Functional and genetic determinants of mutation rate variability in regulatory elements of cancer genomes

Christian A. Lee[1,2,*], Diala Abd-Rabbo[1,*], Jüri Reimand[1,2,3,@]

1. Computational Biology Program, Ontario Institute for Cancer Research, Toronto, ON, Canada
2. Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada
3. Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada
* These authors contributed equally
[@] correspondence: Juri.Reimand@utoronto.ca

ABSTRACT

Background: Cancer genomes are shaped by mutational processes with complex spatial variation at multiple scales. Entire classes of regulatory elements are affected by local variations in mutation frequency. However, the underlying mutational mechanisms with functional and genetic determinants remain poorly understood.
Results: We characterised the mutational landscape of 1.3 million gene regulatory and chromatin architectural elements in 2,419 whole cancer genomes with transcriptional and pathway activity, functional conservation and recurrent driver events. We developed RM2, a statistical model that quantifies mutational enrichment or depletion in classes of genomic elements through genetic, trinucleotide and megabase-scale effects. We report a map of localised mutational processes affecting CTCF binding sites, transcription start sites (TSS) and tissue-specific open-chromatin regions. We show that increased mutational frequency in TSSs correlates with mRNA abundance in most cancer types, while open-chromatin regions are generally enriched in mutations. We identified ~10,000 CTCF binding sites with core DNA motifs and constitutive binding in 66 cell types that represent focal points of local mutagenesis. We detected site-specific mutational signatures, such as SBS40 in open-chromatin regions in prostate cancer and SBS17b in CTCF binding sites in gastrointestinal cancers. We also proposed candidate drivers of localised mutagenesis: *BRAF* mutations associate with mutational enrichments at CTCF binding sites in melanoma, and *ARID1A* mutations with TSS-specific mutations in pancreatic cancer.
Conclusions: Our method and catalogue of localised mutational processes provide novel perspectives to cancer genome evolution, mutagenesis, DNA repair and driver discovery. Functional and genetic correlates of localised mutagenesis provide mechanistic hypotheses for future studies.

Introduction

Genomes accumulate somatic mutations through exposure to exogenous and endogenous mutagens. Subsets of these mutations confer cells select proliferative advantages and drive oncogenesis, while most mutations are functionally neutral passengers [1,2]. The discovery and validation of driver mutations is a major focus of cancer genomics research [3-5]. However, the genome-wide landscape of passenger mutations is also instrumental to our understanding of oncogenesis and tumor evolution [6,7]. Somatic mutation frequencies show complex genomic variation at multiple resolutions [8]. In megabase-scale genomic windows, variations in mutation frequencies are associated with transcriptional activity, chromatin state and DNA replication, as late-replicating and non-transcribed regions are often more mutated than regions of early replication and highly expressed genes [9-12]. At the base pair resolution, certain trinucleotides are preferentially mutated through processes of carcinogen exposures, defective DNA repair pathways, and aberrant DNA replication [13-15]. For example, mutational signatures detected in metastatic tumors are informative of the treatment history of patients [16,17]. In concert, these large-scale and nucleotide-level variations contribute to tumor heterogeneity and leave a footprint of tumor evolution and its cell of origin [12,18,19].

Complex variation in mutation frequencies is also apparent across intermediate genomic resolutions spanning hundreds to thousands of nucleotides. This encapsulates diverse functional genomic elements such as exons, transcription factor binding sites (TFBS) and chromatin architectural elements [8,20]. DNA bound by nucleosomes and transcription factors (TFs) show increased mutation frequencies in cancer genomes [21-23]. Active promoters in melanoma are enriched in UV-induced C>T somatic mutations resulting from differential activity of nucleotide excision repair influenced by DNA-binding of regulatory proteins [24,25]. Likewise, DNA-binding sites of the master transcriptional regulator and chromatin architectural protein CTCF (CCCTC-binding factor) are enriched in somatic mutations in multiple cancer types [21,26-28]. In contrast, certain genomic elements, such as chromatin-accessible regulatory regions [29] and protein-coding exons [30], have been shown to harbor relatively fewer mutations due to increased DNA repair activity. While the majority of such non-coding mutations tend to represent functionally neutral passengers, some regulatory elements at the high end of the mutation frequency spectrum may undergo positive selection due to their effects on cancer phenotypes. For example, the mutation

64 hotspot in the *TERT* promoter creates a TFBS of the ETS TF family that leads to constitutive
65 activation of *TERT* and enables replicative immortality of cancer cells [31-33]. Recent studies have
66 catalogued candidate non-coding driver elements in gene regulatory and chromatin architectural
67 regions of the cancer genome with functional validations of novel elements [34-36] and shown that
68 non-coding mutations converge on molecular pathways and regulatory networks involved in
69 oncogenesis [37,38]. Thus, we need to characterise localised mutational processes to deconvolute
70 the role of carcinogens and endogenous mutational processes and the effects of positive selection
71 in the non-coding genome. However, few dedicated computational methods exist to analyse
72 variation in mutation frequencies at this resolution. As a result, there is a lack of large-scale
73 analyses of the local mutation landscape in pan-cancer WGS datasets, leaving patterns of
74 mutational enrichment and depletion undetected, and the genetic and environmental determinants
75 poorly understood.

76 Here we developed a new statistical framework that quantifies the activity of mutational
77 processes and signatures on specific classes of non-coding elements of the cancer genome. Our
78 model considers local sequence context, megabase-level somatic mutation rates and genetic
79 covariates to control for variation at the trinucleotide and megabase resolution while isolating
80 site-level effects. We performed a systematic analysis of local mutation frequency variation in
81 three classes of gene-regulatory and chromatin architectural genomic elements across 2,419
82 whole cancer genomes of the ICGC/TCGA Pan-cancer Analysis of Whole Genomes (PCAWG)
83 project [3]. We found a pervasive mutational enrichment at these functional non-coding elements
84 that was characterised by specific mutational signatures and transcriptional and pathway-level
85 activities in select cancer types. We detected statistical interactions of local mutagenesis and
86 recurrent genomic alterations that suggest potential genetic mechanisms driving the underlying
87 mutational processes. Our computational framework and systematic analysis reveal the diversity
88 of mutational processes in functional non-coding elements of the cancer genome and their roles
89 in somatic genome evolution, cancer phenotypes and molecular heterogeneity.

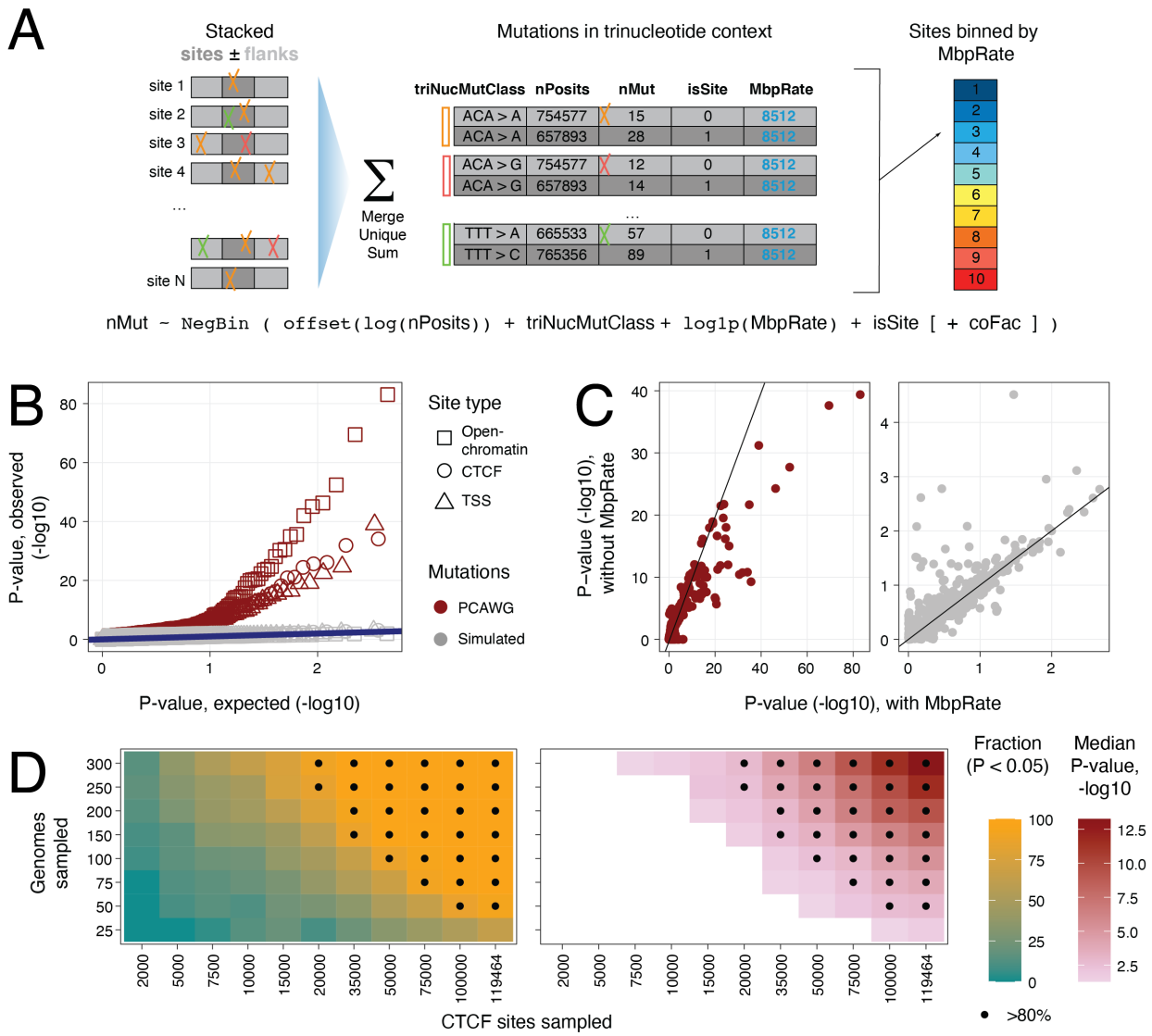

Figure 1. Characterising local mutational processes with RM2. A. Method overview. RM2 studies a set of genomic elements (*i.e.*, sites) and somatic mutations in cancer genomes using a negative binomial regression model. Sites of constant genomic width (dark gray) and two control flanking sequences (light gray) are used (*isSite*). Sites and flanks are collapsed into unique nucleotides and grouped to ten bins using their megabase-scale mutation frequency (*MbpRate*). Mutations in sites and flanks (*nMut*) are grouped by trinucleotide type (*triNucMutClass*). Trinucleotide content corresponding to the potential genomic space for mutations is used as model offset (*nPosits*). Log-likelihood tests are used to compare the mutation frequencies in sites and flanking regions by removing the model cofactor *isSite*. The optional cofactor *coFac* enables interaction analysis of genetic and clinical variables. **B.** QQ-plot shows the observed and expected P-values of true and simulated mutations from PCAWG. No significant signals were identified in simulated data (FDR < 0.05), indicating that our method is well-calibrated. **C.** Comparison of model performance with and without *MbpRate* covariate. Analysis of true (left) and simulated mutations (right) shows the advantage of modelling megabase-scale mutation frequency. **D.** Power analysis of RM2 using down-sampling of CTCF binding sites and liver cancer genomes. Fraction of significant results (left) and median P-value (right) are shown. Panels B-C include total mutations, strand- and signature-specific mutations as in Figure 2. Only total mutations were included for analyses in C-D.

Results

A statistical framework for quantifying localised mutagenesis in cancer genomes

We implemented a statistical model, Regression Models for Local Mutations (RM2), to quantify the local activity of mutational processes in functional elements (*i.e.*, sites) of whole cancer genomes, each spanning tens to hundreds of nucleotides (**Figure 1A**). The model considers a genome-wide set of elements, such as TFBSs isolated from chromatin immunoprecipitation with DNA sequencing (ChIP-seq), detected in thousands to hundreds of thousands of loci. The model uses negative binomial regression to evaluate whether the genomic elements of interest are collectively subject to a different mutation frequency compared to control sequences upstream and downstream of these elements. Somatic single nucleotide variants (SNVs) and small insertions-deletions (indels) were analysed; however, the model can be extended to rare germline variation and other classes of variants such as structural variant breakpoints. The model considers four types of information to evaluate local mutation rates: 1) nucleotide sequence content of genomic elements and control sequences representing the potential space for mutagenesis, grouped by 96 trinucleotide contexts and one indel context (*nPosits*), 2) the counts of observed somatic mutations in the cohort of tumors (*nMut*) in genomic elements and control sequences also grouped by 96 trinucleotide signatures and one indel signature required to derive mutation frequencies (*triNucMut*), 3) megabase-scale background mutation rates of elements computed across the cohort of tumors (*MbpRate*) to account for large-scale mutation correlates, such as transcription and chromatin state, and 4) an optional binary cofactor (*coFac*) to stratify tumors based on their genetic makeup (*e.g.*, presence of a driver mutation) or clinical information (*e.g.*, tumor subtype or stage). Genomic elements and flanking control regions are pooled into equally sized bins based on their megabase-scale mutation frequencies (ten bins by default). Elements and flanking control sequences are distinguished using the binary cofactor *isSite*. The full model is written as follows:

$$nMut \sim \text{NegBin}(\ \text{offset}(\ \log(\ nPosits)\ ) + triNucMut + \text{log1p}(\ MbpRate\ ) + isSite\ [\ + coFac\ ]\ ).$$

120 To determine whether the mutation frequencies of genomic elements differ from flanking
121 sequences given trinucleotide-level and megabase-scale covariates, we evaluate the significance
122 of the cofactor *isSite* using a likelihood-ratio test. Significant and positive coefficients of this
123 cofactor indicate increased mutation rates in genomic elements relative to flanking controls,
124 while negative coefficients indicate a depletion of mutations. Similarly, we can discover
125 potential genetic or clinical interactions with localized activity of mutational processes. Given a
126 binary subgroup classification of tumors (*coFac*), we evaluate the significance of its interaction
127 with local mutation rates (*isSite:coFac*). Positive coefficients of the interaction indicate that the
128 mutation rates in a clinical or genetic tumor subgroup are elevated when accounting for the
129 overall differences of the subgroups. We also extend the analysis to classes of mutations, such as
130 those of COSMIC mutational signatures, by allowing only specific classes to be included in the
131 mutation counts (*nMut*). We evaluated the performance of our method using simulated datasets,
132 power analysis and parameter variations as described below (**Figure 1B-D**).

