

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 selective 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 research³⁻⁵, 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 resolutions⁸. 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 genes⁹⁻¹². At a single base pair resolution, certain trinucleotides are preferentially mutated through processes of carcinogen exposures, defective DNA repair pathways, and aberrant DNA replication¹³⁻¹⁵. 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 genomes²¹⁻²³. 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 regions²⁹ and protein-coding exons³⁰ 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

hotspot in the *TERT* promoter creates a TFBS of the ETS TF family that leads to constitutive activation of *TERT* and enables replicative immortality of cancer cells³¹⁻³³. Recent studies have catalogued candidate non-coding driver elements in gene regulatory and chromatin architectural regions of the cancer genome with functional validations of novel elements³⁴⁻³⁶ and highlighted the convergence of non-coding mutations on molecular pathways and regulatory networks^{37,38}. Thus we need to characterise localised mutational processes to understand the evolution of the somatic genome and the effects of carcinogens and endogenous mutational processes, but also to evaluate the effects of positive selection in the non-coding genome. However, few dedicated computational methods exist to analyse mutation rates at the local resolution of the genome. As a result, there is a lack of large-scale analyses of the local mutation landscape in pan-cancer WGS datasets, leaving the genetic and environmental determinants poorly understood.

Here we developed a new statistical framework that quantifies the activity of mutational processes and signatures on specific classes of non-coding elements of the cancer genome. Our model considers local sequence context, megabase-level somatic mutation rates and genetic covariates to control for variation at the trinucleotide and megabase resolution while isolating site-level effects. We performed a systematic analysis of local mutation rate variation in three classes of gene-regulatory and chromatin architectural genomic elements across 2,500 whole cancer genomes of the ICGC/TCGA Pan-cancer Analysis of Whole Genomes (PCAWG) project³. We found a pervasive mutation enrichment at these functional non-coding elements that was characterised by specific mutational signatures and transcriptional and pathway-level activities of these elements in select cancer types. We detected statistical interactions of local mutagenesis and recurrent genomic alterations that suggest potential genetic mechanisms driving the underlying mutational processes. Our computational framework and systematic analysis reveals the diversity of mutational processes in functional non-coding elements of the cancer genome and their roles in somatic genome evolution, drivers of cancer phenotypes and molecular 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:

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

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

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

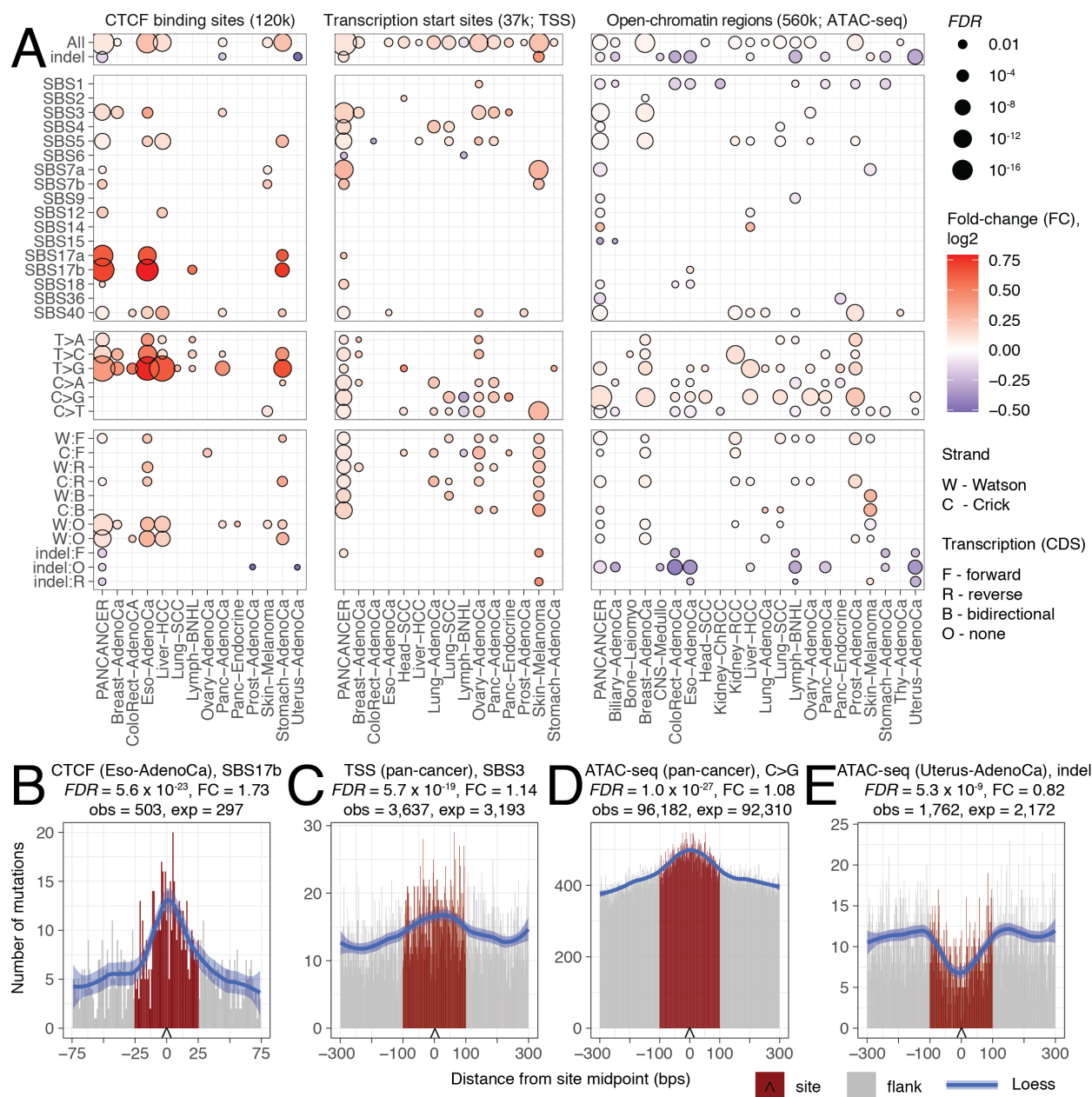


Figure 1. Localised mutation rates in gene-regulatory and chromatin architectural elements of cancer genomes. **A.** Comparison of mutation rates in DNA-binding sites of the CTCF chromatin architectural factor (left), transcription start sites (TSS) and cancer-specific open-chromatin regions (TCGA ATAC-seq) 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.

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

To study localised mutation rates in gene-regulatory and chromatin architectural elements of the genome, we analysed the pan-cancer dataset of 2,514 whole cancer genomes of the PCAWG project³. We individually analysed 25/35 cancer types with at least 25 samples as well as the pan-cancer set containing all tumors of the 35 types (**Supplementary Figure 3A**). 69 hypermutated tumors were excluded to avoid confounding effects. Three classes of genomic elements were analysed: 119,464 CTCF binding sites conserved in at least two cell lines in the ENCODE project³⁹, 37,309 transcription start sites (TSS) of protein-coding genes from the Ensembl database (GRCh37), and a pan-cancer set of 561,057 open-chromatin sites detected across 410 primary tumors of The Cancer Genome Atlas (TCGA) project (*i.e.*, ATAC-seq sites)⁴⁰ (**Supplementary Table 1A**). In total, the analysis included 640,023 unique loci representing 3.9% (120.5 Mbps) of the human genome. In addition to total mutations, we grouped mutations by DNA strand (Watson or Crick), transcription status (forward, reverse, bidirectional or absent), reference and alternative nucleotide pairs, and COSMIC mutational signatures of single base substitutions (SBS) inferred in the PCAWG study¹⁴. Indel mutations were pooled with SNVs and also analysed separately. To obtain a conservative analysis, we excluded a small fraction of tumors as outliers (31 or 1.3%) where even single-sample RM2 analysis revealed highly significant differences in local mutation rates in any of the three classes of sites ($FDR < 0.001$) (**Supplementary Figure 3B**). The final pan-cancer analysis studied 23.0 million mutations including 1.61 million indels detected in 2,421 genomes of 35 cancer types.

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

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

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

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

Pan-cancer open chromatin regions defined as ATAC-seq profiles of primary tumors were also enriched in mutations in 11/25 cancer types and the pan-cancer cohort, especially among C>G mutations (*e.g.*, pan-cancer, $FDR = 1.0 \times 10^{-27}$, $FC = 1.08$) (**Figure 1D**). However, the effect sizes of mutational enrichments were more modest compared to CTCF binding sites and TSSs, potentially due to mixed effects of SNVs and indels: mutation enrichments in open-chromatin sites were primarily driven by SNVs, while in contrast, indel mutations were significantly depleted in the pan-cancer cohort and in 9/25 cancer types, indicating that the open chromatin environment or the binding of regulatory elements may be protective of the mutational processes responsible for generating indels. For example, in uterine adenocarcinoma, 1,762 indel mutations were observed in open-chromatin sites while 2,172 were expected according to RM2 ($FDR = 5.3 \times 10^{-9}$, $FC = 0.82$) (**Figure 1E**). This indel depletion appeared relatively stronger in intergenic sites of open chromatin. Further, mutational signature analysis indicated reduced activity of the aging-related signature SBS1 in open-chromatin sites, apparent in eight cancer types and the pan-cancer cohort (*e.g.*, colorectal adenocarcinoma, $FDR = 3.8 \times 10^{-5}$, $FC = 0.90$). However, analysis of these pan-cancer open chromatin regions is better powered compared to the analysis of TSS loci due to a larger number of sites, thus smaller deviations in local mutation rates were detectable. Our findings contrast an earlier report that indicated broadly decreased mutation rates at chromatin-accessible regulatory elements derived from cell lines²⁹. Comparison of localised mutation rates at TSS loci and open-chromatin sites indicate distinct properties of localised mutagenesis acting on proximal and distal regulatory elements of the genome.

