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

Localised variation of somatic mutation rates affects diverse functional sequence elements in cancer genomes through poorly understood mutational processes. Here, we characterise the mutational landscape of 640,000 gene regulatory and chromatin architectural elements in 2,421 whole cancer genomes using our new statistical model RM2. This method quantifies differential mutation rates and signatures in classes of genomic elements via genetic, trinucleotide and megabase-scale effects. We report a detailed map of localised mutational processes affecting CTCF binding sites, transcription start sites (TSS) and cancer-specific open-chromatin regions. This includes a pan-cancer indel depletion in open-chromatin sites, a TSS-specific mutational process correlated with mRNA abundance in core cellular and cancer-associated processes, a subset of hypermutated, constitutively active CTCF binding sites involved in chromatin architectural interactions, and an enrichment of signature SBS17b in CTCF sites in gastrointestinal cancers. We also detect genetic driver alterations potentially underlying localised mutation rates, including *RAD21* amplifications and *BRAF* mutations associating with mutagenesis of CTCF binding sites, and *SDHA* amplifications indicative of frequent lung cancer mutations in open-chromatin sites. Our framework and the catalogue of localised mutational processes provide novel perspectives to cancer genome evolution and its implications for oncogenesis, tumor heterogeneity and cancer driver gene discovery.

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 rates show complex genomic variation at multiple resolutions [8]. In megabase-scale genomic windows, variations in mutation rates are associated with transcriptional activity, chromatin state and DNA replication as late-replicating and untranscribed regions are often more mutated than regions of early replication and highly expressed genes [9-12]. At a single 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 rates 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 rates 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 carry relatively fewer mutations due to increased DNA repair activity. While most such non-coding mutations are likely functionally neutral passengers induced by localised mutagenesis, 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

58 hotspot in the *TERT* promoter creates a TFBS of the ETS TF family that leads to constitutive
59 activation of *TERT* and enables replicative immortality of cancer cells [31-33]. Recent studies have
60 catalogued candidate non-coding driver elements in gene regulatory and chromatin architectural
61 regions of the cancer genome with functional validations of novel elements [34-36] and highlighted
62 the convergence of non-coding mutations on molecular pathways and regulatory networks [37,38].
63 Thus we need to characterise localised mutational processes to understand the evolution of the
64 somatic genome and the effects of carcinogens and endogenous mutational processes, but also to
65 evaluate the effects of positive selection in the non-coding genome. However, few dedicated
66 computational methods exist to analyse mutation rates at the local resolution of the genome. As a
67 result, there is a lack of large-scale analyses of the local mutation landscape in pan-cancer WGS
68 datasets, leaving the genetic and environmental determinants poorly understood.

69 Here we developed a new statistical framework that quantifies the activity of mutational
70 processes and signatures on specific classes of non-coding elements of the cancer genome. Our
71 model considers local sequence context, megabase-level somatic mutation rates and genetic
72 covariates to control for variation at the trinucleotide and megabase resolution while isolating
73 site-level effects. We performed a systematic analysis of local mutation rate variation in three
74 classes of gene-regulatory and chromatin architectural genomic elements across 2,500 whole
75 cancer genomes of the ICGC/TCGA Pan-cancer Analysis of Whole Genomes (PCAWG) project
76 [3]. We found a pervasive mutation enrichment at these functional non-coding elements that was
77 characterised by specific mutational signatures and transcriptional and pathway-level activities of
78 these elements in select cancer types. We detected statistical interactions of local mutagenesis
79 and recurrent genomic alterations that suggest potential genetic mechanisms driving the
80 underlying mutational processes. Our computational framework and systematic analysis reveals
81 the diversity of mutational processes in functional non-coding elements of the cancer genome
82 and their roles in somatic genome evolution, drivers of cancer phenotypes and molecular
83 heterogeneity.

Results

A statistical framework for quantifying localised mutagenesis in cancer genomes

We implemented a statistical model, Regression Models for Regionalised Mutations (RM2), to quantify the local activity of mutational processes in whole cancer genomes in elements each spanning dozens to hundreds of nucleotides (**Supplementary Figure 1**). The model considers a genome-wide set of genomic elements such as TFBSs detected in thousands to hundreds of thousands of loci using chromatin immunoprecipitation with DNA sequencing (ChIP-seq) and similar techniques. The model uses negative binomial regression to evaluate whether the genomic elements of interest are collectively subject to a different mutation rate compared to control sequences upstream and downstream of these elements. Somatic single nucleotide variants (SNVs) and small insertions-deletions (indels) are analysed, however, the model can be extended to somatic structural variant breakpoints and germline variation. The model considers four types of information to evaluate local mutation rates: a) nucleotide sequence content of genomic elements and control sequences representing the potential space for mutagenesis, grouped by 96 trinucleotide signatures and one indel signature (*nPosits*), b) 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 rates (*triNucMut*), c) 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 d) 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 upstream and downstream control regions are pooled into a user-defined number of bins of equal size based on their megabase-scale mutation rates (ten bins by default). Elements and flanking control sequences are distinguished using the binary cofactor *isElement*. The full model is written as follows:

$$nMut \sim \texttt{NegBin(}\ \texttt{offset(}\ \texttt{log(}\ nPosits\ \texttt{)}\ \texttt{)} + triNucMut + \texttt{log1p(}\ MbpRate\ \texttt{)} + coFac + isElement\ \texttt{)}.$$

To determine whether the mutation rates of the genomic elements differ from the rates of the flanking sequences given trinucleotide-level and megabase-scale covariates, we evaluate the

112 significance of the cofactor *isElement* using a likelihood-ratio test. Significant and positive
113 coefficients of this cofactor indicate increased mutation rates in genomic elements relative to
114 flanking controls, while negative coefficients indicate a depletion of mutations. Similarly, we can
115 discover potential genetic or clinical interactions with localized activity of mutational processes.
116 Given a binary subgroup classification of tumors (*coFac*), we evaluate the significance of its
117 interaction with local mutation rates (*isElement:coFac*). Positive coefficients of the interaction
118 indicate that the mutation rates in a clinical or genetic tumor subgroup are elevated when
119 accounting for the overall differences of the subgroups. Lastly, we extend the mutation rate
120 analysis to subclasses of mutations by allowing only specific classes of mutations to be included
121 in the mutation counts (*nMut*), for example those of specific DNA strand, transcriptional
122 direction, or COSMIC mutational signatures. We evaluated the performance of our method using
123 simulated datasets, power analysis and parameter variations as described below (**Supplementary**
124 **Figure 2**).

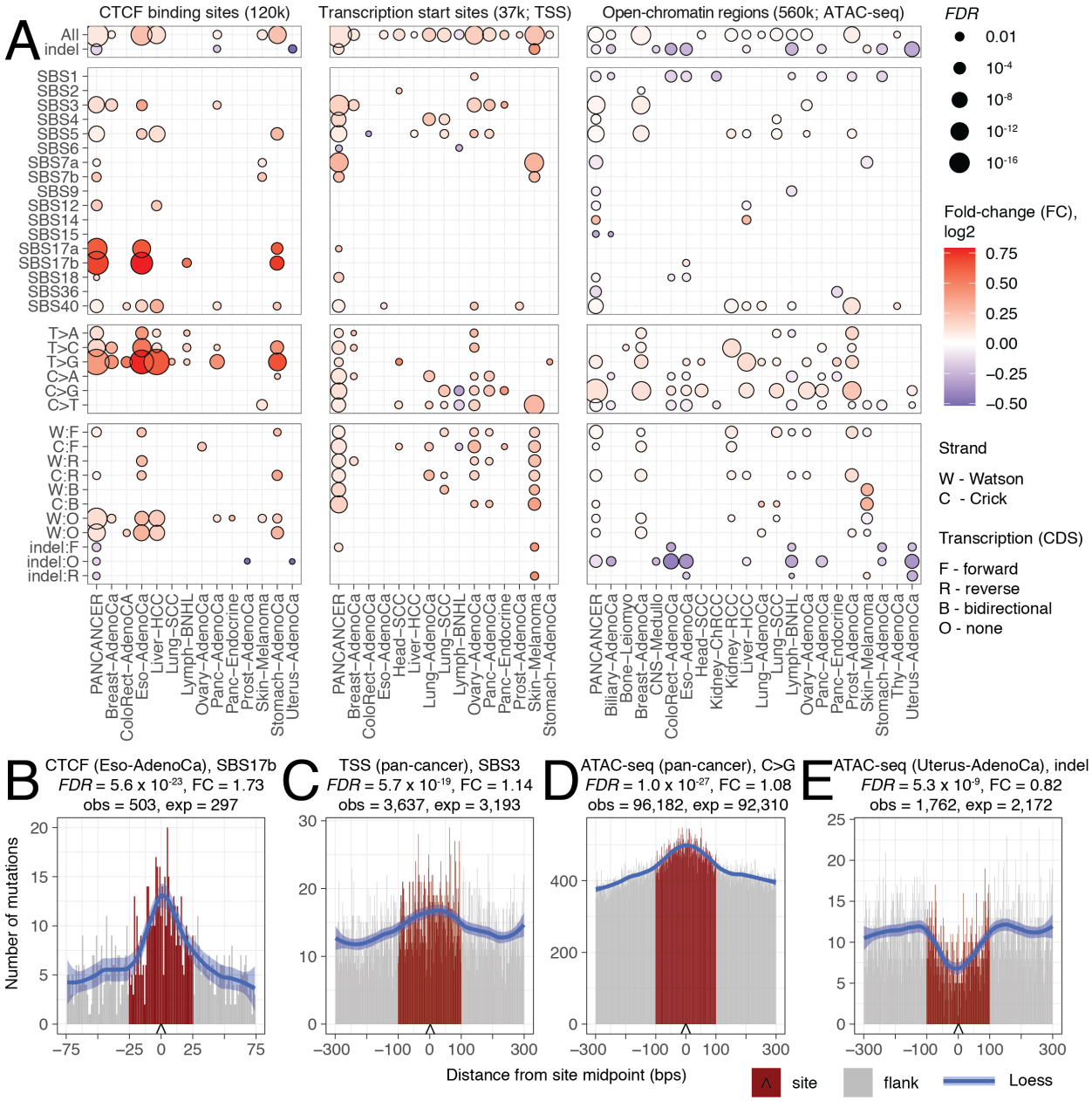

Figure 1. Localised mutation rates in gene-regulatory and chromatin architectural elements of cancer genomes. A. Comparison of mutation rates in DNA-binding sites of the CTCF chromatin architectural factor (left), transcription start sites (TSS) and cancer-specific open-chromatin regions (TCGA ATAC-seq) in 2,421 whole cancer genomes (*FDR* < 0.05). Total mutations (SNVs, indels) and mutations grouped by single base substitution (SBS) signatures, strand and transcriptional direction were analysed. **B-E.** Examples of localised mutation rates and signatures (left to right): enrichment SBS17b in CTCF binding sites in esophageal adenocarcinoma, pan-cancer enrichment of SBS3 mutations in TSSs, pan-cancer enrichment of C>G mutations in open-chromatin regions, and depletion of indels in open-chromatin regions in uterine adenocarcinoma.

