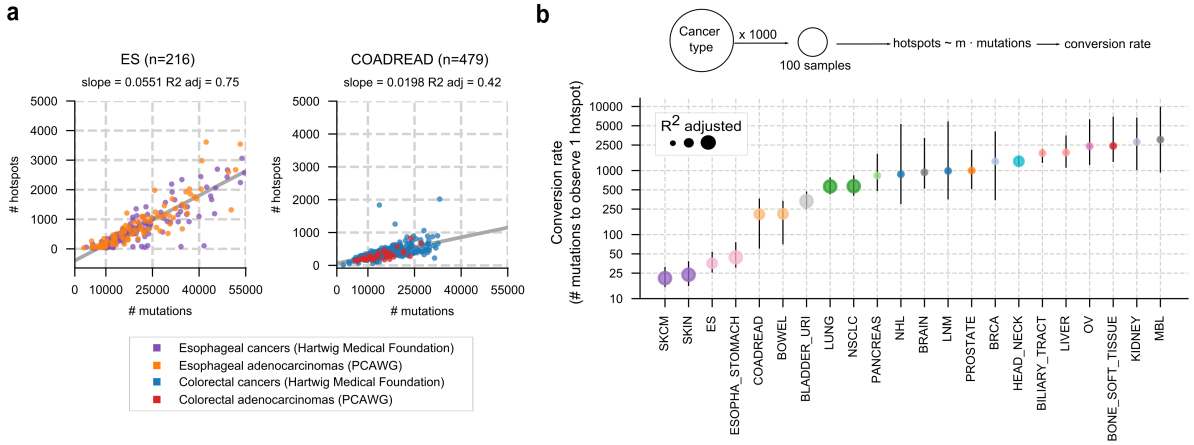


Fig. 2. Cancer types differ in their tendency to form hotspots. a) Scatter plots showing hotspot burden versus mutation burden per sample across the full set of esophageal and colorectal cancers. Fitted regression lines are shown in grey (Methods). **b**) Median conversion rates (number of mutations to observe 1 hotspot) across cancer types with more than 100 samples and at least 750 significant linear models across random replicates (mean=978.9 models per cancer type, range=777-1000; Methods). Dot size represents the goodness of fit of the linear model (R^2 adjusted). Error bars show the range of conversion rates across significant replicates.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

Propensity of mutational processes to form hotspots

We reasoned that mutational processes with different activity across tumour types would present different propensity to form mutational hotspots, thus underlying the observed differences in conversion rates. Supporting this hypothesis, we found that, across cancer types, hotspots were differentially enriched for different types of nucleotide changes, and within particular trinucleotide sequences (Supp. Fig. 6-7). In order to identify the mutational processes contributing hotspots, we conducted a de novo extraction of single base substitution (SBS) signatures in the 96 pyrimidine-centred trinucleotide contexts and decomposed the extracted signatures into their component COSMIC v3.2 GRCh38 SBS signatures to facilitate comparisons across tumour types (Fig. 3a; Supp. Fig. 8; Methods; Supplementary Note 2).

We then estimated the number of hotspots generated by each COSMIC mutational signature in each tumour type (Fig. 3b; Supp. Fig. 9; Methods). SBS1 appeared as an important contributor of hotspots across all cancer types, particularly in tumours of the brain (69.6% of hotspots), pancreas (60.9%; Fig. 3b), prostate (42.3%) and colorectum (39.7%; Fig. 3b), which is in agreement with the observed enrichment of hotspots in C>T transitions in the NpCpG context in these malignancies (Supp. Fig. 6-7). We found SBS17a and SBS17b to contribute a large proportion of hotspots in esophageal cancers (16.8 and 72.3% of hotspots, respectively), stomach tumours (12.1 and 67.4%), and, to a lesser degree, to those in colorectal cancers (1.51 and 18.6%) (Fig. 3b; Supp. Fig. 9). This is in accordance with hotspot enrichment of T>G transversions and T>C transitions, specially within CpTpT contexts in gastrointestinal tumours (Supp. Fig. 6-7). SBS5, an age-related signature of unknown aetiology ubiquitous across cancer types, SBS2 and SBS13 (APOBEC-related), SBS4 (related to tobacco smoking), and SBS7a (caused by UV light damage) were responsible for a high proportion of hotspots of specific cancer types (Fig. 3b; Supp. Fig. 9). Considering that some mutational processes are known to be better defined by extended contexts, e.g., pentanucleotides[7], we checked whether the hotspots contributed by different signatures show preferences for certain nucleotides in a 21 bp-wide window (Supp. Fig. 10). A small preference for specific nucleotides within this extended context (some of them already known) are apparent for some signatures, such as SBS7a and b[7], SBS17a and b[34,35], and SBS93. In every case, however, the contribution of the trinucleotide sequence is clearly stronger than that of the extended context, suggesting that the latter would only play a smaller role in the formation of hotspots, similarly as others have found for overall mutations[7].

While the burden of hotspots contributed by each mutational signature showed, as expected, a good correlation with its activity and the proportion of samples at which it was found active (Supp. Fig. 11), neither of these variables explain the differences in conversion rate observed across malignancies (Fig. 2b). We thus set out to estimate the propensity of signatures to form hotspots --that is, their intrinsic inclination to contribute hotspots, independent of their overall contribution to the mutation burden. We selected 14 mutational signatures with high activity in at least one cancer type (detailed criteria in Methods) and re-ran hotspot identification upon

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

subsampling a fixed number of mutations contributed by each of them. Across a range of 10,000-30,000 mutations sampled from 100 tumours (at equal number of mutations per sample), we observed approximately 1 to 2 orders of magnitude more hotspots contributed by SBS17b, SBS17a and SBS1 than by the other eleven mutational signatures studied across tumour types (Fig. 3c-f). Specifically, for 30,000 mutations across 100 samples, a median of 79, 73 and 40 hotspots were observed for SBS17b, SBS17a and SBS1, respectively (Fig. 3f). Conversely, SBS7a, SBS18, SBS2, SBS93, SBS8, SBS7b, and SBS13 contributed 4-7 hotspots, and SBS5, SBS40, SBS3 and SBS4 generated 1-2 hotspots (Fig. 3f). That is, SBS17b, SBS17a and SBS1 contributed 5-80 times more hotspots than the other signatures under the same conditions. In an orthogonal calculation of the propensity to form hotspots employing the fold change of mutations contributed by each signature across tumours inside and outside hotspots, we obtained very similar results (Supp. Fig. 12,13; Supplementary Note 3). In summary, SBS17b, SBS17a and SBS1 show the highest propensity to form hotspots among mutational processes commonly active in human tissues.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