133

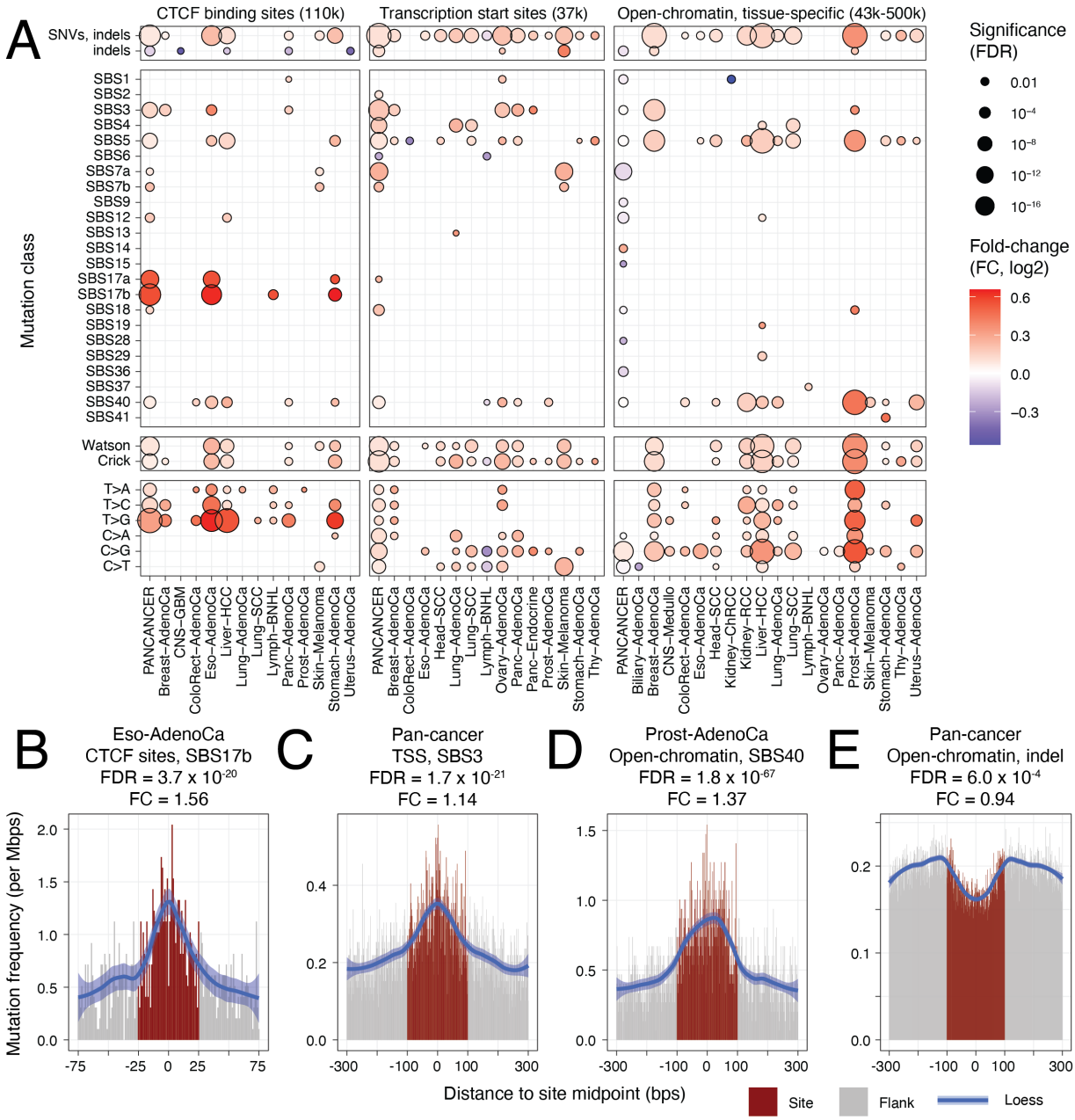

Figure 2. Comprehensive map of mutational processes in gene-regulatory and chromatin architectural elements of cancer genomes. A. Comparison of mutation frequencies in DNA-binding sites of the CTCF chromatin architectural factor (left), transcription start sites (TSS) and cancer-specific open-chromatin sites in 2,419 whole cancer genomes (*FDR* < 0.05). Total mutations (SNVs, indels) and mutations grouped by single base substitution (SBS) signatures, substitution types and DNA strand were analysed. Open-chromatin sites were filtered to exclude TSSs and CTCF sites. **B-E.** Examples of localised mutation frequencies and signatures: **B.** Enrichment SBS17b in CTCF binding sites in esophageal adenocarcinoma, **C.** Pan-cancer enrichment of SBS3 mutations in TSSs. **D.** Enrichment of SBS40 mutations in open-chromatin sites in prostate adenocarcinoma. **E.** Pan-cancer depletion of indel mutations in open-chromatin sites.

Comprehensive map of mutational processes in gene-regulatory and chromatin architectural elements of cancer genomes

137 To study localised mutation frequencies in gene-regulatory and chromatin architectural elements,
138 we used the pan-cancer dataset of 2,514 whole cancer genomes of the PCAWG project [3] with
139 high-confidence SNVs and indels. We individually analysed 25/35 cancer types with at least 25
140 samples, as well as the pan-cancer set of all 35 cancer types (**Supplementary Figure 1A**).
141 Hypermutated tumors (69 or 2.7%) were excluded to avoid confounding effects. To perform a
142 conservative analysis, we excluded a few tumors as outliers (33 or 1.4%) where even single-
143 sample RM2 analysis revealed highly significant mutational enrichments (*FDR* < 0.001)
144 (**Supplementary Figure 1B**). This ensured that our findings were shared across most tumor
145 genomes and did not represent isolated trends representative of few tumors. The final analysis
146 considered 2,419 cancer genomes of 35 cancer types with 22.7 million mutations including 1.62
147 million indels. In addition to all mutations combined, we grouped the mutations by COSMIC
148 mutational signatures of single base substitutions (SBS) inferred in the PCAWG study [14],
149 reference and alternative nucleotide pairs, and DNA strand information. Indel mutations were
150 pooled with SNVs and also analysed separately.

151 Three classes of genomic elements spanning 1,269,347 unique loci and 10% (337.0 Mbps) of the
152 human genome were analysed. These included 119,464 CTCF binding sites conserved in at least
153 two cell lines in the ENCODE project [39], 37,309 transcription start sites (TSS) of protein-coding
154 genes from the Ensembl database (GRCh37), and 1,193,391 unique open-chromatin sites. Open-
155 chromatin sites were analysed in tissue-specific subsets, ranging from 43,000 to 500,000 sites per
156 cancer type (**Supplementary Table 1A**). For most cancer types (17/25), matching open-
157 chromatin sites were adapted from the ATAC-seq profiles of primary tumors of the Cancer
158 Genome Atlas (TCGA) project [40]. The pan-cancer analysis used 500,183 pan-cancer sites from
159 TCGA. For eight cancer types, DNAse-seq profiles of the closest relevant normal tissues of the
160 Roadmap Epigenomics project [41] were used. Open-chromatin sites were filtered to exclude TSSs
161 and CTCF binding sites to allow direct comparison and reduce confounding effects of the three
162 site classes.

163 The analysis revealed a landscape of localised mutational processes in gene-regulatory and
164 chromatin architectural elements of cancer genomes (**Figure 2A**). We found 307 significant
165 differences in mutation frequencies in the three classes of genomic elements (RM2, FDR < 0.05).
166 These affected 21 cancer types and included 23 unique mutational signatures. The majority of

167 findings (281) indicated mutational enrichments at the sites compared to adjacent flanking
168 regions, while a few reduced mutation frequencies were also observed (26 or 8%). The strongest
169 cumulative signals were found in open-chromatin sites in prostate, liver and breast cancers and
170 CTCF binding sites in liver and esophageal cancers.

171 We first focused on the mutational profiles of CTCF DNA-binding sites. Mutational enrichments
172 in CTCF binding sites were the strongest in liver hepatocellular carcinoma (RM2 *FDR* = 1.22 x
173 10^{-12}, fold-change (FC) = 1.08), esophageal adenocarcinoma (*FDR* = 6.0 x 10^{-20}, FC = 1.17) and
174 stomach adenocarcinoma (*FDR* = 6.2 x 10^{-11}, FC = 1.16), and the pan-cancer cohort. Smaller
175 enrichments were detected in melanoma, pancreatic and breast cancer (*FDR* ≤ 0.02). Subgroups
176 of mutations revealed further signals. Strong enrichments of thymine base substitutions (T>G,
177 T>C, T>A) were found in nine cancer types (*e.g.*, T>G in Liver-HCC, *FDR* = 1.3 x 10^{-30}, FC =
178 1.46), while cytosine base substitutions were not enriched. We then asked whether CTCF
179 binding sites were characterised by COSMIC SBS signatures. The strongest site-specific
180 enrichments were found for SBS17 in esophageal and stomach cancer and Non-Hodgkin's
181 lymphoma (SBS17b in esophageal cancer : *FDR* = 3.7 x 10^{-20}, FC = 1.56) (**Figure 2B**).
182 Interestingly, liver and other cancer types with overall frequent CTCF binding site mutations did
183 not show a site-specific enrichment of SBS17. The aetiology of SBS17b is unknown, however it
184 has been linked to acid reflux and oxidative damage to DNA in gastro-esophageal cancers [42,43],
185 and a similar mutational signature found in metastatic tumors has been associated with the
186 effects of nucleoside metabolic inhibitor chemotherapies capecitabine and 5-FU [16]. Our analysis
187 suggests that effects of these mutagens may be especially active at insulator and chromatin
188 architectural elements bound by CTCF in tissues of the digestive system. These results confirm
189 earlier reports of elevated mutation rates in CTCF DNA-binding sites [21,27,28] in a large and
190 diverse dataset of whole cancer genomes, validating our computational model and refining the
191 annotation of the mutational processes associated with CTCF binding sites.

192 TSS of protein-coding genes were significantly enriched in mutations in the pan-cancer cohort
193 (*FDR* = 1.7 x 10^{-37}, FC = 1.07) and in cohorts of 13/25 cancer types, most prominently in
194 melanoma (*FDR* = 3.6 x 10^{-14}, FC = 1.15), breast, head, lung, ovary and pancreatic cancers (*FDR*
195 ≤ 10^{-5}). Enrichments of cytosine base substitutions were found in melanoma (C>T, *FDR* = 1.4 x
196 10^{-15}, FC = 1.19) as well as lung, ovarian and head and neck cancers. In contrast to CTCF

197 binding sites, thymine base substitution frequencies were not elevated. Mutational signature
198 analysis highlighted an elevated activity of the aging-associated signature SBS5 in the pan-
199 cancer cohort (*FDR* = 2.4 x 10^{-13}, FC = 1.07) and in eight cancer types. Signature SBS3,
200 associated with defects of homologous recombination-based DNA damage repair, was enriched
201 at TSSs in the pan-cancer cohort (*FDR* = 1.7 x 10^{-21}, FC = 1.14) (**Figure 2C**) as well as breast,
202 pancreatic and ovarian cancers (*FDR* ≤ 10^{-5}). Spontaneous formation of endogenous DNA
203 double strand breaks at promoters has been associated with the pause and release of RNA
204 polymerase II and linked to chromosomal translocations in cancer [44], suggesting a mechanism of
205 this TSS-specific mutagenesis in cancer genomes. TSSs were enriched in carcinogen-driven
206 mutational signatures, such as the ultraviolet light signature in melanoma (SBS7a: *FDR* = 1.1 x
207 10^{-14}, FC = 1.21) and the tobacco signature in two lung cancer cohorts (SBS4: *FDR* ≤ 10^{-5}).
208 These match the major mutagens and exposures of those cancer types, indicating an overall
209 increased vulnerability of TSSs to mutational processes and carcinogens. This analysis extends
210 previous reports of increased mutation frequencies in promoters in melanoma [24,25] to additional
211 cancer types, demonstrating that TSS-specific mutational processes are widely active in cancer
212 genomes.

213 Tissue-specific open-chromatin regions were also enriched in mutations in 12/25 cancer types,
214 with the strongest signals in prostate cancer (FDR = 9.8 x 10^{-81}, FC = 1.29) as well as liver, lung,
215 breast and kidney cancer (FDR ≤ 10^{-14}) whereas C>G mutations were more frequently observed
216 compared to other base substitutions. Mutational signature analysis revealed the enrichment of
217 SBS40 mutations in open-chromatin regions in nine cancer types and the pan-cancer cohort, with
218 the strongest signal apparent in prostate cancer (FDR = 1.8 x 10^{-67}, FC = 1.33) (**Figure 2D**). The
219 aging-associated signature SBS5 was enriched in open-chromatin sites in eleven cancer types,
220 especially in liver, prostate, breast and lung cancer (FDR ≤ 10^{-9}). In contrast to strong signals in
221 individual cancer types, the analysis of pan-cancer open-chromatin sites and mutations revealed
222 only a trend of elevated mutation frequencies at open-chromatin sites (FDR = 0.056, FC = 1.01).
223 The pan-cancer sites were depleted of indel mutations (*FDR* = 6.0 x 10^{-4}, FC = 0.94) (**Figure**
224 **2E**) and several less-frequent mutational signatures, while those effects were not observed in
225 individual cancer types. The limited findings of the pan-cancer analysis emphasize the
226 importance of integrating open-chromatin profiles of matched cancer types to study the

227 interactions of chromatin and mutagenesis. Overall, open-chromatin regions of individual cancer
228 types are enriched in mutations through specific mutational processes acting at these local scales.