127 **Localised mutation rates in gene-regulatory and chromatin architectural elements of**
128 **cancer genomes**

129 To study localised mutation rates in gene-regulatory and chromatin architectural elements of the
130 genome, we analysed the pan-cancer dataset of 2,514 whole cancer genomes of the PCAWG
131 project [3]. We individually analysed 25/35 cancer types with at least 25 samples as well as the
132 pan-cancer set containing all tumors of the 35 types (**Supplementary Figure 3A**). 69
133 hypermutated tumors were excluded to avoid confounding effects. Three classes of genomic
134 elements were analysed: 119,464 CTCF binding sites conserved in at least two cell lines in the
135 ENCODE project [39], 37,309 transcription start sites (TSS) of protein-coding genes from the
136 Ensembl database (GRCh37), and a pan-cancer set of 561,057 open-chromatin sites detected
137 across 410 primary tumors of The Cancer Genome Atlas (TCGA) project (*i.e.*, ATAC-seq sites)
138 [40] (**Supplementary Table 1A**). In total, the analysis included 640,023 unique loci representing
139 3.9% (120.5 Mbps) of the human genome. In addition to total mutations, we grouped mutations
140 by DNA strand (Watson or Crick), transcription status (forward, reverse, bidirectional or absent),
141 reference and alternative nucleotide pairs, and COSMIC mutational signatures of single base
142 substitutions (SBS) inferred in the PCAWG study [14]. Indel mutations were pooled with SNVs
143 and also analysed separately. To obtain a conservative analysis, we excluded a small fraction of
144 tumors as outliers (31 or 1.3%) where even single-sample RM2 analysis revealed highly
145 significant differences in local mutation rates in any of the three classes of sites ($FDR < 0.001$)
146 (**Supplementary Figure 3B**). The final pan-cancer analysis studied 23.0 million mutations
147 including 1.61 million indels detected in 2,421 genomes of 35 cancer types.

148 We first focused on the mutational profiles of CTCF DNA-binding sites. Overall mutation rates
149 in CTCF binding sites were significantly higher in liver hepatocellular carcinoma (Liver-HCC)
150 (RM2 $FDR = 4.3 \times 10^{-14}$, fold-change (FC) = 1.10), esophageal adenocarcinoma ($FDR = 1.1 \times$
151 10^{-20}, FC = 1.19) and stomach adenocarcinoma ($FDR = 8.3 \times 10^{-14}$, FC = 1.19). The pan-cancer
152 cohort also showed a significant enrichment, likely due to pooled effects of certain cancer types
153 ($FDR = 8.1 \times 10^{-27}$, FC = 1.07). These initial results confirm earlier reports of elevated mutation
154 rates in CTCF DNA-binding sites [21,27,28] and validate our computational model. Smaller
155 increases in mutation rates were also detected in melanoma, pancreatic and breast cancer ($FDR \leq$
156 0.02). Additional signals were observed in specific subgroups of mutations. Grouping the

157 mutations by reference and alternative nucleotides revealed a strong enrichment of T>G
158 mutations (*e.g.*, Liver-HCC, *FDR* = 1.4 x 10^{-33}, FC = 1.56), T>C and T>A mutations.
159 Interestingly, intergenic CTCF binding sites were particularly enriched in mutations in several
160 cohorts. We then asked whether CTCF binding sites were characterised by specific COSMIC
161 SBS mutational signatures. Esophageal and stomach cancers showed a strong enrichment of
162 SBS17 mutations (SBS17b: *FDR* = 5.6 x 10^{-23}, FC = 1.73; and *FDR* = 5.2 x 10^{-8}, FC = 1.66)
163 (**Figure 1B**), while this was not observed in Liver-HCC and other cancer types with frequent
164 CTCF binding site mutations. The etiology of SBS17b is unknown, however it has been linked to
165 acid reflux and oxidative damage to DNA in gastro-esophageal cancers [41,42], and a similar
166 mutational signature found in metastatic tumors has been associated with the effects of
167 nucleoside metabolic inhibitor chemotherapies capecitabine and 5-FU [16]. Our analysis suggests
168 that effects of these mutagens may be especially active at insulator and chromatin architectural
169 elements bound by CTCF in tissues of the digestive system. This analysis refines the annotation
170 of a mutational process acting on the DNA-binding sites of CTCF in a large dataset of whole
171 cancer genomes.

172 Transcription start sites (TSS) of protein-coding genes were significantly enriched in mutations
173 in the pan-cancer cohort (*FDR* = 1.2 x 10^{-35}, FC = 1.07) and in cohorts of 13/25 cancer types,
174 most prominently in melanoma (*FDR* = 1.6 x 10^{-17}, FC = 1.18), breast, head, lung, ovary and
175 pancreatic cancers (*FDR* ≤ 10^{-4}). Stronger enrichments were detected among C>G and C>T
176 mutations. Mutational signature analysis highlighted an elevated rate of the aging-associated
177 signature SBS5 in the pan-cancer cohort (*FDR* = 1.5 x 10^{-11}, FC = 1.06) and cohorts of four
178 cancer types. The signature SBS3, associated with defects of homologous recombination-based
179 DNA damage repair, was observed in the pan-cancer cohort (*FDR* = 5.7 x 10^{-19}, FC = 1.14)
180 (**Figure 1C**) as well as breast and ovarian cancers (*FDR* ≤ 10^{-3}, FC ≥ 1.12). Spontaneous
181 formation of endogenous double strand breaks at promoters has been associated with the pause
182 and release of RNA polymerase II and linked to chromosomal translocations in cancer [43],
183 suggesting a mechanism of TSS-specific mutagenesis in cancer genomes. Further mutational
184 signature enrichments were identified in specific cancer types, such as the ultraviolet light
185 signatures in melanoma (*e.g.*, SBS7a: *FDR* = 3.6 x 10^{-16}, FC = 1.24) and the tobacco-associated
186 signature SBS4 in the two cohorts of lung cancers (*FDR* ≤ 10^{-3}, FC ≥ 1.10). The enriched

187 mutational signatures at TSSs match the major mutagens and exposures of these cancer types,
188 indicating an overall increased vulnerability of TSSs to mutational processes. This analysis
189 confirms previous reports of increased mutation rates in promoters in melanoma [24,25] and
190 indicates that TSS-specific mutational processes are widely active in the pan-cancer context.

191 Pan-cancer open chromatin regions defined as ATAC-seq profiles of primary tumors were also
192 enriched in mutations in 11/25 cancer types and the pan-cancer cohort, especially among C>G
193 mutations (*e.g.*, pan-cancer, *FDR* = 1.0 x 10^{-27}, FC = 1.08) (**Figure 1D**). However, the effect
194 sizes of mutational enrichments were more modest compared to CTCF binding sites and TSSs,
195 potentially due to mixed effects of SNVs and indels: mutation enrichments in open-chromatin
196 sites were primarily driven by SNVs, while in contrast, indel mutations were significantly
197 depleted in the pan-cancer cohort and in 9/25 cancer types, indicating that the open chromatin
198 environment or the binding of regulatory elements may be protective of the mutational processes
199 responsible for generating indels. For example, in uterine adenocarcinoma, 1,762 indel mutations
200 were observed in open-chromatin sites while 2,172 were expected according to RM2 (*FDR* = 5.3
201 x 10^{-9}, FC = 0.82) (**Figure 1E**). This indel depletion appeared relatively stronger in intergenic
202 sites of open chromatin. Further, mutational signature analysis indicated reduced activity of the
203 aging-related signature SBS1 in open-chromatin sites, apparent in eight cancer types and the pan-
204 cancer cohort (*e.g.*, colorectal adenocarcinoma, *FDR* = 3.8 x 10^{-5}, FC = 0.90). However, analysis
205 of these pan-cancer open chromatin regions is better powered compared to the analysis of TSS
206 loci due to a larger number of sites, thus smaller deviations in local mutation rates were
207 detectable. Our findings contrast an earlier report that indicated broadly decreased mutation rates
208 at chromatin-accessible regulatory elements derived from cell lines [29]. Comparison of localised
209 mutation rates at TSS loci and open-chromatin sites indicate distinct properties of localised
210 mutagenesis acting on proximal and distal regulatory elements of the genome.

211 We extended this analysis to benchmark our model and evaluate statistical power. First, to assess
212 model calibration and false positive rates, we analysed a PCAWG dataset of simulated variant
213 calls designed to approximate neutral genome evolution [4]. As expected, analysis of simulated
214 data did not reveal any significant differences in mutation rates in the three classes of elements
215 (all *FDR* > 0.05). Quantile-quantile analysis of *P*-values confirmed that the model is well
216 calibrated for true and simulated mutations (**Supplementary Figure 2A**). Second, we evaluated

217 the statistical power of our model by analyzing down-sampled subsets of liver cancer genomes
218 and CTCF sites (**Supplementary Figure 2B**). For example, the mutation enrichment in CTCF
219 sites was detectable 80% of the time when sampling 75 genomes and 75,000 CTCF binding sites.
220 Third, we varied the parameter corresponding to the normalised width of genomic elements for
221 the three classes (**Supplementary Figure 2C**). Local differences in mutation rates were robustly
222 detected for multiple element widths. However, mutation enrichments in chromatin architectural
223 elements bound by CTCF were generally focused at narrower regions (50 bps) compared to
224 gene-regulatory elements at TSS and open-chromatin regions (200 bps). This is reflective of their
225 different element widths and indicative of differences in underlying mutational processes. In
226 summary, our method provides a versatile and well-calibrated framework for analyzing localised
227 mutation rates and mutational processes in cancer genomes.