Figure 3

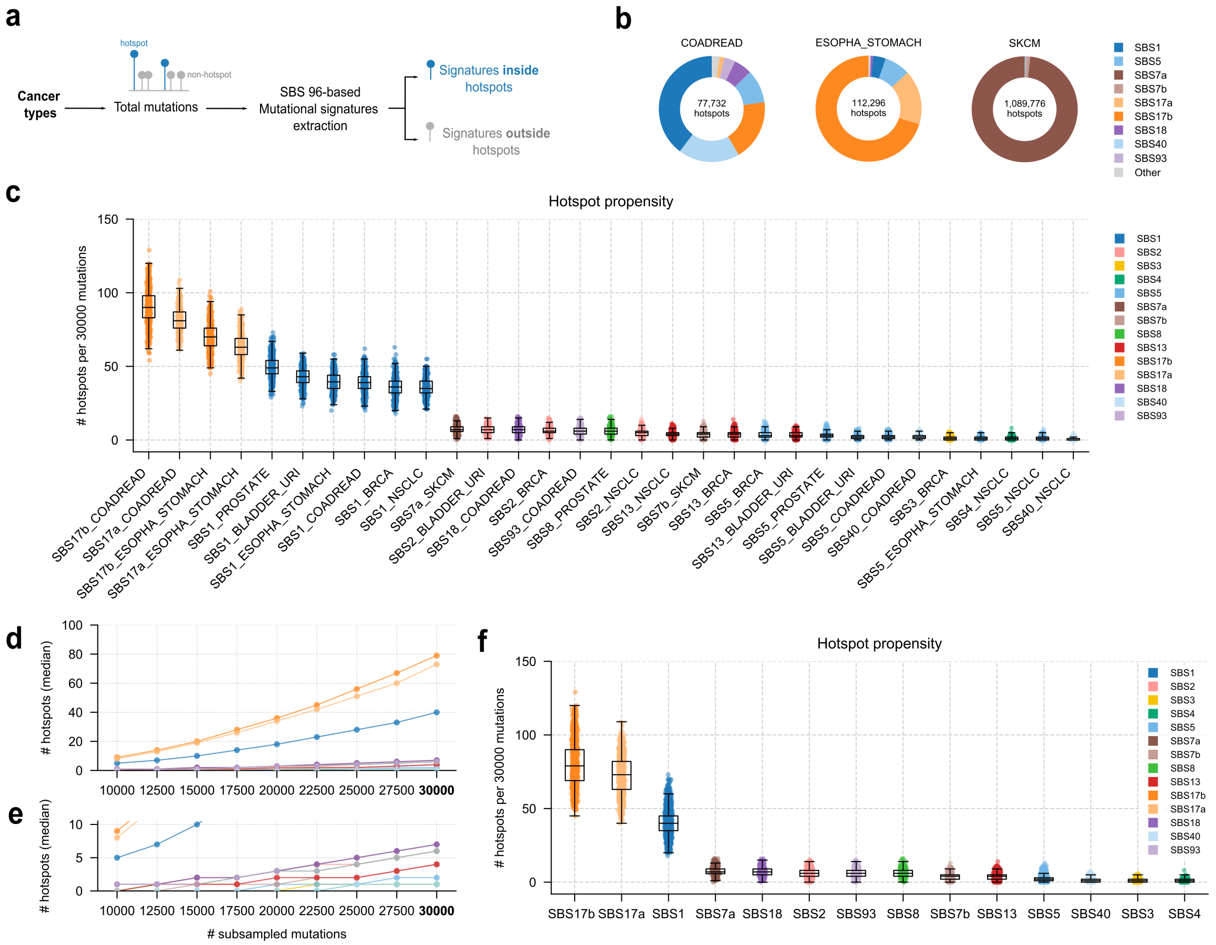


Fig. 3. Mutational processes with increased propensity to form hotspots. a) Graphical definition of the analysis of mutational signatures in the sets of mutations inside and outside hotspots. **b**) Pie charts depicting the relative number of hotspots observed per signature in the cancer type. **c**) Number of hotspots per signature and cancer type observed by subsampling 30,000 total mutations (300 mutations/sample, 100 samples) within each group in the set of mappable megabases (Methods). Cancer type-signature pairs are sorted by descending mean number of observed hotspots. **d**) Median number of observed hotspots per signature across subsamples containing at different mutation burdens (100-300 mutations/sample, 100 samples) (Methods). In order to obtain signature-level estimates, subsamples across different cancer types were merged as listed in Methods. **e**) Zoom into 0-10 hotspots from d. **f**) Number of hotspots per signature observed within 30,000 subsampled mutations (300 mutations/sample, 100 samples) across cancer types merging data shown in c. Error bars show 1.5 times the IQR below and above 1st and 3rd quartiles, respectively. Signatures are sorted in descending order according to the mean number of observed hotspots.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

Signature profile unevenness and trinucleotide abundance affect hotspot formation

We hypothesised that one possible reason for the disparity in propensity to form hotspots of different mutational processes could be the number of genomic positions available to each of them. This is determined, in the first place, by the particular trinucleotide mutational probabilities of the signature that give rise to a particular shape (i.e., skewness or unevenness vs uniformity or evenness) of its trinucleotide profile. We measured the unevenness of the activity of signatures across trinucleotides as the entropy of their mutational profile normalised by their abundance in the human genome, which informs about the mutational probability of the signature across trinucleotides irrespective of the genome composition. The more uneven the profile of a signature, the lower its entropy (Fig. 4a; Supp. Fig. 14a). The three signatures with the highest propensity to form hotspots across tissues (SBS1, SBS17b and SBS17a) showed a low entropy profile (Fig. 4b; Methods). In general, signatures with even profiles, such as SBS3 and SBS5, did not display a high propensity to form hotspots. This can be explained by the fact that the fewer active trinucleotides in a mutational signature (the more uneven its profile), the more likely it is that two mutations contributed by the process map to the same genomic position (Fig. 4b).

The availability of the 96 trinucleotides in the human genome must also influence the propensity of different mutational signatures to form hotspots. To investigate the combined effect of genomic trinucleotide abundance and signatures prolife unevenness, we determined the theoretical (or expected) number of hotspots formed by mutations contributed by several processes (see details in Methods and Supplementary Note 4). The expected number of hotspots contributed by different mutational signatures differed from those observed for the same signatures in two important aspects (Fig. 4c). First, 4.3-75.6 times more hotspots were expected to be contributed by SBS1 than by any other signature at equal sample size and mutation rate (Fig. 4c; Supp. Fig. 14b). SBS17b and SBS17a fell to the second and third position of the ranking under this theoretical model, with 5.8 times fewer expected hotspots than SBS1 (Fig. 4c; Supp. Fig. 14b). The four trinucleotides with highest activity of SBS1 (NpCpG) are comparatively depleted in the human genome (0.4-0.5 % NpCpG vs 1.9-7.9 % non-NpCpG in mappable bins) (Fig. 4d; Supp. Fig. 14c). Conversely, the NpTpT and CpTpN trinucleotides, which concentrate the activity of SBS17b and SBS17a, respectively, show average (or above average) representation in the human genome (Fig. 4d). This difference in trinucleotides availability thus explains why SBS1 has the highest expected propensity to form hotspots (Fig. 4b-c) and suggests that other factors apart from trinucleotide sequence composition underlie the observed propensity of SBS17.

Secondly, for all mutational signatures analysed, we observed more hotspots than expected, with the highest differences for SBS17b, SBS17a and SBS1 (Fig. 4e; Methods). Since the number of hotspots computed via the theoretical model only account for the unevenness of the mutational profile of the signature and the abundance of trinucleotides in the genome, other factors must be at play in the generation of the observed number of hotspots.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

Figure 4

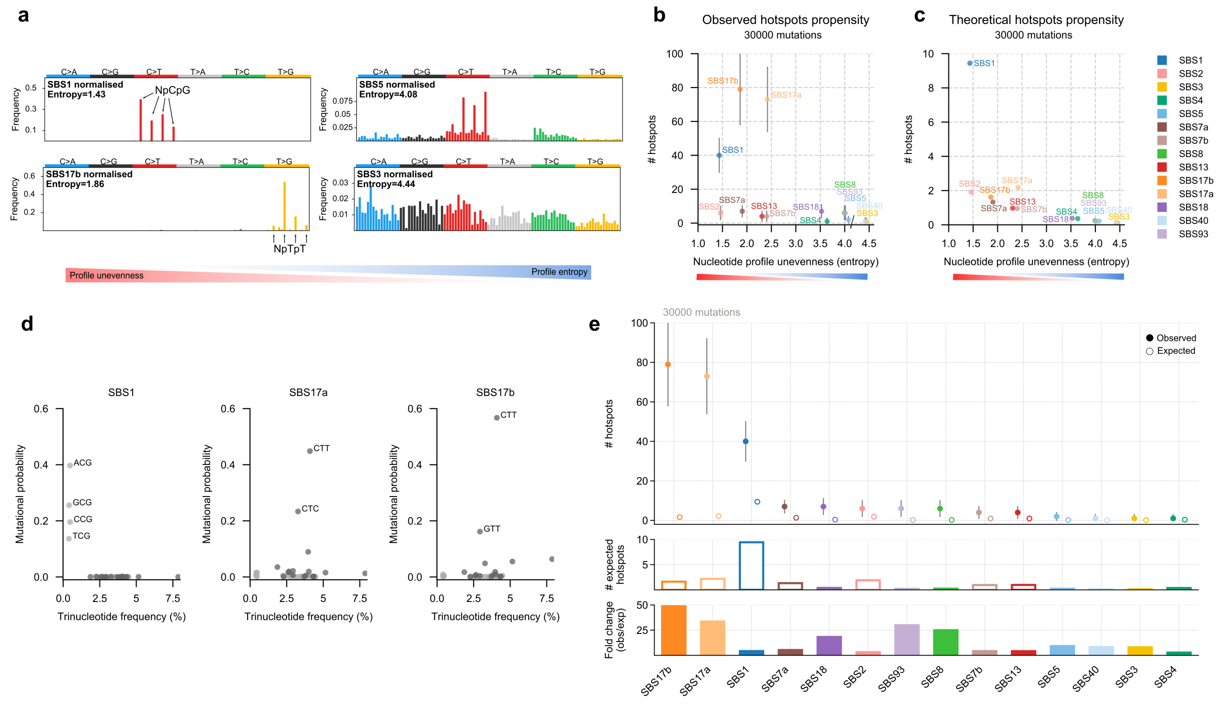


Fig. 4. Contribution of trinucleotide mutational probabilities and trinucleotide abundance to hotspots formation across signatures. a) Normalised trinucleotide profiles of SBS1, SBS17b, SBS5 and SBS3; additional signatures are shown in Supp. Fig. 14a. **b**) Number of observed hotspots (y axis) versus the entropy of the normalised signature profile (x axis). Observed hotspot propensity was computed by subsampling 30,000 total mutations (300 mutations/sample, 100 samples) within mappable megabases (Methods). Error bars show the interquartile ranges (IQR) of the number of hotspots observed across subsamples. **c**) Number of expected hotspot propensity (y axis) versus the entropy of the normalised signature profile (x axis). Expected data was generated using the trinucleotide model of hotspots formation for 300 mutations/sample and 100 samples across mappable megabases, therefore it is comparable to the data in b. **d**) Mutational probability of trinucleotides per signature versus their frequency within mappable megabases. The mutational probability for each trinucleotide was obtained by merging those from the respective three alternates given by the normalised signature profile. **e**) Comparison of observed versus expected hotspot propensity per signature (top-middle) and the fold change of observed versus expected number of hotspots (bottom). Observed dots show median hotspots; error bars show the IQR. Observed and expected data correspond to that shown in b and c.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