229 We used the generated map to benchmark our model and compared it to a PCAWG dataset of
230 simulated variant calls designed to approximate neutral genome evolution [4]. First, analysis of
231 simulated data did not reveal any significant differences in mutation frequencies in the three
232 classes of elements (all $FDR > 0.05$). Quantile-quantile analysis of P-values confirmed that the
233 model is well calibrated for true and simulated mutations (**Figure 1B**). Second, we compared the
234 models with and without accounting for the megabase-scale mutation frequency covariate
235 (MbpRate) and found that including the information led to higher significance in the true
236 mutation dataset and lower significance in the simulated mutation dataset (**Figure 1C**). Third, we
237 evaluated the statistical power of our model by analyzing total mutations in down-sampled
238 subsets of liver cancer genomes and CTCF binding sites (**Figure 1D**). For example, the
239 mutational enrichment in CTCF sites was detectable 80% of the time when sampling 75 genomes
240 and 75,000 CTCF binding sites. Fourth, we varied the parameter corresponding to the normalised
241 width of genomic elements for the three classes (**Supplementary Figure 2A**). Site-specific
242 differences in mutation frequencies were robustly detected for multiple genomic widths of sites.
243 However, mutational enrichments in site bound by CTCF were generally narrower (50 bps)
244 compared to TSS and open-chromatin sites (200 bps), indicating differences in mutational
245 processes. In summary, our method provides a versatile and well-calibrated framework for
246 analysing localised mutational processes in cancer genomes.

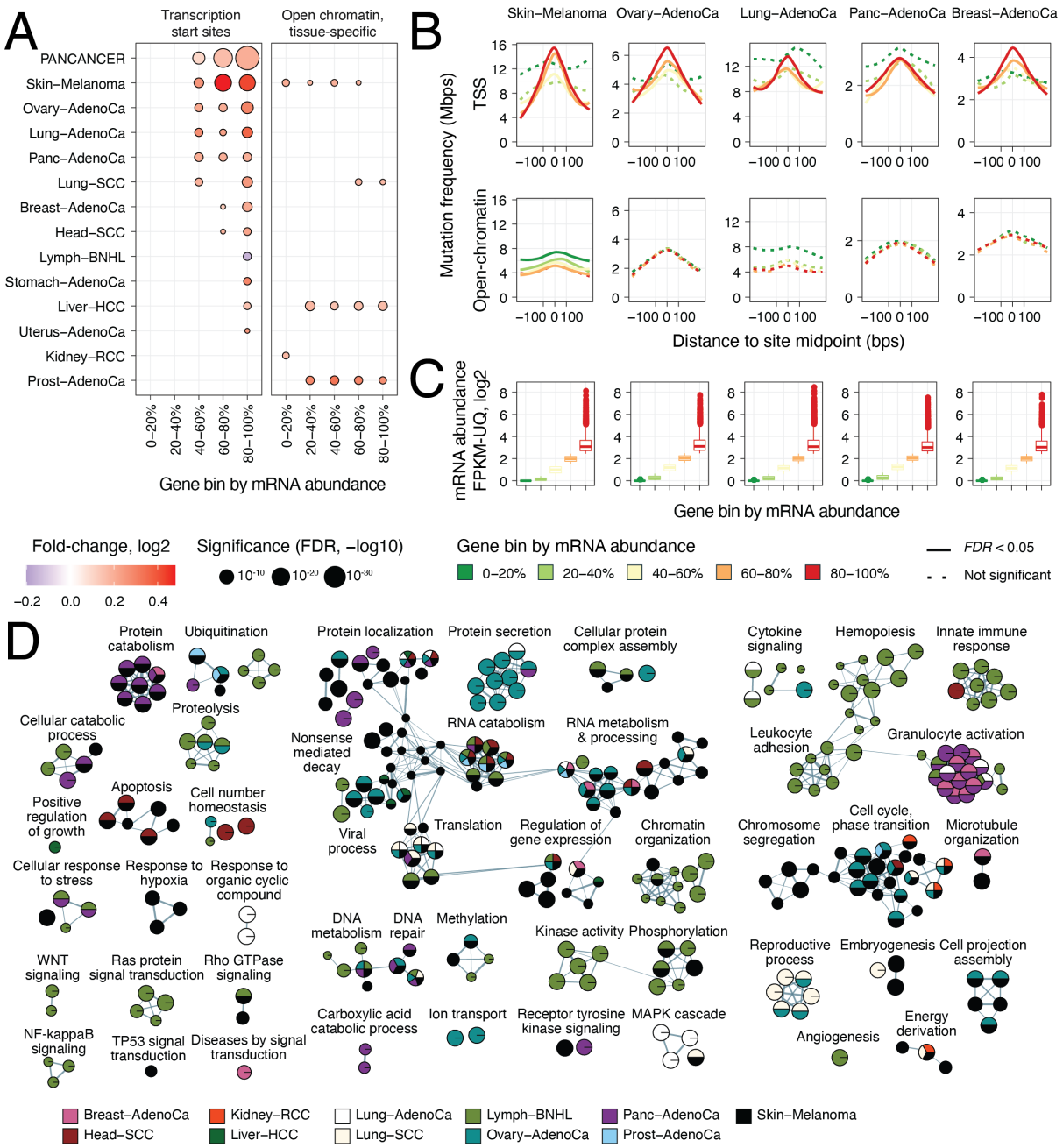

Figure 3. Mutational enrichment at transcription start sites associates with mRNA abundance of target genes and diverse pathways. A. Comparison of mutation frequencies in TSSs and tissue-specific open-chromatin sites grouped by mRNA abundance of target genes in matching tumors (FDR < 0.05). Equal numbers of sites were sampled for an unbiased comparison. **B.** Examples of cancer types with strong transcription-associated mutagenesis in TSSs (top) compared to open-chromatin sites (bottom). Mutation frequencies are shown on the Y-axis with loess smoothing. **C.** Median mRNA abundance of genes in the five bins. **D.** Enrichment map of pathways and processes with frequent mutations at TSSs (FDR < 0.05). Nodes represent pathways and processes that are connected with edges if these include many common genes.

249 **Mutational enrichment at transcription start sites associates with mRNA abundance of**
250 **target genes and diverse pathways**

251 We asked whether the increased mutation frequency at TSSs and open-chromatin sites correlated
252 with transcription of target genes. We used matched RNA-seq data available for 20 cancer types
253 in PCAWG [45] to quantify the tissue-specific activity of regulatory regions. Protein-coding genes
254 were distributed into five equally-sized bins based on their median mRNA abundance in each
255 cancer type such that the first bin included silent genes and the fifth bin included a wide range of
256 highly transcribed genes. Open-chromatin sites were assigned to genes based on their location in
257 gene bodies and promoters, as well as consensus long-range chromatin interactions in promoter-
258 capture Hi-C experiments [46]. This provided a tissue-specific map of TSSs and open-chromatin
259 sites with transcript abundance as a measure of site activity.

260 Higher mutation frequencies in TSSs strongly associated with tissue-specific mRNA abundance
261 of target genes (**Figure 3A**). In the fifth bin of highly transcribed genes, TSSs were consistently
262 enriched in mutations in 12/20 cancer types (*FDR* < 0.05) and the pan-cancer cohort (*FDR* = 6.2
263 x 10^{-38}, FC = 1.15) (**Supplementary Table 1C**). The strongest associations of transcriptional
264 activity and TSS-specific mutagenesis were found in melanoma, ovarian, lung, pancreatic and
265 breast cancer (*FDR* ≤ 10^{-4}, FC ≥ 1.15) (**Figure 3B,C**). The TSSs of genes with intermediate
266 transcript abundance (40-80 percentile) were also enriched in mutations in seven cancer types. In
267 contrast, TSSs of silent and low-abundance genes were not differentially mutated compared to
268 flanking controls. Interestingly, the overall mutation frequency was higher in and around silent
269 TSSs, suggesting that closed chromatin is exposed to lower levels of DNA repair. Of note, the
270 top gene bin was highly variable in mRNA abundance (*e.g.*, range 11—7100 FPKM-UQ in the
271 pan-cancer cohort) **(Figure 3C)**. Therefore, the mutational enrichment in the top gene bin may
272 be partially contributed to by genes with very high expression. TSS bins with higher transcript
273 abundance were also enriched in mutational signatures commonly observed in their respective
274 cancer types (**Supplementary Figure 3A**). This analysis highlights a pan-cancer mutational
275 process localised in TSSs that appears to be driven by transcriptional activity.

276 Compared to TSSs, mutation frequencies in open-chromatin sites were attenuated and did not
277 associate with transcript abundance of target genes (**Figure 3A,B**). Mutational enrichments in

278 multiple gene bins were observed in melanoma, liver cancer and prostate cancer, however the
279 effect sizes were smaller, and most cancer types showed no significant changes in local mutation
280 frequencies. In this unbiased comparison, we included equally-sized bins of TSSs and open-
281 chromatin sites using down-sampling, since open-chromatin sites considerably outnumbered
282 TSSs. An extended analysis of all open-chromatin sites grouped by transcript abundance showed
283 stronger mutational enrichments in most cancer types that is consistent with our analysis above
284 (**Supplementary Figure 2B**), however, no associations with transcript abundance were apparent.
285 This analysis suggests that transcription-associated local mutagenesis is more prominent in TSSs
286 than in general regions of open chromatin.

287 To explore the functional associations of elevated mutation frequencies at TSSs, we performed a
288 pathway enrichment analysis by adapting RM2 to gene sets of GO biological processes and
289 Reactome molecular pathways [47]. This allowed us to test our hypothesis that the TSSs enriched
290 for mutations are concentrated in specific biological processes. We found 336 unique pathways
291 and processes with pronounced enrichments of mutations at TSSs in eleven cancer types (*FDR* <
292 0.05) (**Figure 3D**) (**Supplementary Table 1D**). Half of the pathways were found in at least the
293 melanoma cohort (51%). However, one third of the pathways were found in two or more cancer
294 types, indicating that the pathway associations of elevated mutagenesis at TSSs often apply more
295 generally to multiple cancer types. Translation, ribosome biogenesis and RNA processing were
296 among the largest groups of pathways found. This is expected as the translational machinery is
297 ubiquitously active in proliferating cells and includes many highly expressed genes. Processes
298 related to gene regulation and chromatin organization were also detected. Besides these general
299 housekeeping processes, cancer-related processes and pathways were also enriched in TSS
300 mutations. For example, mitotic cell cycle, apoptosis, DNA repair, angiogenesis, developmental
301 and immune response processes, as well as druggable signalling pathways (*e.g.*, MAPK, Wnt,
302 Notch) were identified in multiple cancer types. The pathway analysis supports our findings of
303 frequent TSS mutations associated with increased transcription. It also highlights a variety of
304 core cellular processes and cancer pathways where mutations accumulate in core promoters
305 across multiple cancer types and may have functional consequences.