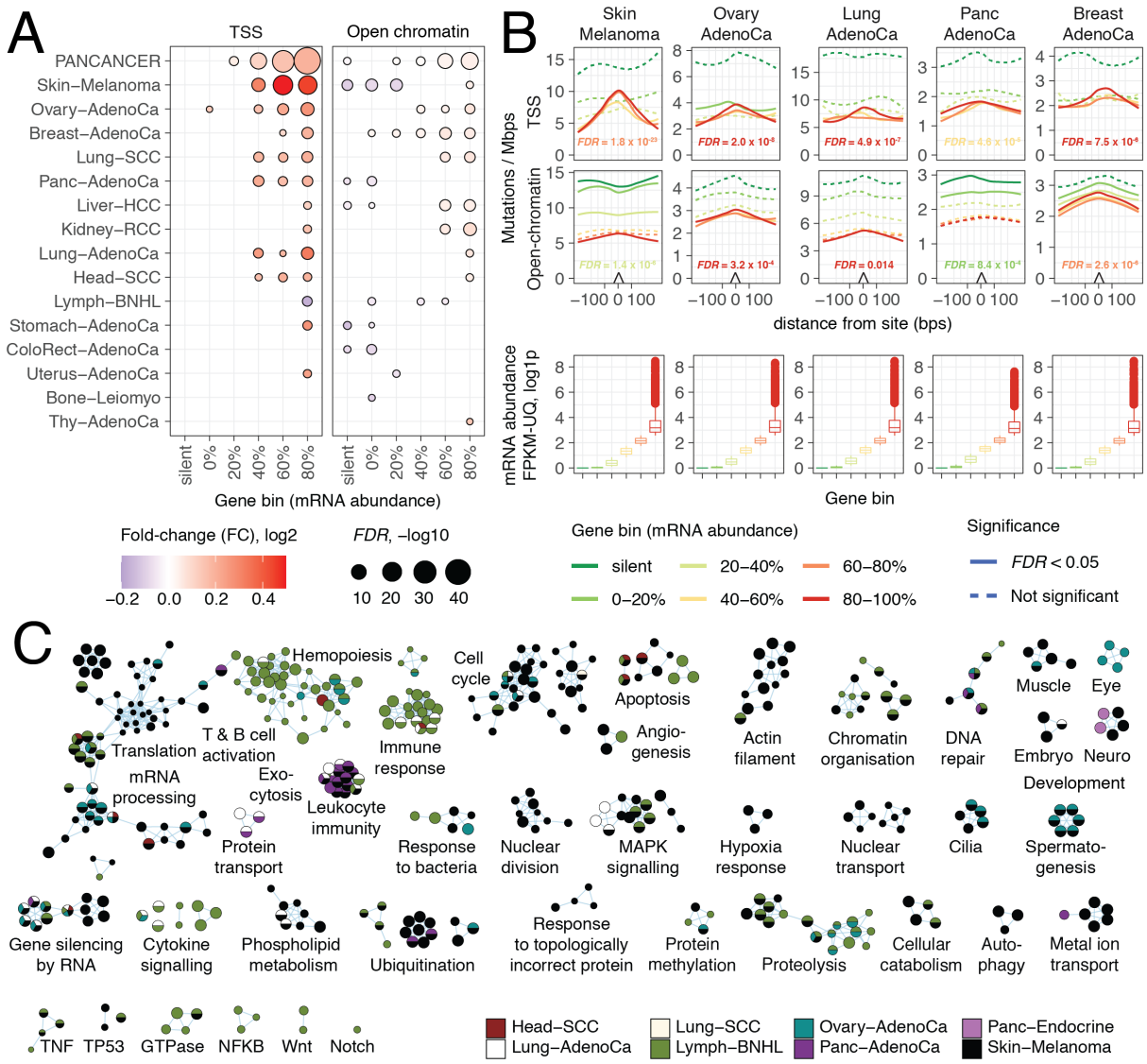

Figure 2. Increased mutation rates at transcription start sites (TSS) of highly expressed genes. A. Comparison of localised mutation rates in TSSs and open-chromatin sites (TCGA ATAC-seq), grouped by mRNA abundance of target genes in matching tumors (*FDR* < 0.05). **B.** Examples of cancer types with strong transcription-associated mutation rate in TSSs (top) compared to open-chromatin sites (middle). Per-patient mutation rates are shown on the Y-axis with loess smoothing. Colors indicate bins of genes based on median mRNA abundance. The *FDR* value of the most significant bin is shown for each cancer type. Bottom: boxplots show mRNA abundance in gene bins. **C.** Enrichment map of pathways and processes with increased mutation rates at TSSs (*FDR* < 0.1). Nodes in the network indicate pathways or processes that are connected with edges if these share many common genes. Node colors indicate the cancer types with frequent TSS-specific mutations.

Functional associations of gene-regulatory and chromatin architectural mutations

231 We asked whether the increased mutational load at TSSs and open-chromatin sites correlated
232 with transcriptional activity of target genes. To precisely quantify mRNA abundance in different
233 cancer types, we used matching RNA-seq data in the PCAWG project [44]. TSS-specific mutation
234 rates were studied in six bins of genes grouped by median mRNA abundance in 19/25 cancer
235 types. TSS-specific mutation rates were significantly increased in highly expressed genes in
236 eleven cancer types (*FDR* < 0.05) and the pan-cancer cohort (*FDR* = 2.2 x 10^{-50}, FC = 1.16)
237 (**Figure 2A**) (**Supplementary Table 1B**). Association of transcriptional activity and TSS-
238 specific mutation rates was the strongest in melanoma and ovarian, lung, pancreatic, and breast
239 adenocarcinoma where genes with the highest mRNA abundance were strongly enriched in TSS-
240 specific mutations (*FDR* ≤ 10^{-4}, FC ≥ 1.15) (**Figure 2B**). Of note, the top gene bin was highly
241 variable in mRNA abundance (*e.g.*, 11-7,100 RPKM-UQ in the pan-cancer cohort). In silenced
242 genes, TSS loci consistently showed no significant differences in mutation rates relative to
243 flanking controls (*FDR* > 0.05), however the overall mutation rate in sites and controls was
244 higher, potentially a result of lower rates of transcription coupled repair in closed chromatin.
245 This analysis highlights a localised mutational process in gene promoters apparent in multiple
246 cancer types that is potentially driven by transcriptional initiation or TF binding at core gene-
247 regulatory promoter elements.

248 Compared to TSSs, mutation rates in open-chromatin sites showed only limited associations with
249 mRNA abundance (**Figure 2B**). The open-chromatin sites of the most highly expressed genes
250 were significantly associated with higher mutation rates, however the effect sizes of fold-changes
251 were consistently lower (*e.g.*, pan-cancer, *FDR* = 9.3 x 10^{-17}, FC = 1.04). We performed a down-
252 sampling analysis of open-chromatin sites and found no significant associations of highly
253 expressed genes and localised mutation rates when considering random subsets of sites
254 comparable with the analysis of TSSs. Since the number of open-chromatin sites is considerably
255 larger, the observed localised increase in mutation rates can be partly attributed to the improved
256 statistical power that allows detection of smaller effects (**Supplementary Figure 4**). Thus, our
257 observed transcription-dependent mutational process appears to be more active at promoters
258 while the broader spectrum of proximal and distal pan-cancer regulatory elements is less
259 affected.

260 To understand the functional associations of TSS mutations, we performed a pathway
261 enrichment analysis by adapting RM2 to gene sets of GO biological processes and Reactome
262 molecular pathways [45]. This allowed us to test our hypothesis that promoters enriched for non-
263 coding mutations are concentrated in specific biological processes. We found 546 unique
264 pathways and processes exhibiting elevated TSS-specific mutation rates (*FDR* < 0.1) (**Figure**
265 **2C**) (**Supplementary Table 1C**), the majority of which were found in the melanoma cohort
266 (70%) while 28% of pathways were found in more than one cancer type. Translation, ribosome
267 biogenesis and RNA processing were among the largest groups of pathways with increased TSS-
268 specific mutation rates. This is expected as the translational machinery is ubiquitously active in
269 proliferating cells and includes many highly expressed genes. Cancer-related processes and
270 pathways were also enriched in TSS mutations, for example mitotic cell cycle, apoptosis, DNA
271 repair, angiogenesis, developmental and immune response processes as well as druggable
272 signalling pathways (*e.g.*, MAPK, Wnt, Notch) were identified in multiple cancer types. The
273 pathway analysis augments our observation of frequent TSS mutations associating with
274 increased transcription and highlights a variety of core cellular processes with high baseline
275 transcription and associated localised mutagenesis active in many cell types. Furthermore, the
276 significant enrichment of TSS-specific mutations in genes of cancer-related processes suggests
277 that some more frequent non-coding mutations at individual promoters are functional and may
278 contribute to cancer driver mechanisms and tumor heterogeneity by altering gene-regulatory
279 circuits in molecular pathways and interaction networks [37].

280 We asked whether the elevated mutation rates in CTCF binding sites were also associated with
281 their functional characteristics. We used the extent of conservation of DNA-binding across cell
282 types as a proxy of site functionality across an extended set of 162,209 CTCF binding sites
283 catalogued in 70 cell lines in ENCODE. We grouped CTCF binding sites into five equal bins
284 based on the number of cell lines where the site was detected, with entirely tissue-specific CTCF
285 binding sites in bin one and the constitutively bound subset of sites detected in most or all cell
286 types in bin five (median conservation in 67 cell lines; range 52-70) (**Figure 3A**). Strikingly, the
287 localised elevations of mutation rates were exclusively detected in the subset of constitutively
288 bound CTCF binding sites, as observed in eleven cancer types and the pan-cancer cohort (*FDR* =
289 3.3×10^{-89}, FC = 1.23) (**Figure 3B**) (**Supplementary Table 1D**). In contrast, all other bins

290 showed no significant enrichment of mutations (*FDR* > 0.05). Additional cancer types with
291 frequent CTCF binding site mutations were uncovered in this more focused analysis, including
292 colorectal and pancreatic cancer and lymphoma (BNHL) (*FDR* ≤ 10^{-6}, FC ≥ 1.18). This extends
293 the established pattern of CTCF mutation enrichment to a subset of pan-tissue genomic elements
294 and highlights the utility of RM2 in detecting mutational patterns. Since CTCF is known as the
295 master regulator of genome architecture, we incorporated a HiC dataset of chromatin long-range
296 interactions [46] to interpret the mutation enrichments at highly conserved CTCF binding sites. We
297 found that these constitutively bound CTCF binding sites were strongly enriched in chromatin
298 loop anchor elements: 30% (9,393/32,442) of the sites of bin five were located within 1 kbps of
299 anchor midpoints (3,973 or 12% expected, Fisher's exact *P* = 0). Contrarily, the majority of
300 CTCF binding sites that were only detected in a few cell types showed no deviation from
301 expected mutation rates and an expected distribution with respect to chromatin loop anchors.
302 This indicates the mutational process primarily affects the subset of CTCF sites that are
303 constitutively bound in most cell types and participate in chromatin architectural and gene
304 regulatory interactions [46,47]. Conservation is a property of functionally integral CTCF binding
305 sites, which upon disruption, can lead to changes in underlying chromatin architecture and gene
306 regulation [48] and is associated with activation of proto-oncogenes [26].

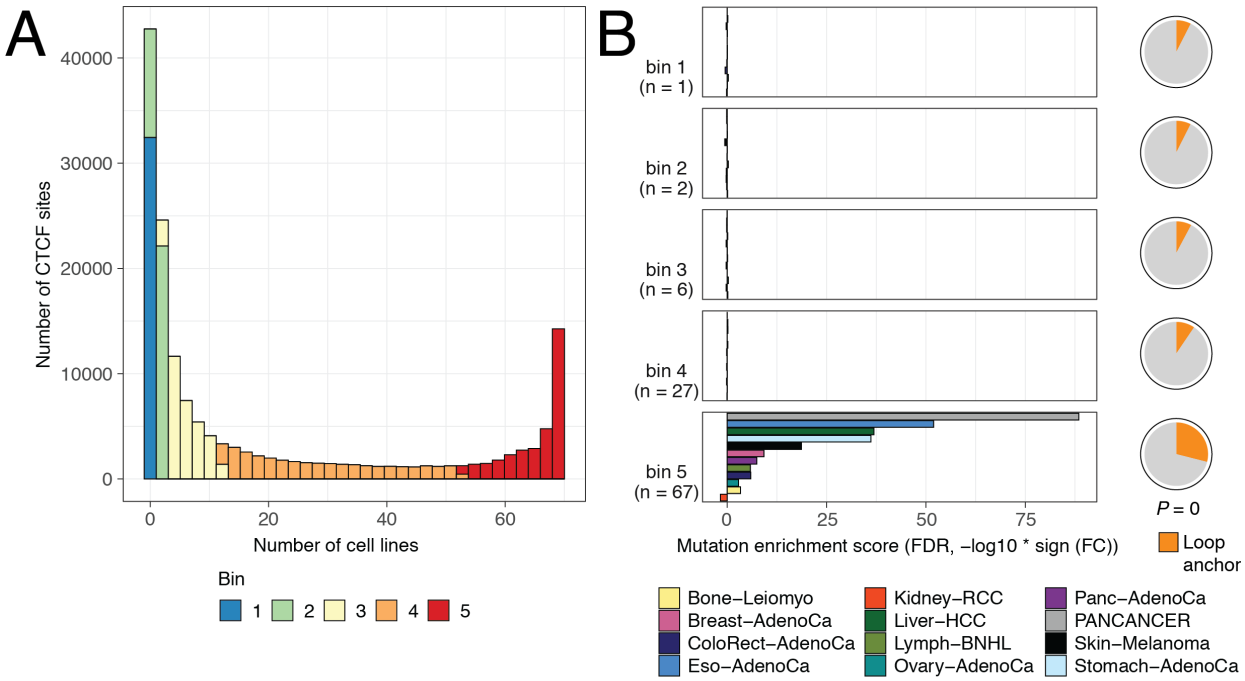

Figure 3. Localised mutation rates at constitutively active binding sites of CTCF. A. Histogram of CTCF binding sites with respect to the number of cell lines with sites observed. CTCF binding sites were grouped into five equal bins based on site activity across the cell lines in ENCODE. The bimodal distribution reveals a subset of sites detected in most or all cell types. **B.** Comparison of mutation rates in the five bins of CTCF binding sites. Mutation enrichment score is on the X-axis. Pie charts (right) show the proportion of sites located at chromatin loop anchors. Median number of cell lines per bin is shown in brackets.