The effect of large scale chromatin features

We reasoned that another factor underlying the differences in hotspot propensity across signatures could be the uneven distribution of their mutations along the human genome. Several megabase scale chromatin features, such as replication time and chromatin compaction are known to play a role in this variability of the mutation rate along the genome[1–4,36–38]. To compute if the distribution of mutations at the megabase scale influences the propensity of different signatures to form hotspots, we counted the number of mutations and hotspots contributed by each signature in autosomal mappable bins of length 1 Mbp. We found that the density of hotspots was positively correlated with that of mutations at the megabase scale (Supp. Fig. 15), as illustrated along chromosome 1 for SBS1 and SBS17b in colorectal cancers (Fig. 5a). As expected, hotspot density per megabase correlated with chromatin accessibility, replication time and level of transcription (Fig. 5a; Supp. Fig.16-18).

In order to quantify the potential effect of the unevenness of the genomic distribution of a signature in the formation of hotspots, we first computed the overdispersion of the distribution of the number of mutations contributed at different megabase bins along the genome (Fig. 5b; Methods). SBS1 mutations exhibit low overdispersion across megabase genomic bins. Others, like SBS17b, show large variability in mutation counts at the megabase scale (Fig. 5b; Supp. Fig. 19). Actually, SBS1, SBS17a and SBS17b, the mutational processes with the highest propensity to form hotspots, appear at opposite ends of the spectrum of megabase mutation overdispersion (Fig. 5c). SBS17a and SBS17b exhibit the highest megabase mutation rate unevenness, followed by SBS18 and SBS4 (Fig. 5c). While differences in the interplay of mutational processes with chromatin features may underlie the dissimilar unevenness observed across signatures, it is worth noticing that the NpTpT trinucleotides targeted by SBS17a and SBS17b show a greater inter-megabase variability than the NpCpG targeted by SBS1 (Supp. Fig. 14d).

Next, we compared the number of hotspots observed across 1 Mb segments of the genome with that expected after accounting for the megabase distribution of mutations and the trinucleotide composition of each segment (Fig. 5d; Methods). The observed-to-expected hotspot fold-change is still greater than 1 for all signatures, as was the case when only the signature profile and the trinucleotide composition were taken into account (Fig. 4e). Nevertheless, the expected number of hotspots (in particular for SBS17a and SBS17b, those with highest megabase mutation rate unevenness) is higher than that in Figure 4e. In other words, the megabase-scale distribution of mutations contributed by different signatures influences their propensity to form hotspots, and this influence appears clearer the higher the unevenness of their mutation rate along the genome. However, an important part of the propensity of signatures to form hotspots (especially in those with higher propensity) remains unexplained.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

Figure 5

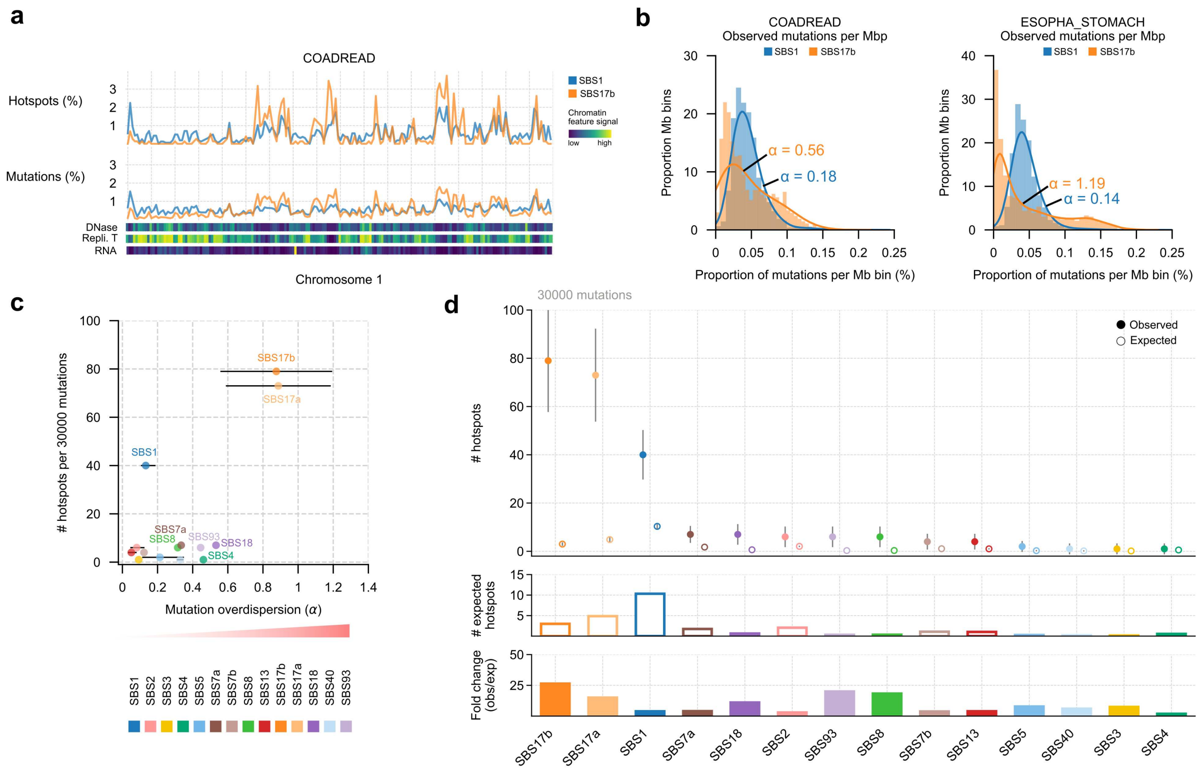


Fig. 5. Contribution of large scale unevenness to hotspots formation across signatures. a) Proportion of observed hotspots (top) and mutations (bottom) in colorectal cancers attributable to SBS1 and SBS17b across mappable megabases of chromosome 1. Normalised epigenomic signals of chromatin accessibility, replication timing and expression per megabase are shown below. **b**) Distribution of the observed proportion of mutations across mappable megabases. Alpha values show the overdispersion of the negative binomial distribution fitted with the mutation counts per megabase (Methods). **c**) Number of observed hotspots versus the overdispersion (unevenness) of mutation counts within genomic megabases. Observed hotspots were computed by subsampling 30,000 total mutations (300 mutations/sample, 100 samples) within mappable megabases as shown in Fig. 4b. **d**) Comparison of observed versus the expected theoretical number of hotspots per signature (top-middle) calculated with the trinucleotide-megabase model accounting for trinucleotide composition and large-scale mutation rate variability. Expected data was generated for 300 mutations/sample and 100 samples across mappable megabases. Dots show the median number of hotspots across cancers. Error bars correspond to the IQR. The fold change of observed versus expected number of hotspots is shown below.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

Missing factors explaining the number of observed hotspots

We quantified the proportion of hotspots contributed by different signatures that remain unexplained after accounting for the signature profile unevenness, the trinucleotide composition of the human genome, and the differential megabase-scale distribution of mutations. On average across 14 mutational signatures, 81.2% (range=46.1-96.2%) of observed hotspots remain unexplained after taking these factors into account (Fig. 6a). The fraction of unexplained hotspots is particularly sizable for SBS17a and SBS17b, where only 6.7% and 3.8% hotspots, respectively, are explained by theoretical models that accommodate previously mentioned covariables of hotspot generation. The proportion of SBS1-contributed hotspots explained by these factors is larger (25.8%), but still far from their observed number (Fig. 6a). This suggests that other sequence and chromatin features below the megabase-scale play a significant role in hotspots formation and their distribution across the genome (Supp. Fig. 20).