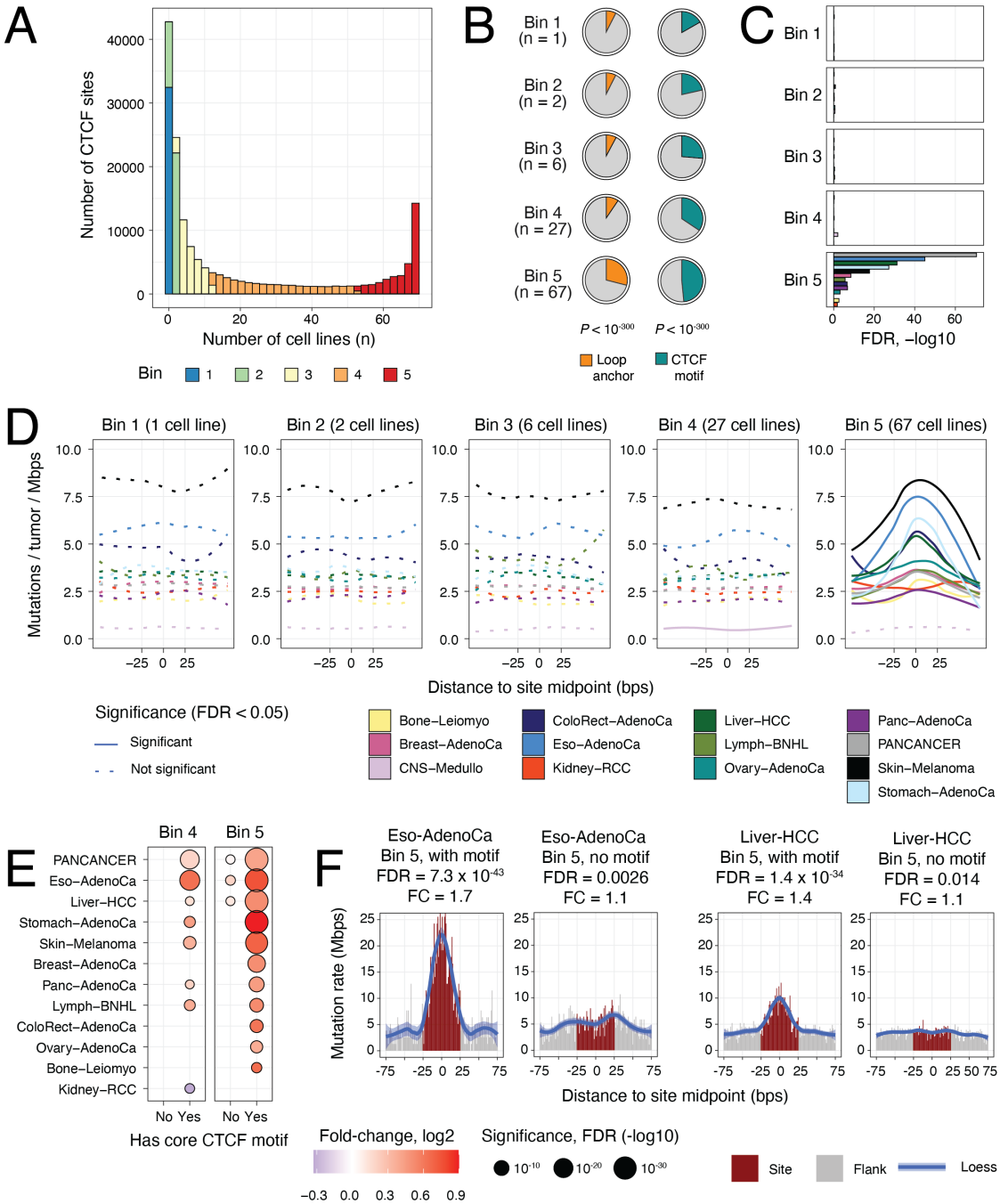

Figure 4. Localised mutational processes at constitutively active binding sites of CTCF. A. Histogram of CTCF binding sites with number of cell lines in ENCODE. Sites were grouped as five equal bins based on conservation across cell lines (colors). Bimodal distribution reveals a subset of sites detected in most or all cell types (in red). **B.** Pie charts show the proportion of CTCF sites in the five bins located at chromatin loop anchors (left) and that matched the core CTCF motif (right). P-values represent the enrichments in the 5th bin compared to all sites (Fisher's exact test). The median numbers of cell lines per bin are shown in brackets. **C.** Significance of the localised mutational enrichments in the five bins of CTCF binding sites. FDR values of the RM2 analysis are shown on the X-axis. Colors correspond to cancer types. **D.** Local mutation frequency in the five bins of CTCF binding sites. Solid lines show statistically significant changes in mutation frequencies compared to flanking controls (RM2 FDR < 0.05). Loess curves were used for smoothing. **E.** Significance of mutational enrichment in highly conserved CTCF binding sites (bins 4-5) grouped by presence or absence of core DNA-binding motif of CTCF in the sites. **F.** Mutations in constitutively bound subsets in CTCF binding sites with and without the core CTCF DNA motifs. Esophageal cancer (left) and liver cancer (right) are compared.

307 **Mutational enrichment at CTCF binding sites is associated with constitutive DNA-binding**
308 **and core sequence motifs**

309 We tested whether the elevated mutation frequencies in CTCF binding sites were associated with
310 functional site characteristics. We used the extent of conservation of DNA-binding across 70 cell
311 lines as a proxy of site activity. We grouped 162,209 unique CTCF binding sites catalogued in
312 ENCODE into five equal bins such that the first bin contained sites observed in a single cell line
313 while the fifth bin included constitutively bound sites with the median site found in 67/70 cell
314 lines (**Figure 4A**). Doing so allowed us to test mutation frequency and genomic associations at
315 different levels of site activity. The constitutively bound CTCF binding sites were enriched in
316 chromatin loop anchors [48] (29% observed *vs.* 12% expected, Fisher's exact $P < 10^{-300}$) and
317 consensus core CTCF DNA-binding motifs (48% observed *vs*. 29% expected, $P < 10^{-300}$)
318 (**Figure 4B**). Together, these features imply that the constitutive bin of sites represents
319 functionally important CTCF binding sites.

320 The constitutive CTCF binding sites were characterised by elevated mutational frequencies in the
321 pan-cancer cohort ($FDR = 1.4 \times 10^{-71}$, FC = 1.18) and in ten cancer types, especially stomach,
322 liver and esophageal cancers as well as melanoma ($FDR \leq 10^{-19}$, FC ≥ 1.25) (**Figure 4C,D**)
323 (**Supplementary Table 1E**). In contrast, all other bins of CTCF sites showed no significant
324 enrichment of mutations ($FDR > 0.05$). Specifically, none of the cancer types showed mutational
325 enrichments in the sites of the fourth bin which still represent a high level of conservation of
326 CTFC binding (median 27/70 cell lines). Mutational signature analysis revealed over-represented
327 signatures such as SBS3, SBS5, SBS7, SBS17 and SBS40 at these sites in various cancer types,
328 however a common signature for constitutive CTCF binding was not apparent (**Supplementary**
329 **Figure 3B**). In summary, this analysis emphasises the potential role of constitutive CTCF
330 binding activity in local mutagenesis.

331 We then examined whether the presence of CTCF DNA-binding motifs in the ChIP-seq peaks
332 played a role in localised mutagenesis. We categorised all the CTCF binding sites into two
333 subsets based on the presence or absence of a consensus core CTCF DNA-binding motif and
334 repeated the conservation-based grouping of the two subsets of sites. The localised mutational
335 process at CTCF binding sites appears to be driven by the combination of constitutive binding
336 and core motif presence (**Figure 4E**) (**Supplementary Table 1F,G**). The constitutively bound

337 CTCF sites matching the DNA motifs were strongly enriched in mutations in ten cancer types
338 and the pan-cancer cohort. In contrast, the constitutively-bound sites lacking the core DNA motif
339 were not enriched. For example, esophageal cancer showed a strong mutational enrichment in
340 the constitutively-bound CTCF sites with the core motif present (*FDR* = 7.3 x 10^{-43}, FC = 1.70),
341 while the signal was clearly attenuated in the constitutively-bound sites with no core motif (*FDR*
342 = 0.0026, FC = 1.14) (**Figure 4F**). This was confirmed in liver cancer (bin 5 with motifs: *FDR* =
343 1.4 x 10^{-34}, FC = 1.44; bin 5 without motifs: *FDR* = 0.014, FC = 1.08). Highly conserved sites
344 with median conservation in 27/70 cell lines also showed motif-driven mutational enrichments.
345 Strikingly, no signal of mutational enrichment was detected in any other bins of sites regardless
346 of motif presence (**Supplementary Figure 4A**). The findings were confirmed in a tissue-specific
347 analysis integrating breast and liver cancer mutations with CTCF binding sites of the cancer cell
348 lines MCF-7 and HepG2, respectively (**Supplementary Figure 4B,C**).

349 Our analysis highlights a minority of CTCF binding sites (9,552 or 5.9%) with constitutive
350 CTCF binding and core DNA motifs that are the primary target of a local mutational process in
351 many cancer types. This analysis refines the established pattern of mutational enrichment in
352 CTCF binding sites by defining a subset of frequently-mutated sites with specific functional
353 properties, such as involvement in chromatin architectural and gene regulatory interactions [48,49].
354 Our observations are consistent with previous findings of CTCF motifs in sites shared among a
355 few cell lines that comprised a particularly strong mutational enrichment indicating higher site
356 activity within cells [50,51]. Conservation is a property of functionally integral CTCF binding sites,
357 which upon disruption, can lead to changes in underlying chromatin architecture and gene
358 regulation [52] and is associated with activation of proto-oncogenes [26]. Therefore, the accumulation
359 of mutations at those sites may have functional consequences in cancer.

360

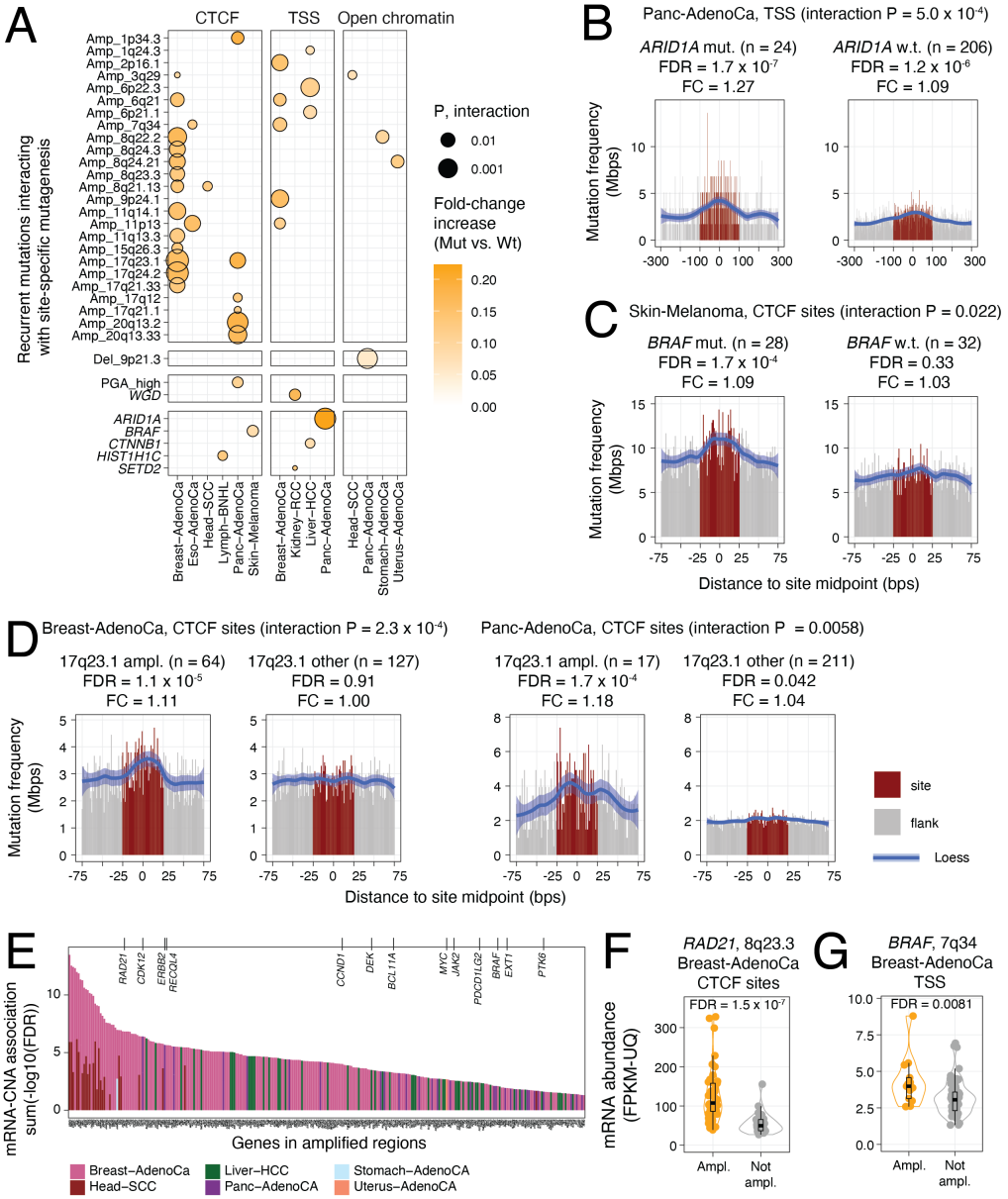

Figure 5. Recurrent driver mutations and copy-number alterations (CNA) associate with localised mutagenesis. A. Dotplot of the statistical interactions of recurrent mutations with increased mutation frequencies at sites (RM2 FDR < 0.05; interaction P < 0.05). Whole-genome duplication (WGD) and CNA burded (median-dichotomised percent genome altered, PGA) are also shown. **B-D.** Examples of increased mutational frequencies associating with recurrent mutations. Tumors with and without recurrent mutations are shown (left *vs.* right). **B.** *ARID1A* mutations in pancreatic cancer associate with enriched mutations at TSSs. **C.** *BRAF* mutations in melanoma associate with enriched mutations at CTCF binding sites. **D.** 17q23.1 amplifications associate with enriched mutations at CTCF binding sites in breast and pancreatic cancer. **E.** Copy-number amplified genes with amplification-driven increases in mRNA abundance. Known cancer genes are shown at the top. **F.** *RAD21* was up-regulated in the set of 8q23.3-amplified breast cancers associated with enriched mutations in CTCF binding sites. **G.** *BRAF* was up-regulated in the set of 7q34-amplified breast cancers associated with enriched mutations in TSSs.

363 **Recurrent driver mutations and copy-number alterations associate with localised**
364 **mutagenesis**

365 To find potential genetic mechanisms of localised mutagenesis, we tested whether the presence
366 of specific features such as recurrent mutations in tumor genomes associated with higher
367 mutation frequencies at gene-regulatory and architectural elements. To enable this analysis, we
368 collected 37 driver genes with frequent SNVs and indels predicted using the ActiveDriverWGS
369 method [34], 70 recurrent copy-number alterations (CNAs) detected in the PCAWG project using
370 the GISTIC2 method [53], as well as two genome-wide measures of aneuploidy, whole-genome
371 duplication (WGD) and percent genome altered (PGA) (**Supplementary Figure 5**). Integrative
372 analysis with RM2 identified 41 significant interactions between genomic features and site-
373 specific elevations in mutation frequencies in nine cancer types, including five driver genes
374 (*ARID1A*, *BRAF*, *CTNNB1*, *HIST1H1C*, *SETD2*), 25 recurrently amplified loci and one deletion
375 (RM2 $FDR < 0.05$, interaction $P < 0.05$) (**Figure 5A**) (**Supplementary Table 1G**).