307

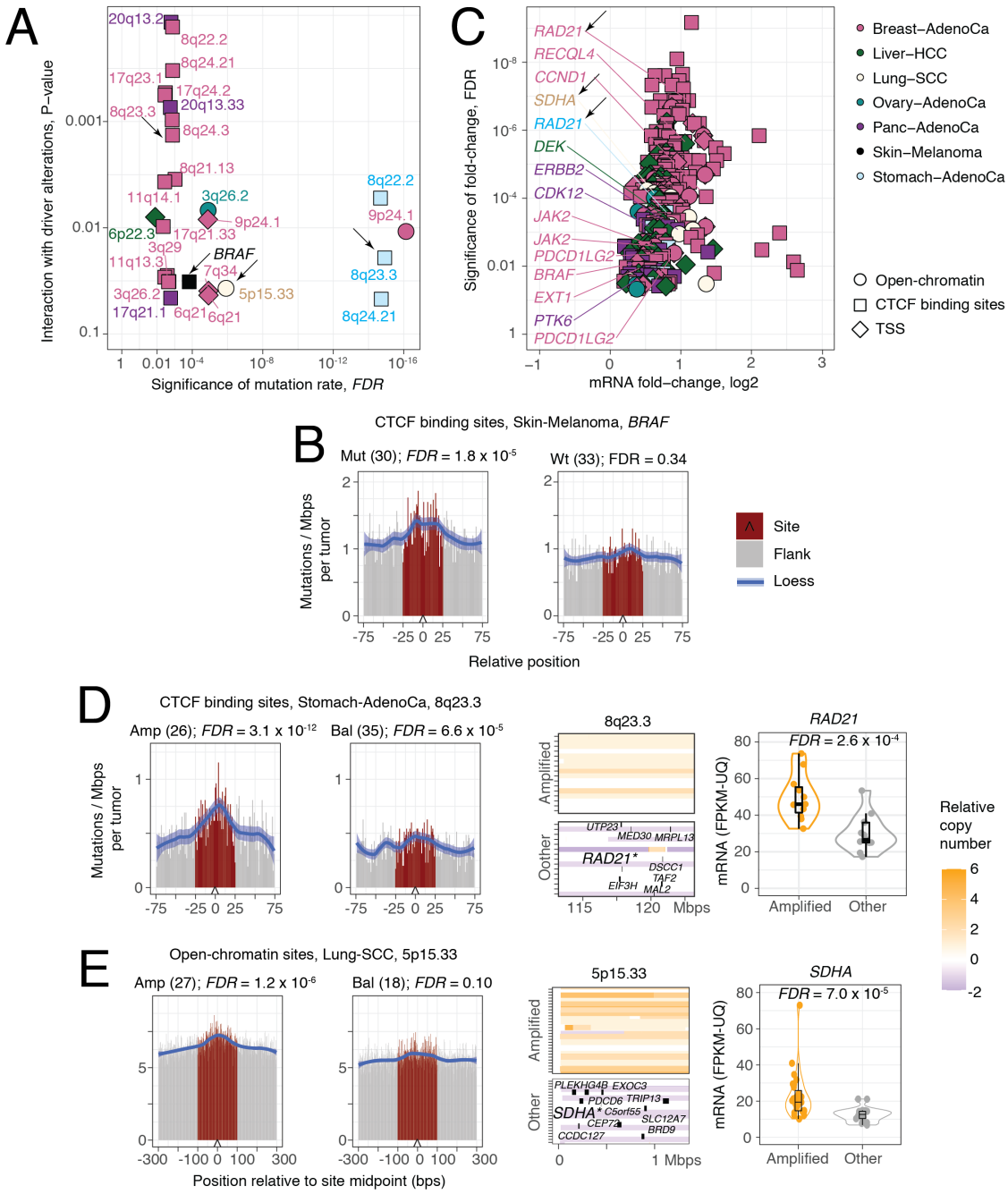

Figure 4. Driver mutations and amplifications associated with gene-regulatory and architectural mutation rates. A. Scatter plot shows the overall significance of localised mutations (X-axis; *FDR* < 0.05) and their statistical interactions with drivers and CNAs (Y-axis; *P* < 0.05). Arrows highlight the cases below. **B.** Presence of *BRAF* driver mutations in melanoma associates with an increased mutation rate at CTCF binding sites. *BRAF*-mutant (left) and wildtype tumors (right) are shown. **C.** Volcano plot shows the genes with amplification-driven increase in mRNA abundance located in the relevant CNA regions (from panel A). Known cancer genes are highlighted. **D.** CTCF binding site mutations in stomach cancer are associated with *RAD21* (cohesin) amplifications in 8q23.3. Left, two plots: localised mutation rates in CTCF binding sites in 8q23.3-amplified and non-amplified tumors. Middle: genomic copy number in amplified and non-amplified tumors. Eight genes with amplification-associated mRNA abundance increase are shown. Asterisks show known cancer genes. Right: *RAD21*, the common DNA-interaction partner of CTCF, is upregulated in 8q23.3-amplified tumors. **E.** Increased localised mutation rates in open-chromatin sites in lung squamous cell carcinoma (Lung-SCC) associate with 5p15.33 amplification. Ten genes in 5p15.33 show increased mRNA abundance in the tumors with 5p15.33 amplifications, including the mitochondrial enzyme and cancer gene *SDHA*.

309 **Localized mutation rates associate with recurrent mutations of *BRAF*, *RAD21* and *SDHA***

310 To find potential genetic mechanisms of localised mutagenesis, we asked whether any recurrent
311 mutations in tumor genomes could explain the observed mutation rate increases in the three
312 classes of genomic elements. We considered 14 cancer types with 15 driver genes with frequent
313 SNVs and indels detected using the ActiveDriverWGS method [34], and 60 recurrent copy-number
314 amplifications detected in the PCAWG project using the GISTIC2 method [49] (**Supplementary**
315 **Figure 5**). We found 27 driver alterations that positively interacted with local mutation rates,
316 including 26 recurrent copy-number amplifications and one driver gene with SNVs and indels
317 (RM2 $FDR < 0.05$, interaction $P < 0.05$) (**Figure 4A**) (**Supplementary Table 1E-F**). Significant
318 interactions were detected in seven cancer types, mostly in breast adenocarcinoma (17) and one
319 to three interactions each in stomach, pancreatic, ovarian, lung and liver cancers and melanoma.
320 Six genomic amplifications were identified twice (3q26.2, 6q21, 8q22.2, 8q23.3, 8q24.21,
321 9p24.1). The majority of interactions (20/27) were found for CTCF binding site mutations. The
322 large number of statistical interactions seen in breast cancer suggests that overall chromosomal
323 instability may contribute to mutational processes at CTCF binding sites [28] in this cancer type.
324 However, several specific examples of potential genetic factors contributing to local mutagenesis
325 were also found.

326 Driver mutations in *BRAF* in melanoma were associated with an increased mutation rate at
327 CTCF binding sites ($FDR = 5.5 \times 10^{-5}$, FC = 1.07, interaction $P = 0.032$). Comparison of driver-
328 mutated and wildtype tumors confirmed the interaction: 30 tumors with *BRAF* mutations showed
329 a significant increase in mutation burden at CTCF sites ($FDR = 1.8 \times 10^{-5}$, FC = 1.10) while the
330 remaining 33 *BRAF*-wildtype tumors showed no significant change ($FDR = 0.34$, FC = 1.02)
331 (**Figure 4B**). The activating V600E amino acid substitutions in the BRAF serine/threonine
332 kinase and proto-oncogene define a druggable subtype of melanoma [50,51]. Ectopic expression of
333 V600E-mutant *BRAF* in epithelial cell lines was shown to induce DNA double strand breaks and
334 reactive oxygen species [52]. In this cohort, 23 melanomas carried V600E substitutions and four
335 additional tumors had V600K substitutions. Therefore, the melanomas defined by *BRAF* driver
336 mutations may have an increased activity of the mutational process acting on CTCF binding
337 sites.

338 To further map the potential genetic mechanisms of local mutagenesis, we determined the genes
339 in the highlighted CNA regions that responded transcriptionally to genomic amplifications. We
340 found 260 unique genes in 19 recurrently amplified regions that were up-regulated in tumors
341 with amplifications (Wilcoxon test, *FDR* < 0.05) (**Figure 4C**) (**Supplementary Table 1G**).
342 These associations were identified for all three categories of genomic elements and six cancer
343 types were represented (breast, liver, lung squamous, ovary, stomach and pancreatic). Eleven
344 known cancer genes were found in single cancer types (*RECQL4*, *CCND1*, *SDHA*, *DEK*,
345 *CDK12*, *ERBB2*, *JAK2*, *PDCD1LG2*, *BRAF*, *EXT1*, *PTK6*) according to the Cancer Gene Census
346 database. Notably, the cohesin subunit *RAD21* was the only gene associated with localized
347 mutagenesis in two cancer types. Thus amplification-driven activation of hallmark cancer genes
348 may directly or indirectly affect the activity of localised mutational processes.