Some chromatin features below the megabase scale are known to interact with different mutational processes influencing their local activity[8]. We thus explored the effect of known small scale chromatin covariates on the formation of SBS1 and SBS17a and SBS17b hotspots. For example, SBS17a and SBS17b hotspots appear increased in colorectal (14.78 and 14.79 times) and esophageal-stomach tumours (3.82 and 3.21 times) at CTCF binding sites with respect to their flanking sequences, significantly beyond the expectation from their sequence composition (Fig. 6b,c; Supp. Fig. 21). These results are consistent with CTCF binding sites bearing clusters of SBS17 mutations[13,14,38]. Similarly, the rate of SBS1 hotspots at CpG islands is, on average, four times higher than their flanking regions across tumour types (mean=4.44, range=3.27-5.16), and also above the rate expected from their sequence composition (Fig. 6e,f; Supp. Fig. 22). This indicates that non-sequence based local features affect SBS17a, SBS17b and SBS1 hotspot formation at CTCF binding sites and CpG islands, respectively.

Importantly, only 0.8% and 1.7% of SBS17a hotspots, and 0.5% and 1.3% of SBS17b hotspots in esophageal-stomach and colorectal cancers, respectively, overlap CTCF binding sites (Fig. 6d), although the majority of CTCF-overlapping hotspots in these two tumour types (86% and 53.5%) are contributed by these two signatures together. Similarly, only 2% of SBS1 hotspots are located within CpG islands across cancer types (range=1.4-2.5%; Fig. 6g), despite the fact that between 15% (breast tumours) and 82% (colorectal adenocarcinomas) of CpG island-overlapping hotspots are attributed to SBS1.

In summary, the results presented here highlight that the observed hotspot propensity remains by large unexplained by currently known factors affecting mutation rates, stressing the need for identifying the broad set of sequence and chromatin features governing hotspots formation to accurately measure neutral mutagenesis.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

Figure 6

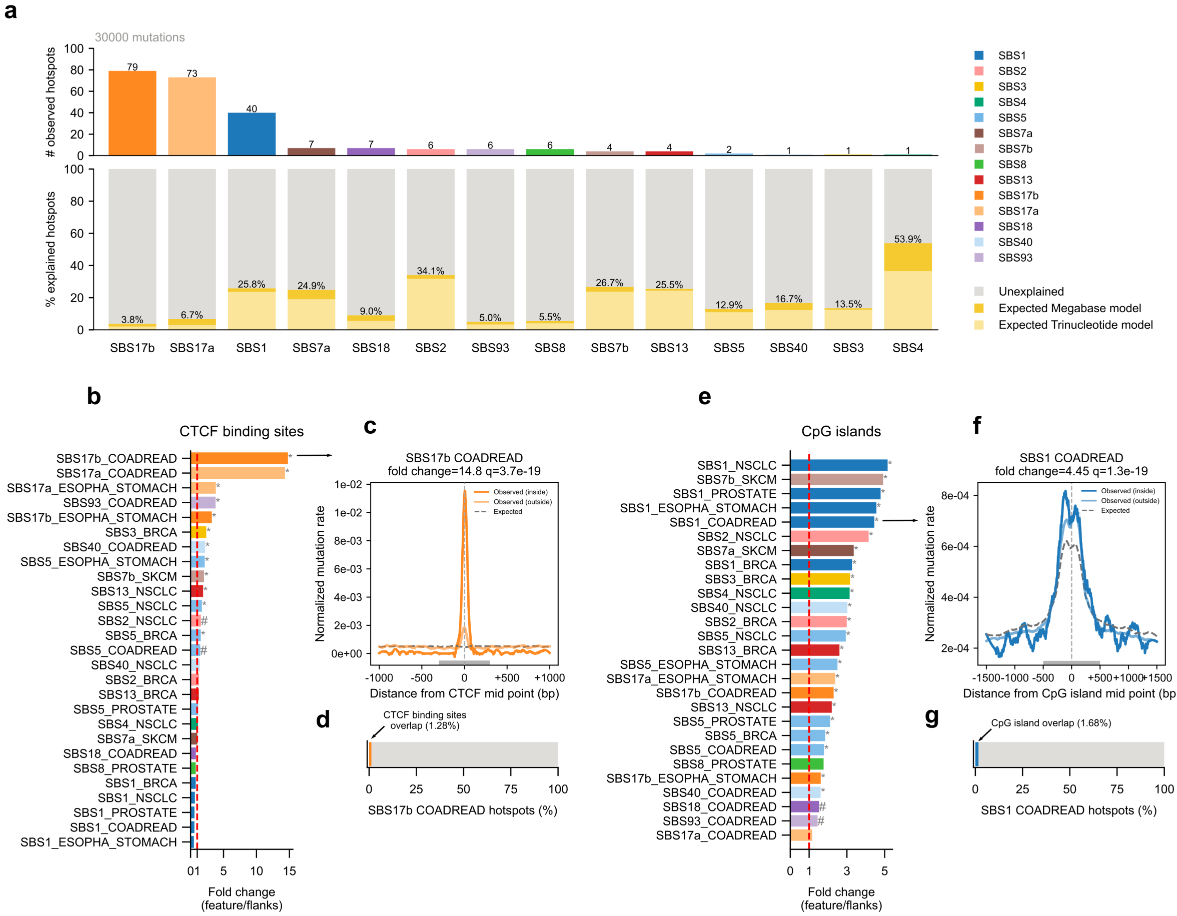


Fig. 6. Trinucleotide composition and large-scale covariates do not account for most observed hotspots. a) Observed hotspots by subsampling 30,000 total mutations (300 mutations/sample, 100 samples) within mappable megabases and the proportion of observed hotspots that are accounted for by the theoretical models (trinucleotide model in light yellow and trinucleotide-megabase model in yellow) at equivalent mutation rates and genomic regions. **b**) Fold changes of piled up observed mutations within CTCF binding sites (600 bp) compared to their flanking 5' and 3' regions (700 bp on each flank). Significance of the observed fold change compared to the expected fold change are shown as p<1e-6 (*); 1e-6<p<0.01 (#). **c**) Normalised mutation rate of SBS17b in colorectal cancers across CTCF binding sites and their flanking regions. Piled-up mutation rates of mutations inside and outside hotspots, together with the expected distribution of mutations inside hotspots, are shown. Observed fold changes and their significance with respect to the expected fold change distribution are depicted on top. Additional plots for all signatures can be found at Supp. Fig. 21. **d**) Proportion of SBS17b colorectal hotspots overlapping tissue-matched CTCF binding sites. **e**) Fold changes of piled up observed mutations within CpG islands (1,000 bp) compared to their flanking 5' and 3' regions (1,000 bp on each flank). Significance of the observed fold change compared to the expected fold change are shown as p<1e-6 (*). **f**) Normalised mutation rate of SBS1 in colorectal cancers across CpG islands and their flanking regions. Piled-up mutation rates of mutations inside and outside hotspots, together with the expected distribution of mutations inside hotspots, are shown. Observed fold changes and their significance with respect to the expected fold change distribution are depicted on top. Additional plots for all signatures can be found at Supp. Fig. 22. **g**) Proportion of SBS1 colorectal hotspots overlapping CpG islands.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

Discussion

This study constitutes, to the best of our knowledge, the largest systematic analysis of the prevalence and causes of hotspots of somatic mutations across cancer genomes. We discovered that SBS1, SBS17a and SBS17b exhibit strikingly increased propensity to form mutational hotspots. SBS1, the ubiquitous clock-like mutational process attributed to 5-methylcytosine deamination[5,6], shows increased propensity to create hotspots across most cancer types analysed. SBS17a and SBS17b, signatures of unknown aetiology, stand out as hotspot-prone both in primary esophageal and stomach cancers[34]. In our dataset, SBS17b is also observed in metastases of patients exposed to capecitabine/5-fluorouracil as part of the treatment of their primary tumours[19,35] (Supplementary Note 2). Other signatures previously related to mutational hotspots, including the dominant signature of UV-light SBS7a [9,10,26,27] and the APOBEC signature SBS2 [24,28,29], also showed increased propensity to form hotspots, although to a lesser degree than SBS1 and SBS17a/b. While this study has focused on hotspots formed by SNVs, the increasing availability of cancer genome sequences, coupled with improvements in mutation calling and the filtering of false positive calls, will pave the way for the exploration of hotspot propensity of other mutational mechanisms (e.g., indels).

These high hotspot propensities are partially explained by very particular combinations of trinucleotide specificities in their profile, genomic sequence composition, distribution of their mutations at the megabase scale and interaction with other chromatin features at smaller scales. In the case of SBS1, the unevenness of the trinucleotide mutational profile and the genome-wide depletion of NpCpG trinucleotides (resulting from the evolution of the sequence of the genome[39,40]) are key determinants of the increase in hotspots formation. SBS1 hotspots are enriched within CpG islands, which is consistent with described CpG island hypermethylation in cancers[41,42]. In the case of SBS17a and SBS17b, the uneven mutational profile and a larger variability of the megabase-scale mutation rate explain only a very small fraction of the observed hotspots. Sequence or chromatin features below the megabase scale[36,38] appear to play a preponderant role in its observed high propensity to form hotspots. Among these smaller scale features, CTCF binding sites are known to contribute to the uneven distribution of SBS17a/b mutations[13,14] and this in turn has now been shown to contribute to their high propensity to form hotspots. Whether this is related with the –yet not established– aetiology of these signatures, and what its implications may be in tumour development remains unknown.