We extended this analysis to benchmark our model and evaluate statistical power. First, to assess model calibration and false positive rates, we analysed a PCAWG dataset of simulated variant calls designed to approximate neutral genome evolution⁴. As expected, analysis of simulated data did not reveal any significant differences in mutation rates in the three classes of elements (all $FDR > 0.05$). Quantile-quantile analysis of *P*-values confirmed that the model is well calibrated for true and simulated mutations (**Supplementary Figure 2A**). Second, we evaluated

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

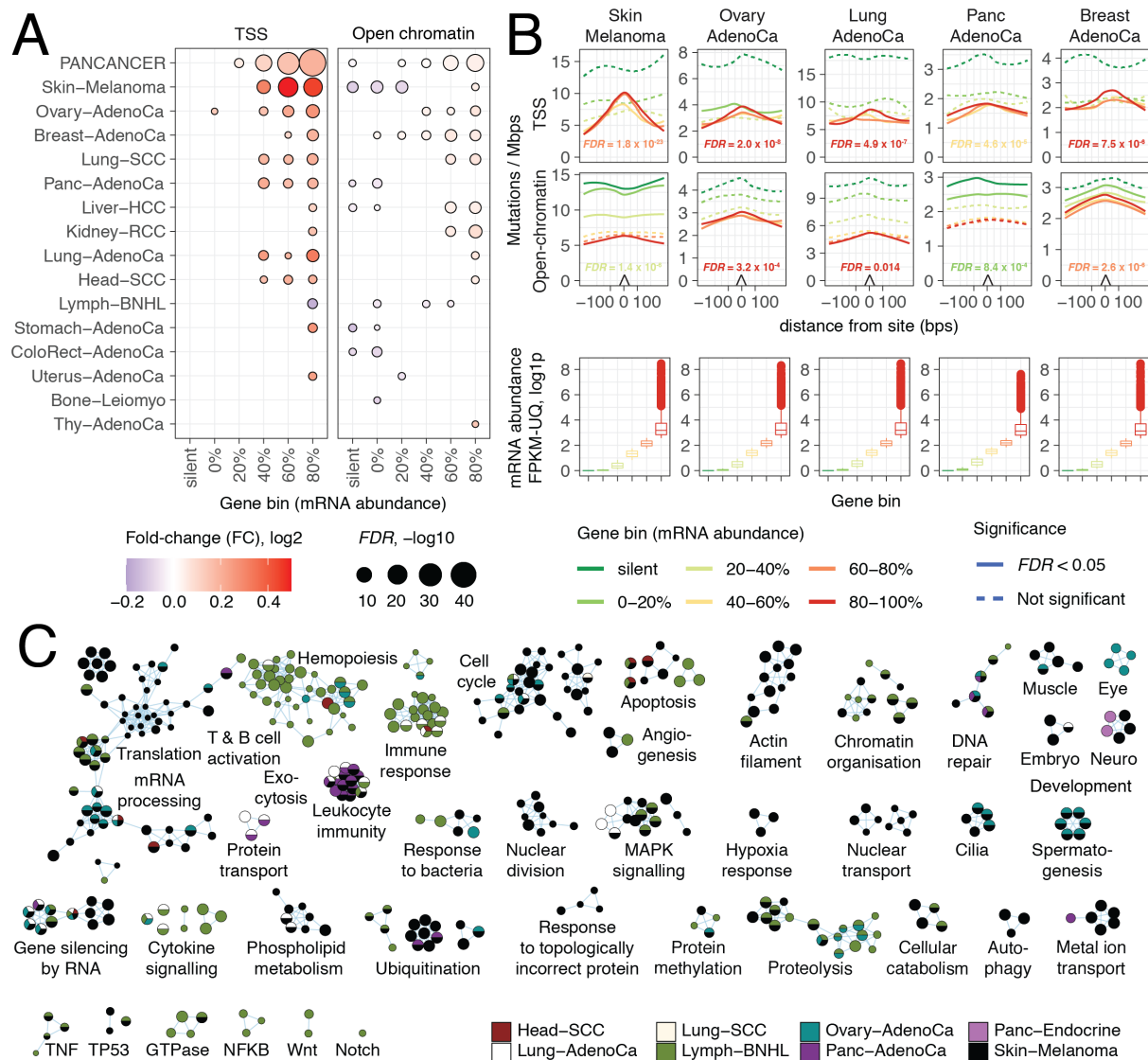


Figure 2. Increased mutation rates at transcription start sites (TSS) of highly expressed genes. **A.** Comparison of localised mutation rates in TSSs and open-chromatin sites (TCGA ATAC-seq), grouped by mRNA abundance of target genes in matching tumors ($FDR < 0.05$). **B.** Examples of cancer types 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.