349 Amplification and transcriptional up-regulation of *RAD21* located at 8q23.3 was associated with
350 increased mutation rates in CTCF binding sites in stomach and breast cancers. *RAD21* encodes a
351 subunit of the cohesin complex that co-binds DNA with CTCF to orchestrate transcriptional
352 insulation and chromatin architectural interactions [46,47]. Stomach cancer genomes with
353 amplifications of 8q23.3 showed a strong increase in mutations in CTCF binding sites (*FDR* =
354 3.1×10^{-12}, FC = 1.24) while the increase was less significant in tumors with no amplification
355 (*FDR* = 6.6×10^{-5}, FC = 1.13; interaction *P* = 0.020) (**Figure 4D**). mRNA abundance of *RAD21*
356 in 8q23.3-amplified stomach cancers was significantly higher compared to non-amplified
357 samples (*FDR* = 2.6×10^{-4}, 46 *vs.* 27 median FPKM-UQ), suggesting that *RAD21* expression is
358 driven by the genomic amplification. This association was also observed in breast cancer where
359 the genomes with 8q23.3 amplifications showed an increased mutation rate in CTCF binding
360 sites (*FDR* = 5.3×10^{-5}, FC = 1.08) that was not apparent in tumors lacking 8q23.3
361 amplifications (*FDR* = 0.36, FC = 0.98; interaction *P* = 5.0×10^{-4}). *RAD21* amplification was
362 also associated with increased mRNA abundance in breast cancer (*FDR* = 1.4×10^{-7}, 108 *vs.* 50
363 median FPKM-UQ), which is also indicative of poor prognosis [53]. In the ENCODE dataset, 51%
364 of CTCF binding sites were also co-bound by cohesin, representing 94% of the high-confidence
365 RAD21 binding sites in ENCODE (60,636 / 64,483), significantly more than expected by chance
366 (666 expected, binomial *P* = 0). Mutations in another cohesin subunit, *STAG2,* have been
367 associated with specific mutational signatures and altered transcription at double strand break

368 sites [54]. In addition to *RAD21*, components of the general transcriptional machinery were
369 amplified in 8q23.3 and showed increased mRNA abundance in breast and stomach cancers,
370 potentially contributing to mutation rates in CTCF binding sites. These included *MED30,* which
371 encodes a subunit of the Mediator transcriptional coactivation complex that interacts with
372 cohesin to connect promoters and enhancers [55], as well as the transcription initiation factor
373 subunit encoded by *TAF2*. Alternative explanations to CTCF binding site mutations were also
374 apparent in our data. For example, interactions with the amplification of 8q24.21 encoding the
375 *MYC* oncogene were identified in both breast and stomach cancer, however these amplification
376 events did not associate with *MYC* up-regulation. This analysis suggests that the elevated
377 mutagenesis at CTCF binding sites may be driven by the genomic amplification and up-
378 regulation of core transcriptional and genome architectural machinery interacting with CTCF.

379 As another example, amplification of 5p15.33 was associated with increased mutation rates at
380 open-chromatin sites in lung squamous cell carcinoma (Lung-SCC) (**Figure 4E**). 27 tumors with
381 this amplification showed an elevated mutation rate in open-chromatin sites (*FDR* = 1.2 x 10^{-6},
382 FC = 1.03) while 18 non-amplified tumors showed no significant difference (*FDR* = 0.10, FC =
383 1.01; interaction P = 0.037). The cancer gene *SDHA* and nine other genes in the region showed
384 significant up-regulation in tumors with this amplification (*FDR* = 7.0 x 10^{-5}, 19 *vs.* 13 FPKM-
385 UQ). *SDHA* encodes a subunit of the mitochondrial succinate dehydrogenase (SDH) complex
386 involved in cellular energy metabolism through the citric acid cycle and the electron transport
387 chain. Germline mutations of the tumor suppressor *SDHA* and the genes encoding other SDH
388 subunits predispose individuals to the neuroendocrine tumors pheochromocytomas and
389 paragangliomas [56,57]. Mutations and inhibition of SDH subunits are associated with increased
390 oxidative stress, production of reactive oxygen species (ROS) and activation of the hypoxia-
391 inducible factor HIF [58,59]. These data suggest that the increased mutation rate in open-chromatin
392 sites in Lung-SCC is associated with the genomic amplification and up-regulation of *SDHA* that
393 leads to the destabilization of the SDH complex and results in increased oxidative damage in
394 open-chromatin sites.

395 In summary, our analysis provides a catalogue of potential genetic mechanisms underlying
396 localised mutation rate variations in cancer genomes. While a subset of these driver mutations
397 and recurrent copy-number amplifications may be directly involved in processes of mutagenesis

398 and DNA repair, others may represent tumor subtypes with specific exposures or endogenous
399 factors.

400

401 **Discussion**

402 The cancer genome is molded by diverse mutational processes that continuously shape its broad
403 megabase-scale features and the fine context of nucleotide signatures. Here we focused on the
404 mutational processes of an intermediate scale that affect thousands of functional genomic
405 elements each spanning dozens to hundreds of nucleotides. Using a novel computational
406 framework, we mapped widespread enrichments and some depletions of mutations in gene-
407 regulatory and chromatin architectural elements. These non-coding mutations associated with
408 transcriptional and pathway activity, trinucleotide mutational signatures, and conservation of site
409 activity across cell types. We found interactions with recurrent driver mutations and copy-
410 number amplifications that provide hypotheses regarding the mechanisms of the underlying
411 mutational processes, for example copy-number alterations and driver mutations in *RAD21*,
412 *SDHA* and *BRAF* were indicative of increased mutation rates in CTCF binding sites and open-
413 chromatin sites. In particular, the finding of *RAD21* amplifications associating with CTCF
414 binding site mutations fits with our observation of the constitutively active, hypermutated subset
415 CTCF binding sites enriched at chromatin loop anchors, as cohesin and CTCF co-bind DNA to
416 facilitate chromatin architectural interactions. Overall, we speculate that the localised mutations
417 represent a functional continuum of passengers and drivers. On the one hand, the mutation rates
418 of these genomic elements likely deviate from the background rates due to focal carcinogen
419 exposures or interactions with transcriptional, DNA replication or repair machinery that make
420 these sites more or less vulnerable to mutations. On the other hand, some of these functional
421 elements may control transcription regulatory interactions or epigenetic states of genes and
422 pathways involved in cancer and their excess mutations reflect positive selection. While it is
423 unlikely that all functional elements of a specific class would positively impact oncogenic
424 processes when mutated, specific subsets of elements may contribute to hallmarks of cancer as
425 suggested by our pathway analysis. Our computational framework and the detailed catalogue of

426 localised mutational processes and genetic interactions detected in a large pan-cancer dataset
427 provides specific hypotheses for further study.

428 Our analysis has certain caveats and limitations. We analysed a generic catalogue of genomic
429 elements that only provides limited representation of the primary tumors in our cohort. To
430 address this limitation, we used matching RNA-seq data to stratify regulatory elements based on
431 their activity in specific cancer types and also considered open-chromatin profiles of primary
432 tumors in the first analysis of this kind. Future analyses will benefit from detailed multi-omics
433 cohorts with matching genomic, transcriptomic and epigenomic profiles of individual tumors.
434 Also, the current framework is designed for genomic elements of uniform width where analysis
435 of elements of variable width, such as exons or non-coding RNAs, may lead to statistical biases.
436 Further, our analysis suggests that different classes of gene-regulatory and architectural elements
437 of the genome may be subject to localised mutational processes that have footprints of different
438 sizes. Thus it is recommended to evaluate that input parameter of the method when analysing
439 new genomic elements. Our method is designed to quantify localized differences of mutation
440 rates acting on an entire class of genomic elements with thousands to hundreds of thousands of
441 genomic loci. It is not powered to evaluate a single genomic element as a potential cancer driver
442 and alternative methods should be used for this purpose. However, we have adapted our method
443 to evaluate TSS-specific mutation rates in gene sets of representing biological processes and
444 pathways with hundreds to thousands of genes.

445 Our study enables a number of future developments. Integrative analysis of whole cancer
446 genome sequences and rich clinical and pathological profiles of tumors [17] may highlight
447 associations of clinical variables and localised mutagenesis and lead to the discovery of novel
448 WGS-based biomarkers. Considering patient lifestyle information, environmental exposures and
449 germline variation in the analysis may elucidate the impact of carcinogens and endogenous DNA
450 repair deficiencies. Our catalogue of genetic associations provides hypotheses on the mutational
451 mechanisms that can be tested experimentally using genome editing and mutagenesis assays.
452 Rare germline variants in the human population [60], *de novo* variants detected in genetic disorders
453 [21] and the widespread somatic genome variation found in healthy tissues [61] provide further
454 avenues to study mutational processes acting at functional non-coding elements. Our study

455 enables a detailed annotation of localised mutational processes in whole genomes to decipher
456 cancer driver mechanisms, molecular heterogeneity and genome evolution.

457

458 **Methods**

459 **Regression models for Regionalised Mutations (RM2).** Local differences in mutation rates in
460 functional genomic elements (*i.e.*, sites) were evaluated using a negative binomial regression
461 model we refer to as RM2. Single nucleotide variants (SNVs) and small insertions-deletions
462 (indels) were analysed. The model simultaneously considers a collection of non-coding sites,
463 such as regulatory elements that are commonly ~10–1,000 bps in length and measured in ChIP-
464 seq and related experimental assays in thousands to hundreds of thousands of genomic loci. Sites
465 were uniformly redefined using their median coordinate and added sequences of fixed width
466 upstream and downstream of the sites (*e.g.*, ±25 bps or 50 bps around the midpoints of CTCF
467 binding sites). Upstream and downstream flanking sequences of these sites were used as control
468 regions to estimate expected mutation rates. Control regions were of equal width to sites such
469 that the upstream and downstream regions combined were twice as wide as the sites. To account
470 for megabase-scale variation in mutation rates, we computed the total log-transformed mutation
471 count for each site within its one-megabase window (i.e., ±0.5 Mbps around site midpoint).
472 Based on this estimate, all sites were distributed into ten equal bins (*MbpRate*). The value of ten
473 bins worked well in our benchmarks and captured variation in smaller and larger cohorts of
474 individual cancer types. However, custom values of this parameter can be used. Mutation rates
475 for sites and flanking sequences for each bin were defined separately and sites and flanking
476 sequences were distinguished by a binary cofactor (*isSite*). To avoid inflated counts, mutations
477 affecting more than one site were counted once. Likewise, mutations affecting the flanking
478 sequence of more than one site were also counted once. Sequence positions were counted
479 separately by their trinucleotide context (*nPosits*) and expanded to three alternative nucleotides
480 to account for the potential sequence space where such single nucleotide variants could occur
481 (*nPosits*). The observed mutations in these contexts were also counted (*nMuts*) and a cofactor
482 was used to add separate weights to different trinucleotide classes (*i.e.,* reference trinucleotide
483 and alternative nucleotide; *triNucMutClass*). Indels were counted under another entry in
484 *triNucMutClass* such that all mutation counts were summed and the entire genomic space was
485 accounted for. An optional binary cofactor (*coFac*) was included to allow the consideration of
486 genetic or clinical covariates of localised mutation rates. To evaluate the significance of localised

487 mutation rates in sites compared to flanking control regions, we first constructed a null model by
488 excluding the term *isSite*:

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

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

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

493 We extended our model to evaluate whether localised mutation rates differed between two
494 subtypes of tumors, such as those defined by clinical annotations or genetic features using the
495 term *coFac.* Trinucleotide sequence content, trinucleotide mutational signatures and megabase-
496 scale covariations of mutation rates were computed separately for the two sets of tumors. To
497 establish the associations of localised mutation rates and driver mutations, we added to the initial
498 model the term *isSite:coFac* mapping the interaction of the tumor subtype and the cofactor
499 distinguishing sites and flanking sequences.

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

501 *coFac* + *isSite* + *isSite:coFac* `)`.