A previous study focused on the prevalence of hotspots across tumours suggested that an important fraction could arise as a result of sequencing errors or artefacts of the alignment or mutation calling[25]. To avoid this concern, here we have focused on hotspots contributed by well-characterised mutational signatures (and by establishing careful filters of genomic regions with mappability issues and hypermutated tumours). Another potential caveat of the analysis based on trinucleotide profile-defined signatures is that some may show preferences for some nucleotides in an extended (e.g., pentanucleotide) context. The signatures included in the study,

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

however, appear to be well defined on the basis of their trinucleotide profile (see above and Supp. Fig. 19), and using this standard representation allows us to use the reference set of signatures in the study.

One of the most interesting findings of our study is that the majority of hotspots contributed by SBS1, SBS17a and SBS17b remain unexplained after accounting for the mutational profile of these signatures, the trinucleotide composition of the human genome and the megabase-scale mutation variability. A large fraction of the observed number of hotspots attributable to these signatures are apparently explained by sequence or chromatin features below the megabase scale. At the end of this study we have focused on a couple of well known features (of few tens of bases) that favour the formation of SBS1 and SBS17b. They are however capable of explaining only a tiny fraction (1-3%) of these hotspots, which suggests that other known or yet unknown features also favour SBS1 and SBS17a/b hotspot formation. These features may exist in scales of length between tens of bases and several kilobases. Probing the effect of candidate features at these scales is not a simple task, however, because they first need to be identified and mapped onto the genome. Then, their instances across the genome can be stacked and their collective influence assessed[8] (e.g., Fig. 6d,g). Carefully estimating the effect of bigger features (in the kilobase-scale) will require larger cohorts of whole-genome sequenced tumours to increase the granularity of mutation rate simulation experiments.

The discovery of the high propensity of several mutational processes (particularly striking for SBS1, SBS17a and SBS17b), to generate passenger hotspots has different implications. It indicates that our current estimates of the mutational probability below the megabase-scale are far from correct, stressing our limited understanding about how somatic mutations occur. Despite big strides in the past decades in explaining the determinants of the megabase scale mutation rate and identifying some smaller scale features influencing it, the landscape of the interactions of different mutagens and DNA repair processes with the chromatin structure and features of the genomic sequence is far from complete. Understanding the distribution of somatic mutations under neutrality is key to accurately identifying driver elements and mutations. While statistical methods typically account for heterogeneity in mutational distribution across large genomic regions (for example, see[4,43]), local scale features are generally not considered. This is particularly relevant for hotspots, which have been exploited as a signal of positive selection[20–22] and are potential targets for the development of precision cancer medicine strategies. Some of the genomic features underlying this variability of mutation rates below the megabase scale remain unknown, but even in those instances for which the role has been demonstrated[9,10,24,26–29], their incorporation into models of background mutation rate remain problematic. This problem transcends the study of mutational hotspots and impacts all efforts to model the background rate of mutations, and is more acute at non-coding regions, where the location of elements perturbing mutational processes is more challenging[44].

Spontaneous 5-methylcytosine deamination, the likely cause of SBS1 [5,6], is a universal ageing-related process affecting not only human somatic cells, but also human germ cells[45], and

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

genomes from other species[46,47]. Thus, our results showing that yet unknown DNA sequence and chromatin features influence the neutral mutational probability of SBS1 at the nucleotide resolution have implications beyond the identification of cancer drivers. Accurate estimates of mutation rates under neutrality are required for models of the evolution of genomic elements across organisms, to understand genome evolution and to study the potential effect of human germline variants. Our work, therefore, evidences the knowledge gap in our understanding on how mutation occurs at base resolution, which is needed to build these accurate models of background mutation rates.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

Methods

Cohorts

We collected somatic mutations from 78 cohorts of whole genome sequenced cancer patients included in IntOGen[32] (release 1 February 2020). Cohorts contained primary and metastatic tumours from adult and paediatric individuals, encompassing a total of 7,507 samples and 83,410,018 somatic mutations. Detailed information about each sequencing cohort and information on how to download them can be found at Supplementary Table S1 and www.intogen.org.

Pre-processing of cohorts

In order to homogenise the datasets for our analysis and minimise the number of false mutation calls, we conducted the following pre-processing on individual cohorts as follows:

- Liftover of somatic mutations to hg38 reference genome. Mutations in those cohorts that used hg19 as reference genome were lifted over to hg38 using pyliftover package version 0.3 (pypi.org/project/pyliftover/) as described in [32]. Only mutations that mapped to hg38 were kept for analysis.
- Filtering of somatic mutations: we removed mutations that a) fell outside of autosomal or sexual chromosomes; b) had the same reference and alternate nucleotides; c) had a reference nucleotide that did not match the annotated hg38 reference nucleotide; d) had an unknown nucleotide --a nucleotide not corresponding to A, C, G, T-- in their tri-nucleotide or penta-nucleotide reference sequence, as stated by their start position; e) were classified as complex indels --indels that are a mixture of insertions and deletions such as GTG>GAAA.
- Filtering of germline variants: our analysis aimed for the identification of hotspots of somatic mutations. In order to decrease contamination of somatic calls by unfiltered germline mutations, we removed mutations overlapping population variants. Briefly, we removed mutations overlapping genomic positions with one or more polymorphic variants (i.e., allele frequency equal or greater than 1%) (see “Mappable genome and high mappability megabase bins” section for complete details).
- Filtering of low mappability sequences: non-mappable regions (i.e., repetitive or non-unique sequences in the genome) are prone to sequencing artefacts. To control such errors, we discarded mutations that 1) fell outside high mappability regions and/or 2) overlapped blacklisted regions of low mappability (see “Mappable genome and high mappability megabase bins” section for complete details).
- Filtering of hypermutated samples: from each cohort, we filtered out hypermutated samples, this is, samples that carried more than 10,000 mutations and exceeded 1.5 times the interquartile range over the 75th percentile, as described in [32].

Cancer types classification

WGS samples were merged into cancer types comprising one or more individual cohorts.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

Cancer type classification was based on the Memorial Sloan Kettering Cancer Center (MSKCC) OncoTree[30] (2021-11-02 release, available at oncotree.mskcc.org). Assignment of each cohort to the different cancer type levels in the OncoTree hierarchy was carried out using the available clinical information of the cohort and can be found in Supplementary Table S1. Ad-hoc cancer types were added in those cases where the OncoTree classification did not fulfil the cohort definition. In order to avoid redundant cancer types --entities containing the same or very similar set of samples-- within our analysis, we simplified the resulting hierarchy into two different levels A (specific) and B (meta-cancer type): level A entities were the most specific annotation available for a group of samples (e.g., melanomas); when two or more level A entities could be merged together according to the hierarchy, a level B annotation was added (e.g., melanomas, basal cell carcinomas, and cutaneous squamous cell carcinoma were grouped in skin cancers). Finally, all samples were merged into the Pancancer level. Cancer types included in the analysis are listed in Supplementary Table S2.

Pre-processing of cancer types

In order to identify the presence of multiple samples originating from the same donor, we conducted a systematic analysis of shared mutations among samples in each cohort and cancer type. Briefly, for every sample in a dataset, we computed the number of equal mutations with any other sample in the group and divided it over the total number of mutations of both samples in the comparison. Samples with more than 10% of shared mutations with any other sample were flagged for manual review. Two samples from M_OS were found to have primary tumour samples sequenced in D_OS and were subsequently removed from the metastatic cohort M_OS.

Mappable genome and high mappability megabase bins

We defined the mappable genome as the fraction of the reference hg38 genome containing mutations within our analysis (2,531,296,367 bp). The mappable genome consisted of regions of high mappability that did not overlap with 1) blacklisted sequences of low mappability and/or 2) genomic positions containing population variants. Regions of high mappability (≥ 0.9) based on 100-mer pileup mappability were computed for hg38 reference genome using The GEnomic Multi-tool[48] (GEM) mappability software version 2013-04-06. BED files containing hg38 blacklisted regions of low mappability were obtained from the ENCODE Unified GRCh38 Blacklist (downloaded from encodeproject.org/files/ENCFF356LFX on 16-06-2020). Positions containing population variants were defined as those overlapping any substitution or short indel with total variant allele frequency above 1% as identified by gnomAD [49] version 3.0 (downloaded from gnomad.broadinstitute.org on 25-06-2020). Additionally, we defined a set of megabases (1 Mbp) of high mappability. We first obtained hg38 genomic bins by partitioning chromosome coordinates in consecutive non-overlapping chunks of 1 Mb length. For each bin, we computed the sequence overlap with the mappable genome using the Python library pybedtools[50,51] and kept those bins within autosomes where at least 90% of their sequence was included in the mappable genome (n=2,222 bins). Mappable bins encompassed a total of 2,065,481,419 bp.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

hg38 nucleotide sequences were retrieved from the Python package bgreference version 0.6.