376 *ARID1A* mutations in pancreatic adenocarcinoma had one of the strongest interactions with
377 localised mutational enrichment (interaction $P = 5.0 \times 10^{-4}$) (**Figure 5B**). The 24 pancreatic
378 cancer genomes with *ARID1A* mutations showed a significant increase in mutation frequency at
379 TSSs (FDR = 1.7×10^{-7}, FC = 1.27). In contrast, the majority of tumors (206) lacking *ARID1A*
380 mutations showed a weaker, albeit significant, mutational enrichment at TSSs, likely explained
381 by the improved statistical power of the larger set of samples analysed (FDR = 1.2×10^{-6}, FC =
382 1.09). No significant interactions were seen with mutations at TSSs and open-chromatin sites
383 (**Supplementary Figure 6A**). *ARID1A* is a tumor suppressor encoding a member of the
384 SWI/SNF chromatin remodeling complex that regulates chromatin accessibility at regulatory
385 elements and is involved in the maintenance of genomic stability and DNA repair [54]. *ARID1A*
386 interacts with the topoisomerase TOP2A to facilitate its DNA binding and DNA decatenation, a
387 process required for proper chromosome segregation in mitosis [55]. The SWI/SNF complex is also
388 involved in DNA double-strand break repair [56] and nucleotide excision repair [57], and loss of
389 *ARID1A* impairs polymerase pausing [58]. *ARID1A* is also one of the most frequently mutated
390 genes in pancreatic adenocarcinoma [59] and in the PCAWG cohort, as 16/24 of tumors carried
391 frameshift or stop-gain mutations that suggest loss of function. The inactivating *ARID1A*

392 mutations may increase the mutation burden at TSSs as a result of impaired repair mechanisms
393 and increased instability, and consequently, further contribute to transcriptional dysregulation.

394 In melanoma, driver mutations in *BRAF* were associated with mutational enrichment in CTCF
395 binding sites (interaction $P = 0.022$) (**Figure 5C**). The 28 tumors with *BRAF* mutations were
396 enriched in CTCF binding site mutations ($FDR = 5.8 \times 10^{-5}$, FC = 1.10), while 32 *BRAF*-
397 wildtype tumors showed no enrichment ($FDR = 0.31$, FC = 1.03). BRAF serine/threonine kinase
398 is a proto-oncogene and the activating V600E mutation defines a druggable subtype of
399 melanoma [60,61]. Previous work shows that ectopic expression of V600E-mutant *BRAF* in
400 epithelial cell lines induces DNA double stranded breaks and the production of reactive oxygen
401 species [62]. In this cohort, 22/28 melanomas carried V600E substitutions and V600K substitutions
402 occurred in three additional tumors. No significant interactions of *BRAF* mutations and localised
403 mutagenesis were found for TSSs and open-chromatin sites (**Supplementary Figure 6B**). This
404 analysis suggests that melanomas defined by *BRAF* driver mutations may have increased activity
405 of a mutational process acting specifically on CTCF binding sites.

406 Amplifications of the 17q23.1 locus associated with an increased mutation frequency in CTCF
407 binding sites in breast cancer (interaction $P = 2.3 \times 10^{-4}$) (**Figure 5D**). The 64 tumors with
408 amplifications were enriched in mutations in CTCF binding sites (FDR = 1.1×10^{-5}, FC = 1.11)
409 while 127 tumors lacking the amplifications showed no enrichment (FDR = 0.91, FC = 1.00).
410 This highest-ranking interaction was also confirmed in the pancreatic cancer cohort (P = 0.0058)
411 where 17 tumors with amplifications showed significantly more mutations at CTCF binding sites
412 (FDR = 1.7×10^{-4}, FC = 1.18) compared to the 211 tumors lacking the amplifications (FDR =
413 0.042, FC = 1.04). Besides 17q23.1, we found 12 amplified loci in breast cancer with significant
414 interactions with CTCF binding sites to potentially explain these effects. Also, chromosomal
415 instability has been associated with mutagenesis of CTCF binding sites in gastrointestinal
416 cancers [28]. However, no significant interactions of high aneuploidy and CTCF binding site
417 mutations for breast cancer were detected in our models. Discovery of the interaction in two
418 cancer types leads to the speculation that the 17q23.1 locus is involved in localised mutagenesis
419 through unknown mechanisms, however more work is needed to investigate this hypothesis in
420 detail.

421 To further decipher the interactions of recurrent genomic amplifications and local mutagenesis,
422 we determined the genes located in the amplified loci that responded transcriptionally to
423 amplifications. Using a cancer-type specific analysis, we found 282 unique genes in 25 amplified
424 regions that were significantly up-regulated in the tumors with these amplifications (Wilcoxon
425 test, *FDR* < 0.05) (**Figure 5E**) (**Supplementary Table 1I**). mRNA abundance associations were
426 identified for all three types of genomic sites and six cancer types. 13 known cancer genes were
427 found (*RAD21*, *CDK12*, *ERBB2*, *RECQL4*, *CCND1, DEK, BCL11A, MYC, JAK2*, *PDCD1LG2,*
428 *BRAF, EXT1, PTK6*).

429 In breast cancer, *RAD21* mRNA abundance was significantly higher in 8q23.3-amplified tumors
430 compared to non-amplified tumors (*FDR* = 1.6×10^{-7}) (**Figure 5F**), indicating that *RAD21* up-
431 regulation is driven in part by the genomic amplification. The 8q23.3 amplification was
432 associated with mutational enrichments in CTCF binding sites in breast cancer (*P* = 0.0089)
433 (**Supplementary Figure 7A**). *RAD21* encodes a subunit of the cohesin complex that co-binds
434 DNA with CTCF to orchestrate transcriptional insulation and chromatin architectural interactions
435 [48,49], suggesting its role in CTCF-related mutagenesis.

436 As another example, *BRAF* was up-regulated in the subset of breast cancers with 7q34
437 amplifications (*FDR* = 0.0081) (**Figure 5G**). The amplification was associated with increased
438 mutagenesis at TSSs in breast cancer (interaction *P* = 0.012) (**Supplementary Figure 7B**). The
439 7q34 amplification was also significantly associated with CTCF binding site mutations in
440 esophageal cancer (interaction *P* = 0.034) (**Supplementary Figure 7B**), however no matching
441 RNA-seq data was available to confirm the transcriptional upregulation of *BRAF*. Thus, our
442 analysis associates *BRAF* amplifications and point mutations to localised mutagenesis in multiple
443 cancer types.

444 In summary, this integrative analysis of mutational processes with recurrent driver mutations,
445 copy number alterations and gene expression data provides a catalogue of somatic alterations
446 that associate with localised mutational enrichments in gene-regulatory and chromatin
447 architectural elements of the cancer genome. Since our evidence remains correlative, further
448 computational and experimental approaches are needed to establish causal relationships. While a
449 subset of these driver mutations and recurrent copy-number amplifications may be directly

450 involved in mutagenesis and DNA repair, others may represent markers of tumor subtypes with
451 specific exposures or endogenous factors. Deeper study of these associations may help decipher
452 genetic mechanisms underlying mutational processes.

453

454

455 **Discussion**

456 The cancer genome is molded by diverse mutational processes that continuously shape its broad
457 megabase-scale features and the fine context of single nucleotides. Here we focused on the
458 mutational processes of an intermediate scale that affect thousands of genomic elements, each
459 spanning tens to hundreds of nucleotides. Such functional elements are widespread in the non-
460 coding genome and play various roles in gene expression control and chromatin architecture.
461 Using our novel computational framework, we characterised the mutational landscape of gene-
462 regulatory and chromatin architectural elements in diverse cancer types and identified their
463 putative functional and genetic determinants. We show that the ubiquitous mutational enrichment
464 in TSSs is associated with high transcription and with biological processes involved in cellular
465 housekeeping, but importantly with a variety of pathways and processes implicated in cancer. In
466 contrast, transcriptionally silent genes showed no local deviations in mutation frequencies. On
467 the other hand, open-chromatin regions were generally enriched in mutations and showed no
468 association with transcript abundance. CTCF binding sites exhibited a particular pattern of site-
469 specific mutagenesis: a small fraction of hypermutated sites defined by a core CTCF DNA motif
470 and constitutive binding across tens of human cell types dominated over the majority of sites that
471 lacked mutational enrichment. Lastly, an integrative analysis of site-specific mutation
472 frequencies with recurrent alterations in cancer genomes allowed us to predict genetic drivers of
473 localised mutational processes. The association of driver mutations in *ARID1A* and *BRAF* with
474 enrichment of site-specific mutations corroborate earlier functional and mechanistic evidence of
475 mutagenesis and DNA repair disruption, lending confidence to our other identified interactions
476 with driver genes and genomic amplifications, and providing mechanistic hypotheses for future
477 studies.

478 Open-chromatin regions were broadly enriched in mutations in most cancer types and the effect
479 appeared to be independent of transcript abundance of predicted target genes. These findings
480 contrast an earlier report that indicated decreased mutation rates in open-chromatin sites of cell
481 lines at a similar resolution of the genome [29]. The use of non-matched chromatin data of cell
482 lines and fewer cancer genomes may explain the depletion signals observed earlier that may be
483 confounded by tissue-specific properties of chromatin state and mutagenesis. Our analysis of the
484 open-chromatin regions of matching cancer types and a larger set of whole cancer genomes

485 likely provides a more accurate view of mutational processes. In our study, the pan-cancer
486 analysis of open-chromatin sites also revealed only attenuated effects of total mutational
487 enrichments and several mutational signatures, as well as indels appeared depleted in sites
488 compared to flanking regions. Contrarily, interpreting individual cancer types with maps of
489 cancer-specific open-chromatin sites revealed robust enrichments of mutations at open-
490 chromatin sites. This highlights the importance of using matched genomic, epigenomic and
491 transcriptomic profiles to study localised mutational processes.

492 We speculate that the local mutational enrichments in gene-regulatory and chromatin
493 architectural sites represent a functional continuum of passenger and driver mutations. On the
494 one hand, the vast majority of mutations enriched across a class of sites likely represent
495 functionally neutral passenger mutations whose frequent occurrence is explained by differences
496 in mutagenesis, DNA repair or carcinogen exposure. For example, previous studies of promoter-
497 specific mutational enrichments in melanoma did not find broad associations of mutations and
498 transcriptional deregulation of target genes [63], suggesting a lack of immediate functional impact.
499 On the other hand, a small subset of genomic elements directly involved in the transcriptional or
500 epigenetic control of hallmark cancer pathways may accumulate functional mutations either by
501 chance or through the elevated activity of site-specific mutational processes. Those may show
502 positive selection at an individual site or across a set of functionally-related sites. For example,
503 while recent large-scale studies of whole cancer genomes have found relatively few individual
504 non-coding regions with driver mutations [4], additional low-frequency candidate drivers were
505 shown to converge onto the regulatory elements of cancer-related biological processes and
506 protein interaction networks [37]. Other recent studies have combined computational analyses and
507 functional validation experiments to nominate new non-coding drivers in cis-regulatory modules
508 and CTCF binding sites, and to associate these with transcriptional deregulation of cancer
509 pathways [34-36]. This suggests that our mechanistic understanding of oncogenesis, progression and
510 metastasis pathways may be refined by analysing the non-coding genome. A better
511 understanding of localised mutagenesis will help deconvolute the effects of mutational processes
512 and positive selection and contribute to cancer driver discovery in the non-coding genome.

513 Our analysis has certain caveats and limitations. We analysed a broad catalogue of genomic
514 elements that provides a limited representation of the heterogeneous dataset of tumor genomes.

515 We used several strategies to prioritise functionally active elements in a tissue-specific manner:
516 TSSs were grouped by target gene transcription in the matching tumor samples, open-chromatin
517 sites were selected from genome-wide profiles of cancers of matching types or related normal
518 tissues, and CTCF binding sites were grouped by their binding conservation across a large panel
519 of human cell types. To better address tumor heterogeneity, future analyses will benefit from
520 detailed multi-omics cohorts where precisely matching genomic, transcriptomic and epigenomic
521 profiles of individual tumors are available. Also, the current method is designed for genomic
522 elements of uniform width and it is not directly applicable to elements of variable width, such as
523 exons or non-coding RNAs. Our analysis suggests that different classes of gene-regulatory and
524 architectural elements of the genome may be subject to localised mutational processes that have
525 footprints of different sizes. Thus, it is recommended to evaluate the relevant input parameter of
526 the method when analysing new classes of genomic elements. Our method is designed to
527 quantify localised differences of mutation frequencies acting on an entire class of genomic
528 elements with thousands to hundreds of thousands of genomic loci. It is not powered to evaluate
529 a single genomic element as a potential cancer driver and alternative methods should be used for
530 this purpose. However, we have adapted our method to evaluate TSS-specific mutation rates in
531 gene sets of biological processes and pathways with hundreds to thousands of genes.

532 Our study opens new avenues for future developments. Integrative analysis of whole cancer
533 genome sequences and rich clinical and pathological profiles of tumors [17] may highlight
534 associations of clinical variables and localised mutagenesis and thus lead to the discovery of
535 novel WGS-based biomarkers. Considering patient lifestyle information, environmental
536 exposures and germline variation in the analysis may elucidate the impact of carcinogens and
537 endogenous DNA repair deficiencies. Our catalogue of genetic associations provides hypotheses
538 on mutational mechanisms that can be tested experimentally using genome editing and
539 mutagenesis assays. Rare germline variants in the human population [64], *de novo* variants detected
540 in genetic disorders [21] and the widespread somatic genome variation found in healthy tissues [65]
541 provide further avenues to study mutational processes acting at functional non-coding elements.
542 Our study provides a detailed annotation of localised mutational processes in whole genomes and
543 enables future work to decipher the interplay of local mutagenesis and cancer driver
544 mechanisms, molecular heterogeneity and genome evolution.