502 We used likelihood ratio tests to compare the models and evaluate the significance of localised
503 mutation rates (H_{alt} *vs.* H_{null} to evaluate the term *isSite*). Chi-square *P*-values from the likelihood
504 ratio tests were reported for each analysis. We also reported coefficient values of the term *isSite*
505 to characterise enrichment or depletion of mutations at sites relative to flanking controls for
506 positive and negative values, respectively. The interactions of driver mutations and mutation
507 rates were evaluated using likelihood ratio tests that compared the models H_{alt} and H_{cof}. Only the
508 models with significant positive coefficients were reported. The expected mutation counts were
509 derived from each model by 1000-fold sampling of mutation counts from the negative binomial
510 distribution informed by the fitted probabilities and theta values derived from the regression
511 models. Fold-change values were derived by dividing median observed and expected mutation
512 counts, and confidence intervals were derived using the 2.5^{th} and 97.5^{th} percentiles of sampled
513 values. Chi-square *P*-values from the models were corrected for multiple testing using the

514 Benjamini-Hochberg procedure where appropriate. Besides modelling total mutations in sites
515 and flanking sequences, we evaluated mutations of multiple sub-classes, such as mutations
516 stratified by transcriptional activity, COSMIC mutational signatures or DNA strands. Mutation
517 subclass analysis was conducted as described above. The same megabase-scale mutation rates
518 estimated for all mutations were used rather than those of the specific subclasses. The method is
519 available at https://github.com/reimandlab/RM2.

520 **Somatic mutations in whole cancer genomes.** Somatic single nucleotide variants (SNVs) and
521 short insertions-deletions (indels) in the genomes of 2,583 primary tumors were retrieved from
522 the uniformly processed dataset of the Pan-cancer Analysis of Whole Genomes (PCAWG)
523 project of the ICGC and TCGA [3]. We used consensus variant calls mapped to the human genome
524 version GRCh37 (hg19). We removed 69 hypermutated tumors with at least 30 mutations per
525 Mbps, resulting in 2,514 tumors and 24.7 million mutations. We also removed tumors for which
526 mutational signature predictions were not available in PCAWG. We analysed tumor genomes of
527 the pooled pan-cancer cohort of multiple cancer types, and also 25 cohorts of specific cancer
528 types with at least 25 samples in the PCAWG cohort. We excluded a small subset of tumors (31
529 or 1.3%) where localised mutation rates were exceptionally strong even when analysing one
530 tumor genome at a time ($FDR < 0.001$, RM2). Based on our initial analyses, we found that the
531 individual contribution of these tumors to the overall analysis would have caused overestimates
532 of mutation rates. To enable this filtering, we performed tumor specific analyses for the three
533 classes of sites (open-chromatin sites, binding sites of CTCF, and TSSs). We analysed each
534 cohort of a cancer type separately and grouped the mutations according to tumor ID, allowing the
535 model to learn an expected background mutation rate in the respective cohort and then test each
536 tumor genome separately for localised mutation rates. To perform this single-tumor analysis in
537 smaller cohorts within the PCAWG dataset (<25 tumors of a given type), we created a meta-
538 cohort by pooling these smaller cohorts. After filtering hypermutated tumors, tumors without
539 PCAWG signatures, and tumors with exceptionally strong signals of localised mutations, we
540 derived a conservative final set of 2,421 genomes of 35 cancer types with 23 million mutations
541 including 1.61 million indels. To evaluate the performance of our model, we also processed a
542 dataset of simulated variant calls for the same set of tumors derived from the PCAWG project
543 (*i.e.*, the Broad dataset) [4].

544 **Mutation features and signatures.** In addition to evaluating total mutations, several classes of
545 mutations were analysed separately. Mutations were mapped to C and T nucleotides and grouped
546 by reference and alternative nucleotides (C>[A,G,T], T>[A,C,G]). Four additional classifications
547 were developed. First, mutations were classified as located either on the Watson (w) strand if the
548 original reference nucleotide was C or T, or the Crick (c) strand if the original reference
549 nucleotide was A or G. Second, transcriptional activity and orientation of the mutated
550 nucleotides was mapped based on the coordinates of protein-coding genes defined in the
551 Ensembl database (GRCh37) using 500 bps flanking sequence at both ends of genes to account
552 for transcriptional initiation and termination. We then classified mutations as forward-transcribed
553 (F), reverse-transcribed (R), bidirectionally transcribed (B), or not transcribed (O). This initial
554 classification did not include information on tissue-specific transcription and was augmented
555 using matching tumor-specific mRNA abundance data, as described below. Third, mutation
556 strand and transcription status were combined into eight categories (w_[F,R,B,O] and
557 c_[F,R,B,O]). Fourth, we classified mutations by the trinucleotide signatures of single base
558 substitutions (SBS) that were derived earlier using the SigProfiler software in the PCAWG
559 project [14]. We assigned each mutation to its most probable signature in the given patient tumor
560 based on its trinucleotide context. For model evaluation, these five major categories of mutations
561 were also derived for the dataset of simulated variant calls.

562 **Chromatin architectural and gene-regulatory genomic elements.** We performed a systematic
563 analysis of three classes of genomic elements: DNA-binding sites of CTCF (CCCTC-binding
564 factor) detected in multiple human cell lines, transcription start sites (TSS) of protein-coding
565 genes, and open-chromatin sites (ATAC-seq sites) detected in human primary tumors. CTCF
566 binding sites were retrieved from the ENCODE project [39]. Sites observed in only one cell line
567 were removed, resulting in 119,464 sites across 70 cell lines. TSS loci of protein-coding genes
568 were retrieved from Ensembl Biomart (GRCh37) and filtered based on location of standard
569 chromosomes (1-22, X, Y), resulting in 37,309 TSSs of 18,710 genes. Open-chromatin sites of
570 410 primary tumors defined by ATAC-seq were retrieved from the TCGA study [40]. We used the
571 pan-cancer set of sites in the GRCh37 genome as defined in the study and filtered sites on non-
572 standard chromosomes and those lacking defined coordinates in GRCh37. This resulted in
573 561,057 open-chromatin sites. For the mRNA-based analysis of mRNA abundance described

574 below, we further filtered open-chromatin sites based on their target genes as defined in the
575 original study. We selected the subset of open-chromatin sites where predicted target genes were
576 available, mapped the gene symbols to ENSG identifiers using Ensembl Biomart (GRCh37,
577 release 100), and removed open-chromatin sites with missing or ambiguous gene symbols,
578 resulting in 438,948 sites annotated to 17,116 genes. Throughout the study, the three classes of
579 sites were normalised to uniform width based on median coordinates. CTCF binding sites were
580 defined using 50 bps (±25 bps) windows around the midpoint of sites. Midpoints of TSS loci
581 were defined in the Ensembl database and we used a 200 bps (±100 bps) window around the
582 TSSs. Open-chromatin sites were also defined using a 200 bps (±100 bps) window around site
583 midpoints. A systematic analysis was used to explore various values of the site width parameters
584 and the final selection was based on the strength of signal and consistency (**Supplementary**
585 **Figure 2**). For CTCF binding site analysis, we also retrieved DNA-binding sites of the cohesin
586 protein RAD21. These sites were also retrieved from the ENCODE dataset and those only
587 observed in single cell lines were filtered, resulting in 64,483 high-confidence sites. The majority
588 (94% or 60,636) of high-confidence RAD21 sites overlapped with high-confidence CTCF sites
589 (*i.e.*, those observed in at least two cell lines in ENCODE). We evaluated the enrichment of
590 RAD21 binding sites in CTCF binding sites with a binomial test, using RAD21-bound fraction
591 of the human genome (kbps) as the expected probability, and total sequence coverage of CTCF
592 sites (kbps) and RAD21-cobound CTCF sites (kbps) as the numbers of tries and successes,
593 respectively.

594 **Grouping gene-regulatory sites by mRNA abundance.** TSS and open-chromatin sites were
595 analysed in groups based on the mRNA abundance of associated genes in matching tumors. TSS
596 target genes were retrieved from the Ensembl database and target genes of open-chromatin sites
597 were retrieved from the original TCGA study. This analysis was carried out in 19 cohorts of
598 cancer types with at least 20 tumor samples with mRNA and WGS data, as well as the pan-
599 cancer cohort of all cancer types. We used the uniformly processed PCAWG RNA-seq dataset [44]
600 (RPKM-UQ) and applied the same filtering of tumor samples described previously and excluded
601 non-coding genes. Additionally, we discarded a subset of genes with duplicated HGNC symbols
602 as well as the genes for which TSS or open-chromatin sites were not mapped. This resulted in
603 mRNA measurements for 20,042 protein-coding genes in 1,267 tumor transcriptomes. Next we

604 derived the gene lists grouped by median mRNA abundance. Six exclusive lists of genes were
605 compiled for each cancer type based on mRNA abundance values in the matching samples,
606 including silent genes (median zero RPKM-UQ) and five lists of non-silent genes of equal size
607 grouped into 20% bins. For the pan-cancer analysis, we binned genes using median mRNA
608 abundance in the entire RNA-seq dataset.

609 **Grouping CTCF binding sites by cell type specificity.** To analyse CTCF binding sites by their
610 tissue and cell type specificity, we grouped all 162,209 CTCF binding sites of the ENCODE
611 dataset into five equally sized bins based on the number of cell lines where the sites were
612 detected. To interpret these CTCF sites, we retrieved chromatin loops in eight cell lines from a
613 Hi-C study [46], used a ±1,000 bps window around loop anchor midpoints to define narrower
614 versions of loop anchors, and counted the number of CTCF binding sites in each bin overlapping
615 these loop anchors. We used a Fisher's exact test to evaluate the enrichment of CTCF binding
616 sites at loop anchors among the CTCF binding sites with constitutive activity across cell types
617 (i.e., the 5th bin of CTCF sites).

618 **Analysis of localised mutation rates in gene-regulatory and chromatin architectural**
619 **elements.** First, we evaluated the localised mutation rates in CTCF binding sites, TSSs and
620 cancer-specific open-chromatin sites (*i.e.*, TCGA ATAC-seq sites) for the pan-cancer cohort and
621 all cohorts of selected cancer types. Total mutations and mutations grouped by COSMIC
622 signatures, mutation and transcription strand, and reference/alternative allele were analysed.
623 Indel mutations were analysed as part of total mutations and also as a separate group. Results
624 were adjusted for multiple testing using the Benjamini-Hochberg false discovery rate procedure
625 and filtered ($FDR < 0.05$). We also analysed the simulated variant call set using the same
626 pipeline and found no significant results, as expected ($FDR < 0.05$). Results of the systematic
627 analysis were visualised as a dot plot. *FDR* values in the main dot plot were capped at 10^{-32} for
628 visualisation purposes. To visualise localised mutation rates, all sites were pooled, aligned using
629 median coordinates and trimmed to uniform lengths. Coordinates were transformed relative to
630 site midpoint. Upstream and downstream flanking sequences of equal length were also
631 considered. Local regression (loess) curves with the span parameter of 33% were used to
632 visualise a smoothened mutation frequency in sites relative to flanking sequences.