Driver gene annotations

Cancer driver gene annotations were collected from two sources: the Compendium of Cancer Genes from the driver discovery pipeline IntOGen[32] (release 1 February 2020) and the COSMIC Cancer Gene Census[31] (CGC) (downloaded on 24-08-2021). The Compendium of Cancer Genes is composed of genes with experimental and/or *in silico* protein-coding driver evidence (n=568 genes). CGC list contains expert-curated genes with experimental driver evidence from sporadic and familial cancers. Only those genes annotated as somatic and having a cancer role different from fusion partners were included (n=589 genes). Our final set of driver genes consisted of 782 genes as listed in Supplementary table S3.

Identification of hotspots of somatic mutations

Genome-wide recurrently mutated positions from independent samples were identified using the new algorithm HotspotFinder version 1.0.0 (Supplementary Note 1), freely available at bitbucket.org/bbglab/hotspotfinder. Hotspots of the four mutation types (SNVs, MNVs, insertions and deletions) were analysed separately. For each cancer type, HotspotFinder was run over the set of filtered mutations after excluding those overlapping coding or non-coding sequences (5'UTR, 3'UTR, splice sites, introns, proximal and distal promoters; Supplementary Note 1) of driver elements. Hotspots were identified as single positions in the genome that contained i) 2 or more mutations of equal alternates (e.g., two C>T transitions) or ii) 2 or more mutations of different alternates (e.g., C>T and C>G). All the analyses included in the present work were carried out using hotspots of equal alternate. Hotspots were annotated with the default mappability, population variants and genomic regions provided within the method (Supplementary Note 1) and those non-overlapping genomic elements were kept. All other parameters were set as default.

Hotspot burden modelling

We modelled the relationship between hotspot burden and mutation burden per sample for each cancer type with univariate ordinary least squares (OLS) regression models using the Python package statsmodels[52].

Estimation of conversion rates

Conversion rates or the number of mutations to observe 1 hotspot were calculated for cancer types with more than 100 individuals through a subsampling experiment. For 1,000 times, we selected 100 random individuals (without replacement) and pooled their SNVs to identify hotspots of equal alternate as previously explained. We then modelled the number of hotspots per individual against their observed mutation burden using OLS regression models[52]. Conversion rates were computed as the inverse of the regression slope of significant models ($p<0.05$) for those cancer types with at least 750 significant linear models across random replicates.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

Enrichment of substitution types in hotspots

Mutations overlapping and non-overlapping hotspots (mutations inside and outside hotspots, respectively) were classified into 6 and 96 pyrimidine-based substitution types based on GRCh38 reference genome using SigProfilerMatrixGenerator[53] version 1.1.26. Hotspot enrichments for each substitution in a cancer type were computed as the ratio (fold change) of the substitution frequency in the set of mutations inside versus the substitution frequency outside. Clustering of cancer types according to enrichments of 6-class based substitutions were computed using the hierarchical clustering function cluster.hierarchy from the Python scipy library[54] with the linkage function “complete”.

Mutational signatures extraction

De novo trinucleotide based SBS mutational signatures (96-mutation types using pyrimidines as reference) were extracted using SigProfiler framework[53,55,56] for the cancer types bearing at least 30 samples and 100,000 total SNVs (Supplementary Note 2). Input GRCh38 96-mutational catalogues were calculated using SigProfilerMatrixGenerator[53] version 1.1.26 and mutational signatures were extracted with SigProfilerExtractor[56] version 1.1.0 (Supplementary Note 2). De novo signatures were decomposed into COSMIC v3.2 GRCh38 reference signatures to allow comparisons across cancer types (Supplementary Note 2). All SBS signature names used in the manuscript correspond to this reference set. Signatures that were present in at least 5% of mutations in a sample were considered active in the sample. At the cancer type level, signatures that were active in at least 5% of samples were considered active in the cancer type.

Assignment of mutational signatures to mutations and hotspots

The probability of each SNV --considering its sample of origin and trinucleotide context-- to arise from each of the decomposed COSMIC signatures in the cancer type was obtained from SigProfilerExtractor (“Decomposed_Mutation_Probabilities.txt” table for the best extracted solution). As a result, a vector of mutational probabilities was generated for each SNV. In those cases where the 1 to 1 attribution of mutations to signatures was required, mutations were credited to the signature showing highest mutational probability (maximum likelihood[36]). Hotspots were assigned to mutational signatures by computing, first, the average mutational probability vector among the mutations contributing to the hotspot, and then selecting the signature with the maximum average probability.

Estimation of hotspot propensity

We set to estimate the propensity of commonly active mutational signatures to form hotspots across cancer types independently of their number of exposed samples and mutation burden contributed to each of them. First, we selected the 7 cancer types with the largest sample size (5,000 or more observed hotspots and prioritising non-meta-cancer types when possible), including: bladder-urinary tract cancers (BLADDER_URI), breast cancers (BRCA), colorectal cancers (COADREAD), oesophagus-stomach cancers (ESOPHA_STOMACH), non-small cell lung cancers (NSCLC), prostate cancers (PROSTATE) and skin melanomas (SKCM). Then, we selected the 14 signatures that fulfilled the following criteria: i) they showed 450 or more

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

attributed hotspots (resulting in at least 1% of the total hotspots in the cancer type) and ii) contributed more than 300 high confidence mutations per sample attributed to the signature via highest probability (maximum likelihood p > 0.5) in at least 101 samples within the cancer type. The signature-cancer type pairs that passed these thresholds included: SBS1 (BLADDER_URI, BRCA, COADREAD, ESOPHA_STOMACH, NSCLC, PROSTATE), SBS13 (BLADDER_URI, BRCA, NSCLC), SBS17a (COADREAD, ESOPHA_STOMACH), SBS17b (COADREAD, ESOPHA_STOMACH), SBS18 (COADREAD), SBS2 (BLADDER_URI, BRCA, NSCLC), SBS3 (BRCA), SBS4 (NSCLC), SBS40 (COADREAD, NSCLC), SBS5 (BLADDER_URI, BRCA, COADREAD, ESOPHA_STOMACH, NSCLC, PROSTATE), SBS7a (SKCM), SBS7b (SKCM), SBS8 (PROSTATE), SBS93 (COADREAD). To estimate the propensity of a signature to form hotspots in a cancer type while accounting for differences in sample size and activity, we subsampled groups of 100 tumours of a given cancer type with a fixed number n (between 100 and 300 mutations/sample or ~ 0.048 and 0.145 mutations/sample·Mbp), of randomly selected high confidence mutations (maximum likelihood p > 0.5) without replacement contributed by the signature under analysis. For each subsample, we then counted the number of observed hotspots (allowing a single hotspot per position) among the $100 \cdot n$subsampled mutations of the signature under analysis across cancer types. This analysis was carried out over the set of high mappable megabase bins (total size 2,065,481,419 bp).

Signatures enrichment in hotspots

For each sample containing hotspots, we first computed the frequency of its active signatures in the sets of hotspot and non-hotspot mutations. These frequencies were obtained by aggregating the signature mutational probability vectors (conveying the relative contribution of all possible signatures to a mutation) across all mutations in the specified set, which were subsequently normalised to 1 (see Assignment of mutational signatures to mutations and hotspots). Then, a signature fold change (FC) in a given sample was computed as the ratio of the normalised frequency of the signature S inside hotspots versus the normalised frequency of the signature outside hotspots:

$$F(S)=\left(Prob(S)_{inside}/\Sigma_T\, Prob(T)_{inside}\right)/\left(Prob(S)_{outside}/\Sigma_T\, Prob(T)_{outside}\right)$$

To obtain the fold change per active signature we calculated the median fold change among active samples. For each signature, we tested whether the magnitude of the frequency inside hotspots deferred from that of non-hotspot mutations across sample-paired observations. We applied a two-sided Wilcoxon rank-sum test using the wilcoxon function from the Python module scipy.stats[54]. The obtained p-values were adjusted for multiple testing using the Benjamini-Hochberg method from statsmodels.sandbox.stats.multicomp function[52] in Python with $\alpha = 0.01$.

Entropy of mutational signatures profiles

The entropy of the 96-channel profiles of SBS mutational signatures (COSMIC v3.2 GRCh38)

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

was calculated after correcting by the trinucleotide content of the mappable genome, i.e. we only account for the relative mutability of each context. This is done by taking the frequency profiles from COSMIC, dividing each trinucleotide frequency by the genome-wide abundance of the corresponding reference triplet and normalising the resulting profile so that the probabilities add up to 1. The entropy was calculated using the scipy.stats.entropy function[54] in Python.