545

Methods

Regression models for Localised Mutations (RM2). Local differences in mutation rates in functional genomic elements (*i.e.*, sites) were evaluated using a negative binomial regression model we refer to as RM2. Single nucleotide variants (SNVs) and small insertions-deletions (indels) were analysed. The model simultaneously considers a collection of sites, such as regulatory elements that are commonly ~10–1,000 bps in length and measured in ChIP-seq and related experimental assays in thousands to hundreds of thousands of genomic loci. Sites were uniformly redefined using their median coordinate and added sequences of fixed width upstream and downstream of the sites (*e.g.*, ±25 bps or 50 bps around the midpoints of CTCF binding sites). Upstream and downstream flanking sequences of these sites were used as control regions to estimate expected mutation rates. Control regions were defined to be of equal width to sites such that the upstream and downstream regions combined were twice as wide as the sites initially. Next, site and flank sequences were pooled, duplicate sequences were removed, and overlapping coordinates of sites were removed from the flanking set such that every nucleotide and every mutation was considered as either as part of a site or a flanking control sequence but never both. The pooling and deduplication steps were used to account for groups of adjacent sites. To account for megabase-scale variation in mutation frequencies, we computed the total log-transformed mutation count for each site within its one-megabase window (i.e., ±0.5 Mbps around site midpoint). Based on this estimate, all sites were distributed into ten equal bins (*MbpRate*). The value of ten bins worked well in our benchmarks and captured variation in smaller and larger cohorts of individual cancer types. However, custom values of this parameter can be used. Mutation rates for sites and flanking sequences for each bin were defined separately and were distinguished by a binary cofactor (*isSite*). Sequence positions were counted separately by their trinucleotide context (*nPosits*) and expanded to three alternative nucleotides to account for the potential sequence space where such single nucleotide variants could occur (*nPosits*). The observed mutations in these contexts were also counted (*nMuts*) and a cofactor was used to add separate weights to different trinucleotide classes (*i.e.,* reference trinucleotide and alternative nucleotide; *triNucMutClass*). Indels were counted under another entry in *triNucMutClass* such that all mutation counts were summed, and the entire genomic space was accounted for. An optional binary cofactor (*coFac*) was included to allow the consideration of genetic or clinical

576 covariates of localised mutation rates. To evaluate the significance of localised mutation rates in
577 sites compared to flanking control regions, we first constructed a null model that excluded the
578 term *isSite*:

579 H_{null} *:* NegBin(*nMut* ~ offset(log(*nPosits*)) + *triNucMutClass* + log1p(*MbpRate*) + *coFac*).

580 The main model representing the alternative hypothesis of a site-specific mutation rate was
581 constructed as follows:

582 H_{alt} *:* NegBin(*nMut* ~ offset(log(*nPosits*)) + *triNucMutClass* + log1p(*MbpRate*) + *coFac* +
583 *isSite*).

584 We extended our model to evaluate whether localised mutation rates differ between two subtypes
585 of tumors, such as those defined by clinical annotations or genetic features using the term *coFac.*
586 Trinucleotide sequence content, trinucleotide mutational signatures and megabase-scale
587 covariations of mutation rates were computed separately for the two sets of tumors. To establish
588 the associations of localised mutation rates and tumor subtypes (or presence of driver mutations),
589 we added to the initial model the term *isSite:coFac* mapping the interaction of the tumor subtype
590 and the cofactor distinguishing sites and flanking sequences, as follows:

591 H_{cof} *:* NegBin(*nMut* ~ offset(log(*nPosits*)) + *triNucMutClass* + log1p(*MbpRate*) +

592 *coFac* + *isSite* + *isSite:coFac*).

593 We used likelihood ratio tests to compare the models and evaluate the significance of localised
594 mutation frequencies (H_{alt} *vs.* H_{null} to evaluate the term *isSite*). Chi-square *P*-values from the
595 likelihood ratio tests were reported for each analysis. We also reported coefficient values of the
596 term *isSite* to characterise enrichment or depletion of mutations at sites relative to flanking
597 controls for positive and negative values, respectively. The interactions of driver mutations and
598 mutation rates were evaluated using likelihood ratio tests that compared the models H_{alt} and H_{cof}.
599 Only the models with significant positive coefficients were reported. The expected mutation
600 counts were derived from each model by 1000-fold sampling of mutation counts from the
601 negative binomial distribution informed by the fitted probabilities and theta values derived from

602 the regression models. Fold-change values were derived by dividing median observed and
603 expected mutation counts, and confidence intervals were derived using the 2.5th and 97.5th
604 percentiles of sampled values. Chi-square *P*-values from the models were corrected for multiple
605 testing using the Benjamini-Hochberg procedure. As an exception, unadjusted P-values were
606 used to quantify the interactions of local mutation frequencies and tumor subtype (*e.g.*, presence
607 of driver mutation) since these analyses were conditional of the significance of main site-specific
608 effects. Besides modelling total mutations in sites and flanking sequences, we evaluated
609 mutations of multiple subclasses, such as mutations stratified by transcriptional activity,
610 COSMIC mutational signatures or DNA strands. Mutation subclass analysis was conducted as
611 described above. The same megabase-scale mutation rates estimated for all mutations were used
612 rather than those of specific subclasses. The method is available at
613 https://github.com/reimandlab/RM2.

614 **Somatic mutations in whole cancer genomes.** Somatic single nucleotide variants (SNVs) and
615 short insertions-deletions (indels) in the genomes of 2,583 primary tumors were retrieved from
616 the uniformly processed dataset of the Pan-cancer Analysis of Whole Genomes (PCAWG)
617 project of the ICGC and TCGA [3]. We used consensus variant calls mapped to the human genome
618 version GRCh37 (hg19). We removed 69 hypermutated tumors with at least 90,000 mutations,
619 resulting in 2,514 tumors and 24.7 million mutations. We also removed 33 tumors for which
620 mutational signature predictions were not available in PCAWG. We analysed tumor genomes of
621 the pooled pan-cancer cohort of multiple cancer types, and also 25 cohorts of specific cancer
622 types with at least 25 samples in the PCAWG cohort. We excluded a small subset of tumors in
623 which localised mutation rates were exceptionally strong even when analysing one tumor
624 genome at a time ($FDR < 0.001$, RM2). The small subset of filtered tumors included 33 tumors
625 (1.4% of the cohort) that provided 38 findings of elevated site-specific mutation frequencies of
626 the three site classes. The outlier tumors showed overall higher mutation burden compared to the
627 remaining pan-cancer cohort (mean 50,238 *vs.* 9,404 mutations). We performed this filtering
628 because based on our initial analyses, we found that the individual contribution of these tumors
629 to the overall analysis would have caused overestimates of mutation rates. The filtering led to a
630 conservative analysis that only revealed associations shared across multiple tumor genomes. To
631 enable this filtering, we performed tumor specific analyses for the three classes of sites (open-

632 chromatin sites, CTCF binding sites, and TSSs). We analysed each cohort of a cancer type
633 separately and grouped the mutations according to tumor ID, allowing the model to learn an
634 expected background mutation rate in the respective cohort and then test each tumor genome
635 separately for localised mutation rates. To perform this single-tumor analysis in smaller cohorts
636 within the PCAWG dataset (<25 tumors of a given type), we created a meta-cohort by pooling
637 these smaller cohorts. After filtering hypermutated tumors, tumors without PCAWG signatures,
638 and tumors with exceptionally strong signals of localised mutations, we derived a conservative
639 final set of 2,419 genomes of 35 cancer types with 22.7 million mutations including 1.61 million
640 indels. To evaluate the performance of our model, we also processed a dataset of simulated
641 variant calls for the same set of tumors derived from the PCAWG project (*i.e.*, the Broad dataset)
642 [4]. The true and simulated datasets were compared for method benchmarking (see below). In
643 addition to evaluating total mutations, several classes of mutations were analysed separately.
644 Mutations were mapped to C and T nucleotides and grouped by reference and alternative
645 nucleotides (C>[A,G,T], T>[A,C,G]). Mutations were also classified as located either on the
646 Watson (w) strand if the original reference nucleotide was C or T, or the Crick (c) strand if the
647 original reference nucleotide was A or G. We also classified mutations by the trinucleotide
648 signatures of single base substitutions (SBS) that were derived earlier using the SigProfiler
649 software in the PCAWG project [14]. We assigned each mutation to its most probable signature in
650 the given patient tumor based on its trinucleotide context. For model evaluation, mutations were
651 also annotated for the simulated dataset.

652 **Chromatin architectural and gene-regulatory genomic elements.** We performed a systematic
653 analysis of three classes of genomic elements: DNA-binding sites of CTCF (CCCTC-binding
654 factor) detected in multiple human cell lines, transcription start sites (TSS) of protein-coding
655 genes, and open-chromatin sites (ATAC-seq sites and DNAse-seq sites) detected in human
656 primary tumors or related normal tissues. CTCF binding sites were retrieved from the ENCODE
657 project [39]. Sites observed in only one cell line were removed for the majority of the study,
658 resulting in 119,464 sites across 70 cell lines. As an exception, the full set of 162,209 CTCF
659 binding sites in ENCODE was used for the analysis of CTCF binding sites that considered site
660 conservation across cell lines (Figure 4), as discussed below. TSS loci of protein-coding genes
661 were retrieved from Ensembl Biomart (GRCh37) and filtered based on location of standard

662 chromosomes (1-22, X, Y), resulting in 37,309 TSSs of 18,710 protein-coding genes. Open-
663 chromatin sites were collected from ATAC-seq profiles of a TCGA study [40] and DNAse-seq
664 profiles of the Roadmap Epigenomics project [41]. We used tissue-specific open-chromatin profiles
665 for every cancer type, first by matching cancer types of the PCAWG project and the TCGA
666 project (17/25 cohorts), and second, by selecting the closest normal tissues or cell lines of the
667 Epigenomics Roadmap project for cancer types for which the cancer-specific open-chromatin
668 profiles were not available (8/25 cohorts) (**Supplementary Table 1A**). The pan-cancer analysis
669 considered pan-cancer open-chromatin sites of the TCGA ATAC-seq study. To better compare
670 the mutation rates of TSSs and open-chromatin sites, we excluded the open-chromatin sites that
671 overlapped with the TSSs or CTCF binding sites. Sites that were originally mapped to the
672 GRCh38 reference genome were re-mapped to GRCh37 using the LiftOver method. Throughout
673 the study, the three classes of sites were normalised to uniform width based on median
674 coordinates. CTCF binding sites were defined using 50 bps (±25 bps) windows around the
675 midpoint of sites. Midpoints of TSS loci were defined in the Ensembl database and we used a
676 200 bps (±100 bps) window around the TSSs. Open-chromatin sites were also defined using a
677 200 bps (±100 bps) window around site midpoints. We systematically explored various values of
678 the site width parameters and the final selection was based on the strength of signal and
679 consistency (**Supplementary Figure 2A**).

680 **Defining target genes of TSSs and open-chromatin sites.** TSS and open-chromatin sites were
681 analysed in groups based on the mRNA abundance of target genes in matching tumors. TSS
682 target genes were retrieved from the Ensembl database and target genes of open-chromatin sites
683 were predicted using two custom strategies. First, direct target genes of open-chromatin sites
684 were defined based on their location in gene promoters, 5' and 3' UTRs, and exons (open-
685 chromatin sites directly overlapping TSSs were excluded, as specified above). Second, a
686 promoter-capture Hi-C dataset from a recent multi-tissue study [46] was used to identify long-range
687 chromatin interactions of open-chromatin sites and target genes. We selected high-confidence
688 interactions with frequency above 10 that were observed in at least 5/28 cell types to define
689 putatively constitutive interactions of open-chromatin sites and target genes.

690 **Grouping TSSs and open-chromatin sites by tissue-specific mRNA abundance.** We used
691 mRNA abundance in matching tumors as a proxy of site functionality of TSSs and open-

692 chromatin sites. To group sites by mRNA abundance, we used the uniformly processed PCAWG
693 RNA-seq dataset [45] (RPKM-UQ) that covered ~50% of the whole cancer genome cohort. We
694 limited the mRNA analysis to the tumors in the WGS dataset and excluded non-coding genes.
695 Cancer type specific analyses were carried out in 20 cohorts of cancer types for which at least 15
696 tumor samples with mRNA and WGS data were available, and also the pooled pan-cancer group
697 as the 21st cohort. We discarded genes with duplicated HGNC symbols and genes for which TSS
698 or open-chromatin sites were not mapped. Together, this resulted in mRNA measurements for
699 20,042 protein-coding genes in 1,267 tumor transcriptomes. Next, for each cancer type, we
700 grouped all the genes into five bins based on their median mRNA abundance and analysed the
701 bins using RM2 separately for TSSs and open-chromatin sites. For the pan-cancer analysis, we
702 binned genes using median mRNA abundance in the entire RNA-seq dataset.

703 **Grouping CTCF binding sites by the conservation of CTCF binding across human cell**
704 **lines.** To analyse CTCF binding sites by their tissue and cell type specificity, we grouped all
705 162,209 CTCF binding sites of the ENCODE dataset into five equally sized bins based on the
706 number of cell lines where a binding event of CTCF to the sites was detected. To interpret these
707 CTCF binding sites, we used information on the three-dimensional genome and CTCF sequence
708 motifs. First, we retrieved chromatin loops in eight cell lines from a Hi-C study [48], used a ±1,000
709 bps window around loop anchor midpoints to define narrower versions of loop anchors, and
710 counted the number of CTCF binding sites in each bin overlapping these loop anchors. We used
711 a Fisher's exact test to evaluate the enrichment of CTCF binding sites at loop anchors among the
712 CTCF binding sites with constitutive activity across cell types (i.e., the 5th bin of CTCF sites).
713 Second, using the motif MA0139.1 of the JASPAR database [66], we detected the presence of the
714 core CTCF motif in CTCF binding sites using the FIMO method [67] and selected significant motif
715 instances ($FDR < 0.01$). Enrichment of CTCF motifs in the 5th bin of sites was identified
716 similarly.