633 **Transcriptomic and functional associations of localised mutation rates.** We evaluated the
634 localised mutation rates in TSSs and open-chromatin sites grouped by mRNA abundance of
635 genes in matching tumor types. Again, the results were adjusted for multiple testing using the
636 Benjamini-Hochberg false discovery rate procedure and filtered ($FDR < 0.05$). To compare
637 different cohorts and subsets of sites, we normalised per-nucleotide mutation counts by dividing
638 these by number of sites in each gene bin, and also by the number of tumors in each cohort.
639 Normalised counts were multiplied by 1e6 to quantify a per-tumor, per-megabase average
640 mutation rate. To study the functional associations of localised mutation rates at CTCF binding
641 sites, we asked whether the extent of conservation of CTCF binding sites in cell types, as
642 observed in ENCODE ChIP-seq experiments, was indicative of the rate of localised mutagenesis
643 at these sites. CTCF binding sites were grouped into five mutually exclusive bins of equal size
644 based on the number of cell types where the sites were observed. Analysis of localised mutation
645 rates in these sites was conducted as described above, findings were corrected for multiple
646 testing correction and filtered to select significant findings ($FDR < 0.05$).

647 **Down-sampling of open-chromatin sites to evaluate mRNA associations.** To evaluate the
648 mRNA associations of mutation rates in open-chromatin sites compared to TSSs, we performed a
649 down-sampling analysis. The analysis was designed to check whether the statistical significance
650 of mRNA associations in open-chromatin sites was systematically amplified due the larger set of
651 open-chromatin sites available for analysis. To this end, RM2 was used to evaluate randomly
652 sampled subsets of open-chromatin sites in all the bins of sites grouped by mRNA abundance.
653 For each bin, we sampled the number of open-chromatin sites that were observed in the
654 equivalent bin of TSSs. The analysis was repeated for 100 random subsets of open-chromatin
655 sites for each bin and median *P*-values and corresponding fold-changes of localised mutation
656 rates were reported. A lenient cut-off was used to filter and visualise results (unadjusted $P < 0.2$).

657 **Identifying pathways with regional mutation rates.** We asked whether the localised mutation
658 rates of TSSs significantly affected genes in specific biological processes and pathways. We
659 repurposed the RM2 model to analyse TSSs of gene sets corresponding to biological processes of
660 Gene Ontology [62] and molecular pathways of the Reactome database [63]. Gene sets were derived
661 from the g:Profiler [64] web server (March 3rd 2020) and subsequently filtered to include 1,871
662 gene sets with 100 to 1,000 genes. Pathway analysis of localised mutation rates was conducted

663 separately for each cancer type, results were corrected for multiple testing using the Benjamini-
664 Hochberg FDR procedure separately for every cancer type and filtered for statistical significance
665 (*FDR* < 0.1). We chose the less stringent significance filter since the mutation rate analysis of
666 functional gene sets was relatively less powered given that fewer sites were considered. The
667 pathways with significant TSS-specific mutation rates were visualised as an enrichment map [65]
668 in Cytoscape and major biological themes were manually curated as described earlier [45]. Nodes
669 in the enrichment map were painted to reflect cancer types where these pathway enrichments
670 were detected following the custom color scheme of the PCAWG project.

671 **Associating regional mutation rates with driver mutations and recurrent copy-number**
672 **alterations.** We asked whether the localised mutation rates in CTCF binding sites, TSSs and
673 open-chromatin sites were associated with driver mutations (*i.e.*, SNVs, indels) or recurrent
674 copy-number alterations (CNAs) in cancer genomes. First we collected a high-confidence set of
675 driver mutations and CNAs in the PCAWG cohort. Driver mutations in exons of protein-coding
676 genes were predicted for each selected cancer type using the ActiveDriverWGS method [34]. We
677 used the PCAWG variant calls after filtering tumors as described above, corrected the results for
678 multiple testing using the Benjamini-Hochberg FDR procedure and selected significant driver
679 genes (*FDR* < 0.05). FDR correction was conducted separately for each cancer type across the
680 pooled set of protein-coding and non-coding genes. Tumors with and without SNVs or indels in
681 predicted driver genes were used for localised mutation rate analysis. Predictions of recurrent
682 CNAs were derived from the pan-cancer dataset of GISTIC2 calls of the PCAWG project [49]. All
683 CNA lesions at 95% confidence scores were considered and amplifications and deletions were
684 analysed separately. High-confidence CNA events were used (GISTIC2 score = 2). Tumors with
685 and without CNAs in the recurrently altered regions as defined by GISTIC2 were used for
686 localised mutation rate analysis. Each cancer type was considered separately. Next we filtered
687 very frequent and infrequent drivers and CNA events to improve the power of the RM2 analysis.
688 We selected the driver genes and CNA regions with at least 25 tumors in the mutated (or copy-
689 number altered) group of tumors and filtered very frequent drivers and CNAs affecting more
690 than 2/3 of the cohort. Each driver gene and recurrent CNA locus in each cancer type was then
691 analysed for associations with localised mutation rates in the three categories of genomic
692 elements (open-chromatin sites, CTCF binding sites, TSSs). The binary co-factor in RM2 was

693 used to indicate the mutated or wildtype status of the tumor with respect to the given recurrent
694 genetic event. We first computed the significance of site-specific localised mutation rates given
695 the presence or absence of driver gene mutations or recurrent CNAs. All combined RM2 results
696 of driver gene mutations, cancer types and genomic sites were adjusted for multiple testing
697 correction using the Benjamini-Hochberg FDR procedure and significant results were selected
698 ($FDR < 0.05$). We then conducted an additional likelihood ratio test to evaluate the significance
699 of the interaction between localised mutation rates and the presence of driver mutations and
700 filtered the results to only include positive and significant interactions (unadjusted $P < 0.05$,
701 main and interaction coefficients > 0). To validate and visualise the detected interactions, we
702 separately analysed individual groups of tumors defined by the presence or absence of driver
703 mutations and CNAs using RM2, and compared the resulting FDR values and fold-changes.

704 **Associating CNAs with mRNA abundance**. To evaluate the functional role of CNAs associated
705 with localised mutation rates, we compared mRNA abundance levels of genes in the CNA loci in
706 groups of tumors defined by the presence or absence of the CNA events, using matching RNA-
707 seq data available in PCAWG [44]. Genes in CNA loci were retrieved from the PCAWG GISTIC2
708 dataset and genes with low mRNA abundance were removed (median FPKM-UQ < 1). mRNA
709 abundance levels of genes in CN amplified and non-amplified (*i.e.*, balanced and deleted) tumors
710 were compared using the nonparametric Wilcoxon test. One-sided tests were used, assuming that
711 change in mRNA abundance would match the underlying copy-number change (*i.e.*, copy
712 number amplifications were tested for increase in mRNA abundance of matching genes). Results
713 were adjusted for multiple testing correction using the Benjamini-Hochberg FDR procedure and
714 significant results were selected ($FDR < 0.05$). To confirm the CNAs, we retrieved the consensus
715 dataset of CNA calls in each tumor from the PCAWG study [49] and visualised the detected CNA
716 segments normalised by tumor ploidy predictions in subsets of tumors defined by the presence or
717 absence of these CNA events. Known cancer genes of the COSMIC Cancer Gene Census
718 database [66] (v91, downloaded 14.05.2020) were identified among the genes with mRNA/CNA
719 associations.

720 **Method benchmarking and power analysis.** We evaluated the performance of our method and
721 statistical power using simulated variant calls, different parametrizations and down-sampling of
722 input datasets. First, to evaluate method calibration and false-positive rates, we performed a

723 systematic analysis of open-chromatin sites, TSSs, and CTCF binding sites in a comparable set
724 of simulated variant calls from PCAWG. This simulated variant set was derived earlier from the
725 same set of tumor genomes using trinucleotide-informed shuffling of mutations [4]. Simulated
726 variant calls were analysed similarly to true variant calls for total mutation counts, reference and
727 alternative nucleotide combinations, predicted mutational signatures and transcription and
728 mutation strand properties. Results from RM2 were adjusted for multiple testing using the
729 Benjamini-Hochberg FDR procedure separately for results derived from true and simulated
730 variant calls. As expected, simulated variant calls revealed no statistically significant results of
731 localised mutation rates in any cancer type, site type or mutation subset ($FDR < 0.05$). We then
732 visualised the distribution of log-transformed P-values derived from true and simulated variant
733 calls using quantile-quantile (QQ) plots and found that highly significant P-values were detected
734 in true datasets while the P-values derived from simulated variant calls were uniformly
735 distributed as expected. These analyses show that our model is well-calibrated and is not subject
736 to inflated false-positive findings. Second, to evaluate the statistical power of RM2, we
737 performed systematic down-sampling by randomly selecting subsets of sites and tumors for
738 localised mutation rate analysis. We focused on the PCAWG liver hepatocellular carcinoma
739 (Liver-HCC) cohort of 300 samples and CTCF sites. A series of down-sampling configurations
740 were used (2000, 5000, …, 100,000 sites sampled; 25, 50, …, 300 genomes sampled). Each
741 configuration was tested 100 times in different subsets of data. For a power analysis, we
742 evaluated the fraction of runs that revealed a significant enrichment of somatic mutations at
743 CTCF sites ($P < 0.05$) and the median P-value of these 100 runs. Third, we evaluated the
744 parameter values of RM2 that determine the genomic width of each site and the control regions
745 of upstream and downstream flanking sequences. As expected, site-specific mutation rates were
746 consistently identified at multiple values of the width parameter for each class of site (open-
747 chromatin sites, CTCF binding sites, TSSs), indicating robustness of our analysis to different
748 parameter values. However, different site classes showed preferences towards shorter sites
749 (CTCF binding sites: 20-100 bps) or longer sites (open-chromatin sites and TSS: 200-800 bps),
750 likely due to different underlying mutational processes. For the entire study, the optimal genomic
751 size of every site class was selected based on the strongest effect size and significance across
752 multiple cancer types. The value of 50 bps (±25 bps) was selected for CTCF sites. For open-
753 chromatin sites and TSSs, we selected the common site width of 200 bps (±100 bps) that showed

754 strong effects in both TSSs and open-chromatin sites, to increase comparability of the two
755 classes.

756

757 **Acknowledgments.** This work was supported by the Canadian Institutes of Health Research
758 (CIHR) Project Grant to J.R. and the Investigator Award to J.R. from the Ontario Institute for
759 Cancer Research (OICR). C.A.L. partially was supported by a Graduate Student Fellowship from
760 the Department of Medical Biophysics, University of Toronto. Funding to OICR is provided by
761 the Government of Ontario.