Theoretical models of hotspots formation

We devised a method to compute the expected number of hotspots generated by a given mutational process in a given DNA region, in a theoretical scenario whereby all samples in the cohort have identical mutation rates and positions mutate independently. Briefly, when the regional mutation rate is uniform, we can estimate the probability that each position in the region undergoes mutation consistently with the relative mutability of each trinucleotide context dictated by the mutational process. The expected presence of a hotspot at a given position can then be calculated as the probability of at least two samples getting the same mutation at a given position. Because of statistical independence across positions, the regional expectation is given as the sum of expectations across positions. Equipped with this method, we provide theoretical hotspot rate estimates associated with the 14 different mutational signatures from which we calculated hotspot propensity. Theoretical hotspot rates were computed in two scenarios: i) genome wide based on the trinucleotide composition alone and ii) per megabase, considering cancer type-specific variation in the distribution of mutations per signature across megabases (Supplementary Note 4). In the second model, the megabase mutation rates per signature and cancer type were calculated from the total number of mutations attributed to the signature by maximum likelihood. The same signature-cancer type pairs used to calculate observed hotspot propensity were analysed. For both models, comparisons between observed and expected hotspot propensity were carried out using 30,000 total mutations (300 muts/sample across mappable 1 Mbp bins), which was equivalent to approximately 0.14524 muts/Mbp. To match the definition of hotspots with the theoretical models, only 1 observed hotspot per genomic position was considered (e.g., 2 C>A mutations and 2 C>T within the same position resulted in 1 hotspot).

Analysis of large scale chromatin features

Chromatin accessibility and gene expression data for different tissues and cell lines matching the cancer types under analysis were obtained from the Epigenome Roadmap Project[57] at egg2.wustl.edu/roadmap/web_portal (see Supplementary Table 5 for complete details and urls). Replication timing data for 7 cell lines from solid tissues was obtained from ENCODE[58] at hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq (Supplementary Table 5). For each of these three chromatin features (mapped to hg19 reference genome), we obtained the average signal across hg19 mappable megabases --liftovered from the original hg38 mappable megabase coordinates-- as follows. For chromatin accessibility, we first computed the average counts across megabases from genome-wide fold-enrichment DNase counts tracks (BIGWIG format) per epigenome. Then, for each megabase, we computed the average DNase-seq signal across the different epigenomes linked to a cancer

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

type. Similarly, megabase gene expression signals were computed from normalised coverage genome tracks (BIGWIG format). In this case, if stranded libraries were available for an epigenome, we first added up the absolute RNA-seq signals from the negative and positive strands and then computed the average signal per epigenome and megabase. The cancer type signal per megabase was obtained by calculating the mean megabase RNA-seq signal from the different epigenomes linked to the cancer type. For replication timing data, we used the percentage-normalised Repli-seq signal tracks (BIGWIG format). Following the same approach, we obtained the average Repli-seq signal across cell lines per megabase. Signals extracted from BIGWIG files were handled using the Python package pyBigWig[59].

In order to investigate the relationship between the number of hotspots and non-hotspot mutations with chromatin accessibility, gene expression and replication timing per signature and cancer type, we first intersected mutations inside and outside hotspots with mappable megabase bins. Next, we added up the vector of mutational probabilities of each mutation to arise from a signature in the cancer type (see Assignment of mutational signatures to mutations and hotspots). For each signature, we normalised its signal across megabases for mutations inside and outside hotspots. We then categorised mappable megabases into 10 percentiles (deciles) according to the distribution of the chromatin feature signal across megabases and plotted the signature activity inside and outside hotspots on each decile.

Overdispersion of mutational signatures across the genome

For each signature, we fitted the distribution of their attributed mutation counts across mappable megabase bins by fitting a negative binomial regression model, which yields an overdispersion parameter. Briefly, the overdispersion parameter, referred to as α throughout, measures the excess variance over the mean, i.e., excess variance over the variance that we would expect if the mutation counts were Poisson distributed. Specifically, if μ denotes the mean, in our negative binomial regression setting the relationship between the variance v and the mean μ is given by the equation

$$v=\mu+\alpha\,\mu^2$$

The Python function statsmodels.discrete.discrete_model.NegativeBinomial was used[52].

Analysis of small scale chromatin features

CpG islands coordinates mapped to hg38 reference genome were downloaded from[60] at haowulab.org/software/makeCGI/model-based-cpg-islands-hg38.txt on 02-03-2022. In order to homogenise the CpG islands under analysis, only those within autosomes, 200-1,000 bp long and showing a 90% or larger overlap with the mappable genome were used for the analysis (n=30,513 elements) (Supplementary Table 5). CTCF binding sites were defined as CTCF ChIP peaks in a tissue or cell line matching the cancer type type. hg38 ChIP peak coordinates were downloaded from ReMap2022[61] at remap.univ-amu.fr on 08-03-2022 (Supplementary Table 5). Only CTCF peaks within autosomes, 200-600 bp long, and showing 90% or more overlap to the mappable genome were kept for analysis (mean n=41,307.5, range=28,802-51,635 CTCF

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

binding sites per cancer type). Intersections between annotations were carried out using pybedtools[50,51]. The enrichment of hotspots within CpG islands and CTCF binding sites was computed as follows. For each individual small scale chromatin feature, we first constructed a window of length L where the feature of length $L_{feature}$ was positioned in the centre surrounded by two flanking sites of equal size $\left(L-L_{feature}\right)/2$. Window length was defined as 3,000 bp for CpG islands and 2,000 bp for CTCF binding sites. Feature length was set as the maximum size encompassing all individual features under analysis as: CpG islands=1,000 bp, CTCF=600 bp. For each active signature in a cancer type and the corresponding matched small scale chromatin feature, we intersected inside and outside hotspot mutations attributed to the signature by maximum likelihood with each window. We obtained the expected distribution of mutations per signature by randomising 1,000 times the observed number of mutations inside hotspots across the window according to the signature trinucleotide probabilities and the window sequence composition. To compute the hotspot enrichment in the feature (fold change), we piled-up mutations in equal positions across windows and calculated the ratio of mutations inside the feature versus its flanks. The significance of the observed fold changes was estimated by fitting simulated fold changes to a gaussian kernel density estimate distribution and deriving the upper quantile of the observed fold change. Resulting p-values were adjusted for multiple testing using the statsmodels Benjamini-Hochberg function[52] with α=0.01. To visualise the results across piled-up features, observed and expected mutation counts per position were normalised to the respective total number of mutations in the set across the window. A smoothing Savitzky-Golay filter of length 301 bp for CpG islands and 101 bp for CTCF binding sites was applied using the scipy.signal.savgol_filter function[54].

Sequence logos

We computed the frequency of the nucleotide sequence composition around hotspots attributed to a signature across cancer types. The same signature-cancer type groups as those used to calculate hotspot propensity were used. For each hotspot in the signature, we retrieved the 10 bp 5' and 3' flanking sequences (considering the strand containing a pyrimidine in the hotspot position) from bgreference package and built a 21 bp window centred at the hotspot. We then computed the information content over the nucleotide frequency with respect to the nucleotide hg38 mappable genome frequency at each position across the window. Logo plots and information content were generated using the Python package logomaker[62] version 0.8.

Additional software used

The following Python packages were used across different analyses, including: matplotlib[63], numpy[64], and pandas[65].

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

Supplementary material

Supplementary Table S1. Sequencing cohorts used in the analysis.
Supplementary Table S2. Cancer types used in the analysis.
Supplementary Table S3. List of cancer driver genes.
Supplementary Table S4. Observed hotspots across cancer types.
Supplementary Table S5. Annotations of large and small scale chromatin covariates.
Supplementary Note 1. HotspotFinder.
Supplementary Note 2. Mutational signatures extraction.
Supplementary Note 3. Comment on hotspot propensity.
Supplementary Note 4. Theoretical models of hotspots formation.

Data availability

Somatic mutations were retrieved from several sources as listed in Supplementary Table S1. Large and small scale features (DNase, Repli-seq, RNA-seq, CTCF ChIP, CpG islands) are listed in Supplementary Table S5.

Code availability

The code used within this work is freely available from external sources (see Methods) or has been developed in-house. HotspotFinder algorithm is available for download at bitbucket.org/bbglab/hotspotfinder. The in-house code containing the analyses and figures will be available at the time of publication.