717 **Analysis of localised mutation frequencies in gene-regulatory and chromatin architectural**
718 **elements.** First, we evaluated the localised mutation rates in CTCF binding sites, TSSs and
719 cancer-specific open-chromatin sites for the pan-cancer cohort and all cohorts of selected cancer
720 types. Total mutations and mutations grouped by COSMIC signatures and reference/alternative
721 allele were analysed. Indel mutations were analysed as part of total mutations and also as a

722 separate group. Results were adjusted for multiple testing using the Benjamini-Hochberg false
723 discovery rate procedure and filtered (*FDR* < 0.05). We also analysed the simulated variant call
724 set using the same pipeline and found no significant results, as expected (*FDR* < 0.05). Results of
725 the systematic analysis were visualised as a dot plot. *FDR* values in the main dot plot were
726 capped at 10^{-32} for visualisation purposes. To visualise localised mutation rates, all sites were
727 pooled, aligned using median coordinates and trimmed to uniform lengths. Coordinates were
728 transformed relative to site midpoint. Upstream and downstream flanking sequences of equal
729 length were also considered. Local regression (LOESS) curves with the span parameter of 33%
730 were used to visualise a smoothened mutation frequency in sites relative to flanking sequences.

731 **Associating mutations in TSSs and open-chromatin sites with transcript abundance.** To
732 study the associations of mRNA abundance and mutations in TSSs and open-chromatin sites, we
733 grouped the sites by mRNA abundance of genes in matching tumor types, and then performed
734 the RM2 analysis separately for each bin and cancer type. Since RM2 is better powered to
735 discover localised mutation rate differences in larger groups of sites, we used down-sampling of
736 TSSs and open-chromatin sites in the mRNA-defined gene bins to provide better comparisons. In
737 each bin and cancer type, we randomly sampled 4,500 sites for RM2 analysis and repeated this
738 procedure over 100 iterations and selected the RM2 result with the median P-value for reporting.
739 The number for down-sampling was chosen as the closest smallest common value for bin sizes
740 for TSSs and open-chromatin sites. The results were adjusted for multiple testing using the
741 Benjamini-Hochberg false discovery rate procedure and filtered (*FDR* < 0.05).

742 **Identifying pathways with frequent mutations in transcription start sites.** We asked whether
743 the localised changes in mutation frequencies of TSSs affected specific biological processes and
744 pathways. We repurposed the RM2 model to analyse TSSs of gene sets corresponding to
745 biological processes of Gene Ontology [68] and molecular pathways of the Reactome database [69].
746 Gene sets were derived from the g:Profiler [70] web server (March 3rd, 2020) and subsequently
747 filtered to include 1,871 gene sets with 100 to 1,000 genes while smaller and larger gene sets
748 were excluded. Results were corrected for multiple testing using the Benjamini-Hochberg FDR
749 procedure separately for every cancer type and filtered for statistical significance (*FDR* < 0.05).
750 The pathways with significantly higher TSS-specific mutation frequencies were visualised as an
751 enrichment map [71] in Cytoscape and major biological themes were manually curated as described

752 earlier [47]. Nodes in the enrichment map were painted to reflect cancer types where these pathway
753 enrichments were detected according to the custom color scheme of the PCAWG project.

754 **Associating elevated mutation frequencies in CTCF binding sites with constitutive CTCF**
755 **binding and core sequence motifs.** To study the functional associations of localised mutation
756 frequencies at CTCF binding sites, we first analysed five groups of CTCF binding sites defined
757 based on the counts of cell lines where the sites were observed. These bins, ranging from sites of
758 single cell lines to sites constitutively active in all or most cell lines, were analysed in all cohorts
759 of individual cancer types and the pan-cancer cohort. Findings were corrected for multiple
760 testing and filtered to select significant findings (*FDR* < 0.05). To perform the motif-based
761 analysis, we first split sites into two groups based a significant match to the core CTCF motif
762 using FIMO [67] (FDR < 0.01). Then, we repeated the binning of sites by shared activity, as
763 described above, to ensure all bins of sites within each group (with or without motif) were of
764 equal size. The ten bins of CTCF sites were also analysed using RM2. Lastly, to perform a
765 tissue-specific analysis of matching cancer types and cell lines, we selected CTCF binding sites
766 for two cancer cell lines in the ENCODE dataset: MCF-7, a breast cancer cell line, and HepG2, a
767 liver cancer cell line. The CTCF binding sites found in at least these two cell lines were also split
768 into ten bins based on conservation of CTCF binding and presence of the core sequence motif,
769 and then analysed with RM2 using mutations of the PCAWG dataset for the Breast-AdenoCa
770 cohort (with MCF-7) and Liver-HCC cohort (with HepG2).

771 **Associating elevated mutation frequencies in sites with driver mutations and recurrent**
772 **copy-number alterations.** We tested whether the localised changes in mutation frequencies in
773 CTCF binding sites, TSSs and open-chromatin sites were associated with driver mutations (*i.e.*,
774 SNVs, indels) or recurrent copy-number alterations (CNAs). First we collected a high-
775 confidence set of driver mutations and CNAs in the PCAWG cohort. Driver mutations in exons
776 of protein-coding genes were predicted for each selected cancer type using the
777 ActiveDriverWGS method [34]. For driver analysis, we used the PCAWG variant calls after
778 filtering tumors as described above, corrected the results for multiple testing using the
779 Benjamini-Hochberg FDR procedure and selected significant driver genes (*FDR* < 0.05). FDR
780 correction was conducted separately for each cancer type across the pooled set of protein-coding
781 and non-coding genes. Tumors with and without SNVs or indels in predicted driver genes were

782 used for localised mutation rate analysis. Predictions of recurrent CNAs were derived from the
783 pan-cancer dataset of GISTIC2 calls of the PCAWG project [53]. All lesions at 95% confidence
784 scores were considered and amplifications and deletions were analysed separately. High-
785 confidence CNA events were used (GISTIC2 score = 2). Tumors with and without CNAs in the
786 recurrently altered regions, as defined by GISTIC2, were used for localised mutation rate
787 analysis. Two additional chromosomal features were used. First, we grouped the tumors by
788 presence or absence of Whole Genome Duplication (WGD) as determined in the PCAWG
789 project [53]. Second, we computed the Percent Genome Altered (PGA) metric for every tumor as a
790 proxy of aneuploidy. PGA was computed as the percentage of the autosomal genome affected by
791 CNA segments whose total copy number deviated from the global reference (two copies for non-
792 WGD genomes; four copies for WGD genomes). PGA metrics were median-dichotomised
793 separately for every cancer type, resulting in two subgroups of tumors: the PGA-high group and
794 the PGA-low group corresponding to above-median and below median CNA burden of tumors.
795 Following the construction of these genetic features (*i.e.*, driver gene mutations, recurrent CNAs,
796 WGD, PGA-high), we filtered overly frequent (more than 2/3 of the cohort) and infrequent
797 features (less than 15 tumors) to improve the power of the RM2 analysis. For each cancer type,
798 the relevant features were then analysed for localised elevations in mutation frequencies in the
799 three categories of genomic elements (open-chromatin sites, CTCF binding sites, TSSs). The
800 binary co-factor in RM2 was used to indicate the presence or absence of the given feature in the
801 given tumor genome. We first computed the significance of site-specific localised mutation rates
802 given the presence or absence of the feature. To validate this joint model, subgroups of tumors
803 with and without the defining feature were also analysed separately using RM2. All combined
804 RM2 results concerning individual features, cancer types and genomic sites were then adjusted
805 for multiple testing correction using the Benjamini-Hochberg FDR procedure and significant
806 results were selected ($FDR < 0.05$). We filtered the results to only include positive and
807 significant interactions of genetic features and elevated localised mutation frequencies
808 (interaction $P < 0.05$, main and interaction coefficients > 0) and compared the FDR values, fold-
809 change values and visualisations of localised mutation frequencies to validate these findings.

810 **Associating mutagenesis-related copy number amplifications with mRNA abundance**. To
811 evaluate the functional role of recurrent copy number amplifications associated with increased

812 mutation frequencies in the gene-regulatory and chromatin architectural sites, we studied the
813 genes located in the amplified regions and retrieved from the PCAWG GISTIC2 dataset. We
814 compared the mRNA abundance levels of the genes between groups of tumors defined by the
815 presence or absence of the amplifications, using matching RNA-seq data available in PCAWG
816 [45]. Genes with low mRNA abundance were removed from the analysis (median FPKM-UQ < 1).
817 mRNA abundance levels of genes in amplified and non-amplified (*i.e.*, balanced and deleted)
818 tumors were compared using one-sided non-parametric Wilcoxon tests, assuming that an
819 increase in mRNA abundance would match the underlying copy number amplification. Results
820 were adjusted for multiple testing correction using the Benjamini-Hochberg FDR procedure and
821 significant results were selected ($FDR < 0.05$). We ranked the resulting genes based on the total
822 significance of CNA/mRNA associations across all cancer types and site types by summing the
823 negative log10-transformed significant FDR-values of each gene. Known cancer genes of the
824 COSMIC Cancer Gene Census database [72] (v91, downloaded 14.05.2020) were highlighted.

825 **Method benchmarking and power analysis.** We evaluated the performance of our method and
826 statistical power using simulated variant calls, different parametrizations and down-sampling of
827 input datasets. First, to evaluate method calibration and false-positive rates, we performed a
828 systematic analysis of open-chromatin sites, TSSs, and CTCF binding sites in a comparable set
829 of simulated variant calls from PCAWG. This simulated variant set was derived earlier from the
830 same set of tumor genomes using trinucleotide-informed shuffling of mutations [4]. Simulated
831 variant calls were analysed similarly to true variant calls for total mutation counts, reference and
832 alternative nucleotide combinations and predicted mutational signatures. Results from RM2 were
833 adjusted for multiple testing using the Benjamini-Hochberg FDR procedure separately for results
834 derived from true and simulated variant calls. As expected, the simulated variant calls revealed
835 no statistically significant results of localised mutation rates in any cancer type, site type or
836 mutation subset ($FDR < 0.05$). We then visualised the distribution of log-transformed *P*-values
837 derived from true and simulated variant calls using quantile-quantile (QQ) plots and found that
838 highly significant *P*-values were detected in true datasets while the *P*-values derived from
839 simulated variant calls were uniformly distributed as expected. These analyses show that our
840 model is well-calibrated and is not subject to inflated false-positive findings. Second, to evaluate
841 the statistical power of RM2, we performed a systematic down-sampling analysis by randomly

842 selecting subsets of sites and tumors for localised mutation rate analysis. We focused on the
843 PCAWG liver hepatocellular carcinoma (Liver-HCC) cohort of 300 samples and CTCF sites. A
844 series of down-sampling configurations were used (2000, 5000, …, 100,000 sites sampled; 25,
845 50, …, 250, 300 genomes sampled), as well as the full datasets of sites and genomes. Each
846 configuration was tested 100 times with different random subsets of data. As a power analysis,
847 we evaluated the fraction of runs that revealed a significant enrichment of somatic mutations at
848 CTCF sites ($P < 0.05$) and the median P-value of these 100 runs. Third, we evaluated the
849 parameter values of RM2 that determine the genomic width of each site and the control regions
850 of upstream and downstream flanking sequences. As expected, site-specific mutation rates were
851 consistently identified at multiple values of the width parameter for each class of site (open-
852 chromatin sites, CTCF binding sites, and TSSs), indicating robustness of our analysis to different
853 parameter values. However, different site classes showed preferences towards shorter sites
854 (CTCF binding sites: 20-100 bps) or longer sites (open-chromatin sites and TSS: 200-800 bps),
855 likely due to different underlying mutational processes. For the entire study, the optimal genomic
856 size of every site class was selected based on the strongest effect size and significance across
857 multiple cancer types. The value of 50 bps (±25 bps) was selected for CTCF sites. For open-
858 chromatin sites and TSSs, we selected the common site width of 200 bps (±100 bps) that showed
859 strong effects in both TSSs and open-chromatin sites, to increase comparability of the two
860 classes. We evaluated the effect of grouping sites by their megabase-scale mutation frequencies
861 (*i.e.*, the MbpRate covariate of RM2). We repeated the systematic analysis using a simplified
862 RM2 model that excluded the MbpRate covariate and compared the resulting P-values and FDR
863 values derived from the original and simplified models. The original model (including MbpRate)
864 outperformed the simplified model (excluding MbpRate) by assigning stronger significance to
865 findings in the dataset of true mutations and reduced significance to findings in the simulated
866 dataset, indicating that accounting for megabase-scale mutation frequency as covariate improves
867 sensitivity and specificity of the model.

868 **Acknowledgments.** This work was supported by a Project Grant from the Canadian Institutes of
869 Health Research (CIHR) to J.R., an Operating Grant from the Cancer Research Society to J.R.,
870 and the Investigator Award to J.R. from the Ontario Institute for Cancer Research (OICR).
871 C.A.L. was partially supported by a Graduate Student Fellowship from the Department of

Medical Biophysics, University of Toronto. Funding to OICR is provided by the Government of Ontario.

Author contributions. J.R., C.A.L. and D.A.R. designed and implemented the method. J.R. and C.A.L. analysed the data. J.R. wrote the manuscript with significant input from C.A.L. and D.A.R. J.R. conceived and supervised the project. All authors reviewed and edited the manuscript and approved the final version.