762

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

767

768 **References**

769 1 Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719-
770 724, doi:10.1038/nature07943 (2009).
771 2 Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558,
772 doi:10.1126/science.1235122 (2013).
773 3 ICGC-TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis
774 of whole genomes. *Nature* **578**, 82-93, doi:10.1038/s41586-020-1969-6 (2020).
775 4 Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,693 cancer whole
776 genomes. *Nature* **in press** (2020).
777 5 Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and
778 Mutations. *Cell* **174**, 1034-1035, doi:10.1016/j.cell.2018.07.034 (2018).
779 6 Kumar, S. *et al.* Passenger Mutations in More Than 2,500 Cancer Genomes: Overall
780 Molecular Functional Impact and Consequences. *Cell* **180**, 915-927 e916,
781 doi:10.1016/j.cell.2020.01.032 (2020).
782 7 Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **in press** (2020).
783 8 Supek, F. & Lehner, B. Scales and mechanisms of somatic mutation rate variation across
784 the human genome. *DNA Repair (Amst)*, 102647, doi:10.1016/j.dnarep.2019.102647
785 (2019).
786 9 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-
787 associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).
788 10 Reijns, M. A. M. *et al.* Lagging-strand replication shapes the mutational landscape of the
789 genome. *Nature* **518**, 502-506, doi:10.1038/nature14183 (2015).
790 11 Schuster-Bockler, B. & Lehner, B. Chromatin organization is a major influence on
791 regional mutation rates in human cancer cells. *Nature* **488**, 504-507,
792 doi:10.1038/nature11273 (2012).
793 12 Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of
794 cancer. *Nature* **518**, 360-364, doi:10.1038/nature14221 (2015).
795 13 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**,
796 415-421, doi:10.1038/nature12477 (2013).
797 14 Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature*
798 **578**, 94-101, doi:10.1038/s41586-020-1943-3 (2020).
799 15 Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents.
800 *Cell* **177**, 821-836 e816, doi:10.1016/j.cell.2019.03.001 (2019).
801 16 Pich, O. *et al.* The mutational footprints of cancer therapies. *Nature genetics* **51**, 1732-
802 1740, doi:10.1038/s41588-019-0525-5 (2019).
803 17 Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature*
804 **575**, 210-216, doi:10.1038/s41586-019-1689-y (2019).
805 18 Kübler, K. *et al.* Tumor mutational landscape is a record of the pre-malignant state.
806 *bioRxiv* **517565**, doi:https://doi.org/10.1101/517565 (2019).
807 19 Jiao, W. *et al.* A deep learning system accurately classifies primary and metastatic
808 cancers using passenger mutation patterns. *Nat Commun* **11**, 728, doi:10.1038/s41467-
809 019-13825-8 (2020).
810 20 Gonzalez-Perez, A., Sabarinathan, R. & Lopez-Bigas, N. Local Determinants of the
811 Mutational Landscape of the Human Genome. *Cell* **177**, 101-114,
812 doi:10.1016/j.cell.2019.02.051 (2019).

813 21 Kaiser, V. B., Taylor, M. S. & Semple, C. A. Mutational Biases Drive Elevated Rates of
814 Substitution at Regulatory Sites across Cancer Types. *PLoS Genet* **12**, e1006207,
815 doi:10.1371/journal.pgen.1006207 (2016).
816 22 Yazdi, P. G. *et al.* Increasing Nucleosome Occupancy Is Correlated with an Increasing
817 Mutation Rate so Long as DNA Repair Machinery Is Intact. *PLoS One* **10**, e0136574,
818 doi:10.1371/journal.pone.0136574 (2015).
819 23 Hara, R., Mo, J. & Sancar, A. DNA damage in the nucleosome core is refractory to repair
820 by human excision nuclease. *Mol Cell Biol* **20**, 9173-9181, doi:10.1128/mcb.20.24.9173-
821 9181.2000 (2000).
822 24 Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & Lopez-Bigas, N.
823 Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature*
824 **532**, 264-267, doi:10.1038/nature17661 (2016).
825 25 Perera, D. *et al.* Differential DNA repair underlies mutation hotspots at active promoters
826 in cancer genomes. *Nature* **532**, 259-263, doi:10.1038/nature17437 (2016).
827 26 Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome
828 neighborhoods. *Science* **351**, 1454-1458, doi:10.1126/science.aad9024 (2016).
829 27 Katainen, R. *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nature*
830 *genetics* **47**, 818-821, doi:10.1038/ng.3335 (2015).
831 28 Guo, Y. A. *et al.* Mutation hotspots at CTCF binding sites coupled to chromosomal
832 instability in gastrointestinal cancers. *Nat Commun* **9**, 1520, doi:10.1038/s41467-018-
833 03828-2 (2018).
834 29 Polak, P. *et al.* Reduced local mutation density in regulatory DNA of cancer genomes is
835 linked to DNA repair. *Nature biotechnology* **32**, 71-75, doi:10.1038/nbt.2778 (2014).
836 30 Frigola, J. *et al.* Reduced mutation rate in exons due to differential mismatch repair.
837 *Nature genetics* **49**, 1684-1692, doi:10.1038/ng.3991 (2017).
838 31 Bell, R. J. *et al.* The transcription factor GABP selectively binds and activates the mutant
839 TERT promoter in cancer. *Science* **348**, 1036-1039, doi:10.1126/science.aab0015 (2015).
840 32 Huang, F. W. *et al.* Highly recurrent TERT promoter mutations in human melanoma.
841 *Science* **339**, 957-959, doi:10.1126/science.1229259 (2013).
842 33 Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding
843 somatic mutations and gene expression alterations across 14 tumor types. *Nature genetics*
844 **46**, 1258-1263, doi:10.1038/ng.3141 (2014).
845 34 Zhu, H. *et al.* Candidate Cancer Driver Mutations in Distal Regulatory Elements and
846 Long-Range Chromatin Interaction Networks. *Mol Cell*,
847 doi:10.1016/j.molcel.2019.12.027 (2020).
848 35 Corona, R. I. *et al.* Non-coding somatic mutations converge on the PAX8 pathway in
849 ovarian cancer. *Nat Commun* **11**, 2020, doi:10.1038/s41467-020-15951-0 (2020).
850 36 Liu, E. M. *et al.* Identification of Cancer Drivers at CTCF Insulators in 1,962 Whole
851 Genomes. *Cell Syst* **8**, 446-455 e448, doi:10.1016/j.cels.2019.04.001 (2019).
852 37 Reyna, M. A. *et al.* Pathway and network analysis of more than 2,500 whole cancer
853 genomes. *Nature Communications* **in press** (2020).
854 38 Mazrooei, P. *et al.* Cistrome Partitioning Reveals Convergence of Somatic Mutations and
855 Risk Variants on Master Transcription Regulators in Primary Prostate Tumors. *Cancer*
856 *Cell* **36**, 674-689 e676, doi:10.1016/j.ccell.2019.10.005 (2019).
857 39 Encode Project Consortium. An integrated encyclopedia of DNA elements in the human
858 genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

859 40 Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers.
860 *Science* **362**, doi:10.1126/science.aav1898 (2018).
861 41 Dvorak, K. *et al.* Bile acids in combination with low pH induce oxidative stress and
862 oxidative DNA damage: relevance to the pathogenesis of Barrett's oesophagus. *Gut* **56**,
863 763-771, doi:10.1136/gut.2006.103697 (2007).
864 42 Tomkova, M., Tomek, J., Kriaucionis, S. & Schuster-Bockler, B. Mutational signature
865 distribution varies with DNA replication timing and strand asymmetry. *Genome Biol* **19**,
866 129, doi:10.1186/s13059-018-1509-y (2018).
867 43 Dellino, G. I. *et al.* Release of paused RNA polymerase II at specific loci favors DNA
868 double-strand-break formation and promotes cancer translocations. *Nature genetics* **51**,
869 1011-1023, doi:10.1038/s41588-019-0421-z (2019).
870 44 PCAWG Transcriptome Core Group *et al.* Genomic basis of RNA alterations in cancer.
871 *Nature* **in press** (2020).
872 45 Reimand, J. *et al.* Pathway enrichment analysis and visualization of omics data using
873 g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protoc* **14**, 482-517,
874 doi:10.1038/s41596-018-0103-9 (2019).
875 46 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles
876 of chromatin looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).
877 47 Wendt, K. S. *et al.* Cohesin mediates transcriptional insulation by CCCTC-binding factor.
878 *Nature* **451**, 796-801, doi:10.1038/nature06634 (2008).
879 48 Khoury, A. *et al.* Constitutively bound CTCF sites maintain 3D chromatin architecture
880 and long-range epigenetically regulated domains. *Nat Commun* **11**, 54,
881 doi:10.1038/s41467-019-13753-7 (2020).
882 49 Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **in**
883 **press** (2020).
884 50 The Cancer Genome Atlas Network. Genomic Classification of Cutaneous Melanoma.
885 *Cell* **161**, 1681-1696, doi:10.1016/j.cell.2015.05.044 (2015).
886 51 Chapman, P. B. *et al.* Improved survival with vemurafenib in melanoma with BRAF
887 V600E mutation. *N Engl J Med* **364**, 2507-2516, doi:10.1056/NEJMoa1103782 (2011).
888 52 Sheu, J. J. *et al.* Mutant BRAF induces DNA strand breaks, activates DNA damage
889 response pathway, and up-regulates glucose transporter-1 in nontransformed epithelial
890 cells. *Am J Pathol* **180**, 1179-1188, doi:10.1016/j.ajpath.2011.11.026 (2012).
891 53 Xu, H. *et al.* Enhanced RAD21 cohesin expression confers poor prognosis and resistance
892 to chemotherapy in high grade luminal, basal and HER2 breast cancers. *Breast Cancer*
893 *Res* **13**, R9, doi:10.1186/bcr2814 (2011).
894 54 Meisenberg, C. *et al.* Repression of Transcription at DNA Breaks Requires Cohesin
895 throughout Interphase and Prevents Genome Instability. *Mol Cell* **73**, 212-223 e217,
896 doi:10.1016/j.molcel.2018.11.001 (2019).
897 55 Kagey, M. H. *et al.* Mediator and cohesin connect gene expression and chromatin
898 architecture. *Nature* **467**, 430-435, doi:10.1038/nature09380 (2010).
899 56 Dahia, P. L. Pheochromocytoma and paraganglioma pathogenesis: learning from genetic
900 heterogeneity. *Nature reviews. Cancer* **14**, 108-119, doi:10.1038/nrc3648 (2014).
901 57 Burnichon, N. *et al.* SDHA is a tumor suppressor gene causing paraganglioma. *Hum Mol*
902 *Genet* **19**, 3011-3020, doi:10.1093/hmg/ddq206 (2010).
903 58 Guzy, R. D., Sharma, B., Bell, E., Chandel, N. S. & Schumacker, P. T. Loss of the SdhB,
904 but Not the SdhA, subunit of complex II triggers reactive oxygen species-dependent

905 hypoxia-inducible factor activation and tumorigenesis. *Mol Cell Biol* **28**, 718-731,
906 doi:10.1128/MCB.01338-07 (2008).
907 59 Ishii, T. *et al.* A mutation in the SDHC gene of complex II increases oxidative stress,
908 resulting in apoptosis and tumorigenesis. *Cancer Res* **65**, 203-209 (2005).
909 60 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in
910 141,456 humans. *Nature* **581**, 434-443, doi:10.1038/s41586-020-2308-7 (2020).
911 61 Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of
912 somatic mutations in normal human skin. *Science* **348**, 880-886,
913 doi:10.1126/science.aaa6806 (2015).
914 62 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene
915 Ontology Consortium. *Nature genetics* **25**, 25-29, doi:10.1038/75556 (2000).
916 63 Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res* **46**, D649-
917 D655, doi:10.1093/nar/gkx1132 (2018).
918 64 Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler--a web-based toolset
919 for functional profiling of gene lists from large-scale experiments. *Nucleic acids research*
920 **35**, W193-200, doi:10.1093/nar/gkm226 (2007).
921 65 Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a
922 network-based method for gene-set enrichment visualization and interpretation. *PloS one*
923 **5**, e13984, doi:10.1371/journal.pone.0013984 (2010).
924 66 Futreal, P. A. *et al.* A census of human cancer genes. *Nat Rev Cancer* **4**, 177-183,
925 doi:10.1038/nrc1299 (2004).

926