Acknowledgements

The authors wish to acknowledge the contribution of patients, families and biomedical researchers who shared, processed and sequenced the data used within the study. The results published here are in part based on the data generated by the Pan-Cancer Analysis of Whole Genomes, St Jude Children's Research Hospital, PedcBioPortal, Therapeutically Applicable Research to Generate Effective Treatments program. This publication and the underlying study have been made possible partly on the basis of the data that Hartwig Medical Foundation has made available to the study. We acknowledge the technical contributions of Iker Reyes-Salazar and Loris Mularoni to HotspotFinder algorithm and annotations. N.L.-B. acknowledges funding from the European Research Council (consolidator grant 682398) and ERDF/Spanish Ministry of Science, Innovation and Universities – Spanish State Research Agency/DamReMap Project (RTI2018-094095-B-I00) and Asociación Española Contra el Cáncer (AECC) (GC16173697BIGA). IRB Barcelona is a recipient of a Severo Ochoa Centre of Excellence Award from the Spanish Ministry of Economy and Competitiveness

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

(MINECO; Government of Spain) and is supported by CERCA (Generalitat de Catalunya). C.A.-P. is supported by “la Caixa” Foundation (ID 100010434) fellowship (LCF/BQ/ES18/11670011).

Author contributions

C.A.-P, A.G.-P. and N.L.-B. conceptualised the project. C.A.-P curated the sequencing data for the analysis and implemented HotspotFinder. F.M. implemented the theoretical models of hotspots formation and generated the expected hotspot counts. C.A.-P carried out all the remaining analyses. C.A.-P prepared the figures. All authors participated in the design of the analyses and figures and in the interpretation of the results. C.A.-P, A.G.-P. and F.M. drafted the manuscript. All authors reviewed and edited the final manuscript. A.G.-P. and N.L.-B. supervised the project.

References

1. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015).
2. Stamatoyannopoulos, J. A. *et al.* Human mutation rate associated with DNA replication timing. *Nat. Genet.* **41**, 393–395 (2009).
3. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).
4. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
5. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
6. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
7. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
8. Gonzalez-Perez, A., Sabarinathan, R. & Lopez-Bigas, N. Local determinants of the mutational

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

landscape of the human genome. *Cell* **177**, 101–114 (2019).

9. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**, 264–267 (2016).
10. Perera, D. *et al.* Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature* **532**, 259–263 (2016).
11. Kaiser, V. B., Taylor, M. S. & Semple, C. A. Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types. *PLoS Genet.* **12**, e1006207 (2016).
12. Frigola, J., Sabarinathan, R., Gonzalez-Perez, A. & Lopez-Bigas, N. Variable interplay of UV-induced DNA damage and repair at transcription factor binding sites. *Nucleic Acids Res.* **49**, 891–901 (2021).
13. Katainen, R. *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.* **47**, 818–821 (2015).
14. Guo, Y. A. *et al.* Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. *Nat. Commun.* **9**, 1520 (2018).
15. Zou, X. *et al.* Short inverted repeats contribute to localized mutability in human somatic cells. *Nucleic Acids Res.* **45**, 11213–11221 (2017).
16. Georgakopoulos-Soares, I., Morganella, S., Jain, N., Hemberg, M. & Nik-Zainal, S. Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *Genome Res.* **28**, 1264–1271 (2018).
17. Nowarski, R. & Kotler, M. APOBEC3 cytidine deaminases in double-strand DNA break repair and cancer promotion. *Cancer Res.* **73**, 3494–3498 (2013).
18. Mas-Ponte, D. & Supek, F. DNA mismatch repair promotes APOBEC3-mediated diffuse hypermutation in human cancers. *Nat. Genet.* **52**, 958–968 (2020).
19. Pich, O. *et al.* The mutational footprints of cancer therapies. *Nat. Genet.* **51**, 1732–1740 (2019).
20. Miller, M. L. *et al.* Pan-Cancer Analysis of Mutation Hotspots in Protein Domains. *Cell Syst.* **1**, 197–209 (2015).
21. Chang, M. T. *et al.* Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.* **34**, 155–163 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

22. Juul, R. I., Nielsen, M. M., Juul, M., Feuerbach, L. & Pedersen, J. S. The landscape and driver potential of site-specific hotspots across cancer genomes. *NPJ Genom. Med.* **6**, 33 (2021).
23. Stobbe, M. D. *et al.* Recurrent somatic mutations reveal new insights into consequences of mutagenic processes in cancer. *PLoS Comput. Biol.* **15**, e1007496 (2019).
24. Hess, J. M. *et al.* Passenger hotspot mutations in cancer. *Cancer Cell* **36**, 288-301.e14 (2019).
25. Smith, T. C. A., Carr, A. M. & Eyre-Walker, A. C. Are sites with multiple single nucleotide variants in cancer genomes a consequence of drivers, hypermutable sites or sequencing errors? *PeerJ* **4**, e2391 (2016).
26. Fredriksson, N. J. *et al.* Recurrent promoter mutations in melanoma are defined by an extended context-specific mutational signature. *PLoS Genet.* **13**, e1006773 (2017).
27. Elliott, K. *et al.* Elevated pyrimidine dimer formation at distinct genomic bases underlies promoter mutation hotspots in UV-exposed cancers. *PLoS Genet.* **14**, e1007849 (2018).
28. Buisson, R. *et al.* Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science* **364**, (2019).
29. Shi, M.-J. *et al.* Identification of new driver and passenger mutations within APOBEC-induced hotspot mutations in bladder cancer. *Genome Med.* **12**, 85 (2020).
30. Kundra, R. *et al.* Oncotree: A cancer classification system for precision oncology. *JCO Clin. Cancer Inform.* **5**, 221–230 (2021).
31. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
32. Martínez-Jiménez, F. *et al.* A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**, 555–572 (2020).
33. Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
34. Secrier, M. *et al.* Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nat. Genet.* **48**, 1131–1141 (2016).
35. Christensen, S. *et al.* 5-Fluorouracil treatment induces characteristic T>G mutations in human

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

cancer. *Nat. Commun.* **10**, 4571 (2019).

36. Morganella, S. *et al.* The topography of mutational processes in breast cancer genomes. *Nat. Commun.* **7**, 11383 (2016).
37. Tomkova, M., Tomek, J., Kriaucionis, S. & Schuster-Böckler, B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. *Genome Biol.* **19**, 129 (2018).
38. Otlu-Saritas, B. *et al.* Topography of mutational signatures in human cancer. *BioRxiv* (2022) doi:10.1101/2022.05.29.493921.
39. Pich, O. *et al.* Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes. *Cell* **175**, 1074-1087.e18 (2018).
40. Long, H. *et al.* Evolutionary determinants of genome-wide nucleotide composition. *Nat. Ecol. Evol.* **2**, 237–240 (2018).
41. Costello, J. F. *et al.* Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. *Nat. Genet.* **24**, 132–138 (2000).
42. Weisenberger, D. J. *et al.* CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nat. Genet.* **38**, 787–793 (2006).
43. Martincorena, I. *et al.* Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029-1041.e21 (2017).
44. Sherman, M. A. *et al.* Genome-wide mapping of somatic mutation rates uncovers drivers of cancer. *Nat. Biotechnol.* (2022) doi:10.1038/s41587-022-01353-8.
45. Moore, L. *et al.* The mutational landscape of human somatic and germline cells. *Nature* **597**, 381–386 (2021).
46. Cagan, A. *et al.* Somatic mutation rates scale with lifespan across mammals. *Nature* **604**, 517–524 (2022).
47. Fryxell, K. J. & Zuckerkandl, E. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol. Biol. Evol.* **17**, 1371–1383 (2000).
48. Marco-Sola, S., Sammeth, M., Guigó, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

alignment by filtration. *Nat. Methods* **9**, 1185–1188 (2012).

49. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
50. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
51. Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423–3424 (2011).
52. Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. in *Proceedings of the 9th Python in Science Conference* 92–96 (SciPy, 2010). doi:10.25080/Majora-92bf1922-011.
53. Bergstrom, E. N. *et al.* SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics* **20**, 685 (2019).
54. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
55. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246–259 (2013).
56. Islam, S. M. A. *et al.* Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *BioRxiv* (2020) doi:10.1101/2020.12.13.422570.
57. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
58. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
59. Ryan, D., Grüning, B. & Ramirez, F. Pybigwig 0.2.4. *Zenodo* (2016) doi:10.5281/zenodo.45238.
60. Wu, H., Caffo, B., Jaffee, H. A., Irizarry, R. A. & Feinberg, A. P. Redefining CpG islands using hidden Markov models. *Biostatistics* **11**, 499–514 (2010).
61. Hammal, F., de Langen, P., Bergon, A., Lopez, F. & Ballester, B. ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.14.507952; this version posted September 17, 2022. 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-NC-ND 4.0 International license.

binding sequencing experiments. *Nucleic Acids Res.* **50**, D316–D325 (2022).

62. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274 (2020).
63. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
64. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
65. Reback, J. *et al.* pandas-dev/pandas: Pandas 1.4.4. *Zenodo* (2022) doi:10.5281/zenodo.7037953.