880 **References**

881 1 Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719-
882 724, doi:10.1038/nature07943 (2009).
883 2 Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558,
884 doi:10.1126/science.1235122 (2013).
885 3 ICGC-TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis
886 of whole genomes. *Nature* **578**, 82-93, doi:10.1038/s41586-020-1969-6 (2020).
887 4 Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,693 cancer whole
888 genomes. *Nature* **578**, 102–111 (2020).
889 5 Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and
890 Mutations. *Cell* **174**, 1034-1035, doi:10.1016/j.cell.2018.07.034 (2018).
891 6 Kumar, S. *et al.* Passenger Mutations in More Than 2,500 Cancer Genomes: Overall
892 Molecular Functional Impact and Consequences. *Cell* **180**, 915-927 e916,
893 doi:10.1016/j.cell.2020.01.032 (2020).
894 7 Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **in press** (2020).
895 8 Supek, F. & Lehner, B. Scales and mechanisms of somatic mutation rate variation across
896 the human genome. *DNA Repair (Amst)*, 102647, doi:10.1016/j.dnarep.2019.102647
897 (2019).
898 9 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-
899 associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).
900 10 Reijns, M. A. M. *et al.* Lagging-strand replication shapes the mutational landscape of the
901 genome. *Nature* **518**, 502-506, doi:10.1038/nature14183 (2015).
902 11 Schuster-Bockler, B. & Lehner, B. Chromatin organization is a major influence on
903 regional mutation rates in human cancer cells. *Nature* **488**, 504-507,
904 doi:10.1038/nature11273 (2012).
905 12 Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of
906 cancer. *Nature* **518**, 360-364, doi:10.1038/nature14221 (2015).
907 13 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**,
908 415-421, doi:10.1038/nature12477 (2013).
909 14 Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature*
910 **578**, 94-101, doi:10.1038/s41586-020-1943-3 (2020).
911 15 Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents.
912 *Cell* **177**, 821-836 e816, doi:10.1016/j.cell.2019.03.001 (2019).
913 16 Pich, O. *et al.* The mutational footprints of cancer therapies. *Nature genetics* **51**, 1732-
914 1740, doi:10.1038/s41588-019-0525-5 (2019).
915 17 Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature*
916 **575**, 210-216, doi:10.1038/s41586-019-1689-y (2019).
917 18 Kübler, K. *et al.* Tumor mutational landscape is a record of the pre-malignant state.
918 *bioRxiv* **517565**, doi:https://doi.org/10.1101/517565 (2019).
919 19 Jiao, W. *et al.* A deep learning system accurately classifies primary and metastatic
920 cancers using passenger mutation patterns. *Nat Commun* **11**, 728, doi:10.1038/s41467-
921 019-13825-8 (2020).
922 20 Gonzalez-Perez, A., Sabarinathan, R. & Lopez-Bigas, N. Local Determinants of the
923 Mutational Landscape of the Human Genome. *Cell* **177**, 101-114,
924 doi:10.1016/j.cell.2019.02.051 (2019).

925 21 Kaiser, V. B., Taylor, M. S. & Semple, C. A. Mutational Biases Drive Elevated Rates of
926 Substitution at Regulatory Sites across Cancer Types. *PLoS Genet* **12**, e1006207,
927 doi:10.1371/journal.pgen.1006207 (2016).
928 22 Yazdi, P. G. *et al.* Increasing Nucleosome Occupancy Is Correlated with an Increasing
929 Mutation Rate so Long as DNA Repair Machinery Is Intact. *PLoS One* **10**, e0136574,
930 doi:10.1371/journal.pone.0136574 (2015).
931 23 Hara, R., Mo, J. & Sancar, A. DNA damage in the nucleosome core is refractory to repair
932 by human excision nuclease. *Mol Cell Biol* **20**, 9173-9181, doi:10.1128/mcb.20.24.9173-
933 9181.2000 (2000).
934 24 Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & Lopez-Bigas, N.
935 Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature*
936 **532**, 264-267, doi:10.1038/nature17661 (2016).
937 25 Perera, D. *et al.* Differential DNA repair underlies mutation hotspots at active promoters
938 in cancer genomes. *Nature* **532**, 259-263, doi:10.1038/nature17437 (2016).
939 26 Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome
940 neighborhoods. *Science* **351**, 1454-1458, doi:10.1126/science.aad9024 (2016).
941 27 Katainen, R. *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nature*
942 *genetics* **47**, 818-821, doi:10.1038/ng.3335 (2015).
943 28 Guo, Y. A. *et al.* Mutation hotspots at CTCF binding sites coupled to chromosomal
944 instability in gastrointestinal cancers. *Nat Commun* **9**, 1520, doi:10.1038/s41467-018-
945 03828-2 (2018).
946 29 Polak, P. *et al.* Reduced local mutation density in regulatory DNA of cancer genomes is
947 linked to DNA repair. *Nature biotechnology* **32**, 71-75, doi:10.1038/nbt.2778 (2014).
948 30 Frigola, J. *et al.* Reduced mutation rate in exons due to differential mismatch repair.
949 *Nature genetics* **49**, 1684-1692, doi:10.1038/ng.3991 (2017).
950 31 Bell, R. J. *et al.* The transcription factor GABP selectively binds and activates the mutant
951 TERT promoter in cancer. *Science* **348**, 1036-1039, doi:10.1126/science.aab0015 (2015).
952 32 Huang, F. W. *et al.* Highly recurrent TERT promoter mutations in human melanoma.
953 *Science* **339**, 957-959, doi:10.1126/science.1229259 (2013).
954 33 Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding
955 somatic mutations and gene expression alterations across 14 tumor types. *Nature genetics*
956 **46**, 1258-1263, doi:10.1038/ng.3141 (2014).
957 34 Zhu, H. *et al.* Candidate Cancer Driver Mutations in Distal Regulatory Elements and
958 Long-Range Chromatin Interaction Networks. *Mol Cell*,
959 doi:10.1016/j.molcel.2019.12.027 (2020).
960 35 Corona, R. I. *et al.* Non-coding somatic mutations converge on the PAX8 pathway in
961 ovarian cancer. *Nat Commun* **11**, 2020, doi:10.1038/s41467-020-15951-0 (2020).
962 36 Liu, E. M. *et al.* Identification of Cancer Drivers at CTCF Insulators in 1,962 Whole
963 Genomes. *Cell Syst* **8**, 446-455 e448, doi:10.1016/j.cels.2019.04.001 (2019).
964 37 Reyna, M. A. *et al.* Pathway and network analysis of more than 2,500 whole cancer
965 genomes. *Nature Communications* **11**, 729 (2020).
966 38 Mazrooei, P. *et al.* Cistrome Partitioning Reveals Convergence of Somatic Mutations and
967 Risk Variants on Master Transcription Regulators in Primary Prostate Tumors. *Cancer*
968 *Cell* **36**, 674-689 e676, doi:10.1016/j.ccell.2019.10.005 (2019).
969 39 Encode Project Consortium. An integrated encyclopedia of DNA elements in the human
970 genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

971 40 Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers.
972 *Science* **362**, doi:10.1126/science.aav1898 (2018).
973 41 Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human
974 epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).
975 42 Dvorak, K. *et al.* Bile acids in combination with low pH induce oxidative stress and
976 oxidative DNA damage: relevance to the pathogenesis of Barrett's oesophagus. *Gut* **56**,
977 763-771, doi:10.1136/gut.2006.103697 (2007).
978 43 Tomkova, M., Tomek, J., Kriaucionis, S. & Schuster-Bockler, B. Mutational signature
979 distribution varies with DNA replication timing and strand asymmetry. *Genome Biol* **19**,
980 129, doi:10.1186/s13059-018-1509-y (2018).
981 44 Dellino, G. I. *et al.* Release of paused RNA polymerase II at specific loci favors DNA
982 double-strand-break formation and promotes cancer translocations. *Nature genetics* **51**,
983 1011-1023, doi:10.1038/s41588-019-0421-z (2019).
984 45 PCAWG Transcriptome Core Group *et al.* Genomic basis of RNA alterations in cancer.
985 *Nature* **in press** (2020).
986 46 Jung, I. *et al.* A compendium of promoter-centered long-range chromatin interactions in
987 the human genome. *Nature genetics* **51**, 1442-1449, doi:10.1038/s41588-019-0494-8
988 (2019).
989 47 Reimand, J. *et al.* Pathway enrichment analysis and visualization of omics data using
990 g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protoc* **14**, 482-517,
991 doi:10.1038/s41596-018-0103-9 (2019).
992 48 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles
993 of chromatin looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).
994 49 Wendt, K. S. *et al.* Cohesin mediates transcriptional insulation by CCCTC-binding factor.
995 *Nature* **451**, 796-801, doi:10.1038/nature06634 (2008).
996 50 Kaiser, V. B. & Semple, C. A. Chromatin loop anchors are associated with genome
997 instability in cancer and recombination hotspots in the germline. *Genome Biol* **19**, 101,
998 doi:10.1186/s13059-018-1483-4 (2018).
999 51 Pinoli, P., Stamoulakatou, E., Nguyen, A. P., Rodriguez Martinez, M. & Ceri, S. Pan-
1000 cancer analysis of somatic mutations and epigenetic alterations in insulated
1001 neighbourhood boundaries. *PLoS One* **15**, e0227180, doi:10.1371/journal.pone.0227180
1002 (2020).
1003 52 Khoury, A. *et al.* Constitutively bound CTCF sites maintain 3D chromatin architecture
1004 and long-range epigenetically regulated domains. *Nat Commun* **11**, 54,
1005 doi:10.1038/s41467-019-13753-7 (2020).
1006 53 Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **in**
1007 **press** (2020).
1008 54 Wu, R. C., Wang, T. L. & Shih Ie, M. The emerging roles of ARID1A in tumor
1009 suppression. *Cancer Biol Ther* **15**, 655-664, doi:10.4161/cbt.28411 (2014).
1010 55 Dykhuizen, E. C. *et al.* BAF complexes facilitate decatenation of DNA by topoisomerase
1011 IIalpha. *Nature* **497**, 624-627, doi:10.1038/nature12146 (2013).
1012 56 Park, J. H. *et al.* Mammalian SWI/SNF complexes facilitate DNA double-strand break
1013 repair by promoting gamma-H2AX induction. *EMBO J* **25**, 3986-3997,
1014 doi:10.1038/sj.emboj.7601291 (2006).

1015 57 Zhao, Q. *et al.* Modulation of nucleotide excision repair by mammalian SWI/SNF
1016 chromatin-remodeling complex. *The Journal of biological chemistry* **284**, 30424-30432,
1017 doi:10.1074/jbc.M109.044982 (2009).
1018 58 Trizzino, M. *et al.* The Tumor Suppressor ARID1A Controls Global Transcription via
1019 Pausing of RNA Polymerase II. *Cell Rep* **23**, 3933-3945,
1020 doi:10.1016/j.celrep.2018.05.097 (2018).
1021 59 Cancer Genome Atlas Research Network. Electronic address, a. a. d. h. e. & Cancer
1022 Genome Atlas Research, N. Integrated Genomic Characterization of Pancreatic Ductal
1023 Adenocarcinoma. *Cancer Cell* **32**, 185-203 e113, doi:10.1016/j.ccell.2017.07.007 (2017).
1024 60 The Cancer Genome Atlas Network. Genomic Classification of Cutaneous Melanoma.
1025 *Cell* **161**, 1681-1696, doi:10.1016/j.cell.2015.05.044 (2015).
1026 61 Chapman, P. B. *et al.* Improved survival with vemurafenib in melanoma with BRAF
1027 V600E mutation. *N Engl J Med* **364**, 2507-2516, doi:10.1056/NEJMoa1103782 (2011).
1028 62 Sheu, J. J. *et al.* Mutant BRAF induces DNA strand breaks, activates DNA damage
1029 response pathway, and up-regulates glucose transporter-1 in nontransformed epithelial
1030 cells. *Am J Pathol* **180**, 1179-1188, doi:10.1016/j.ajpath.2011.11.026 (2012).
1031 63 Fredriksson, N. J. *et al.* Recurrent promoter mutations in melanoma are defined by an
1032 extended context-specific mutational signature. *PLoS Genet* **13**, e1006773,
1033 doi:10.1371/journal.pgen.1006773 (2017).
1034 64 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in
1035 141,456 humans. *Nature* **581**, 434-443, doi:10.1038/s41586-020-2308-7 (2020).
1036 65 Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of
1037 somatic mutations in normal human skin. *Science* **348**, 880-886,
1038 doi:10.1126/science.aaa6806 (2015).
1039 66 Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription
1040 factor binding profiles. *Nucleic Acids Res* **48**, D87-D92, doi:10.1093/nar/gkz1001 (2020).
1041 67 Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given
1042 motif. *Bioinformatics* **27**, 1017-1018, doi:10.1093/bioinformatics/btr064 (2011).
1043 68 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene
1044 Ontology Consortium. *Nature genetics* **25**, 25-29, doi:10.1038/75556 (2000).
1045 69 Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res* **46**, D649-
1046 D655, doi:10.1093/nar/gkx1132 (2018).
1047 70 Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler--a web-based toolset
1048 for functional profiling of gene lists from large-scale experiments. *Nucleic acids research*
1049 **35**, W193-200, doi:10.1093/nar/gkm226 (2007).
1050 71 Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a
1051 network-based method for gene-set enrichment visualization and interpretation. *PLoS*
1052 *One* **5**, e13984, doi:10.1371/journal.pone.0013984 (2010).
1053 72 Futreal, P. A. *et al.* A census of human cancer genes. *Nat Rev Cancer* **4**, 177-183,
1054 doi:10.1038/nrc1299 (2004).

1055