We asked whether the increased mutational load at TSSs and open-chromatin sites correlated with transcriptional activity of target genes. To precisely quantify mRNA abundance in different cancer types, we used matching RNA-seq data in the PCAWG project⁴⁴. TSS-specific mutation rates were studied in six bins of genes grouped by median mRNA abundance in 19/25 cancer types. TSS-specific mutation rates were significantly increased in highly expressed genes in eleven cancer types ($FDR < 0.05$) and the pan-cancer cohort ($FDR = 2.2 \times 10^{-50}$, $FC = 1.16$) (**Figure 2A**) (**Supplementary Table 1B**). Association of transcriptional activity and TSS-specific mutation rates was the strongest in melanoma and ovarian, lung, pancreatic, and breast adenocarcinoma where genes with the highest mRNA abundance were strongly enriched in TSS-specific mutations ($FDR \leq 10^{-4}$, $FC \geq 1.15$) (**Figure 2B**). Of note, the top gene bin was highly variable in mRNA abundance (e.g., 11-7,100 RPKM-UQ in the pan-cancer cohort). In silenced genes, TSS loci consistently showed no significant differences in mutation rates relative to flanking controls ($FDR > 0.05$), however the overall mutation rate in sites and controls was higher, potentially a result of lower rates of transcription coupled repair in closed chromatin. This analysis highlights a localised mutational process in gene promoters apparent in multiple cancer types that is potentially driven by transcriptional initiation or TF binding at core gene-regulatory promoter elements.

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

To understand the functional associations of TSS mutations, we performed a pathway enrichment analysis by adapting RM2 to gene sets of GO biological processes and Reactome molecular pathways⁴⁵. This allowed us to test our hypothesis that promoters enriched for non-coding mutations are concentrated in specific biological processes. We found 546 unique pathways and processes exhibiting elevated TSS-specific mutation rates ($FDR < 0.1$) (**Figure 2C**) (**Supplementary Table 1C**), the majority of which were found in the melanoma cohort (70%) while 28% of pathways were found in more than one cancer type. Translation, ribosome biogenesis and RNA processing were among the largest groups of pathways with increased TSS-specific mutation rates. This is expected as the translational machinery is ubiquitously active in proliferating cells and includes many highly expressed genes. Cancer-related processes and pathways were also enriched in TSS mutations, for example mitotic cell cycle, apoptosis, DNA repair, angiogenesis, developmental and immune response processes as well as druggable signalling pathways (*e.g.*, MAPK, Wnt, Notch) were identified in multiple cancer types. The pathway analysis augments our observation of frequent TSS mutations associating with increased transcription and highlights a variety of core cellular processes with high baseline transcription and associated localised mutagenesis active in many cell types. Furthermore, the significant enrichment of TSS-specific mutations in genes of cancer-related processes suggests that some more frequent non-coding mutations at individual promoters are functional and may contribute to cancer driver mechanisms and tumor heterogeneity by altering gene-regulatory circuits in molecular pathways and interaction networks³⁷.

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

showed no significant enrichment of mutations ($FDR > 0.05$). Additional cancer types with frequent CTCF binding site mutations were uncovered in this more focused analysis, including colorectal and pancreatic cancer and lymphoma (BNHL) ($FDR \leq 10^{-6}$, $FC \geq 1.18$). This extends the established pattern of CTCF mutation enrichment to a subset of pan-tissue genomic elements and highlights the utility of RM2 in detecting mutational patterns. Since CTCF is known as the master regulator of genome architecture, we incorporated a HiC dataset of chromatin long-range interactions⁴⁶ to interpret the mutation enrichments at highly conserved CTCF binding sites. We found that these constitutively bound CTCF binding sites were strongly enriched in chromatin loop anchor elements: 30% (9,393/32,442) of the sites of bin five were located within 1 kbps of anchor midpoints (3,973 or 12% expected, Fisher's exact $P = 0$). Contrarily, the majority of CTCF binding sites that were only detected in a few cell types showed no deviation from expected mutation rates and an expected distribution with respect to chromatin loop anchors. This indicates the mutational process primarily affects the subset of CTCF sites that are constitutively bound in most cell types and participate in chromatin architectural and gene regulatory interactions^{46,47}. Conservation is a property of functionally integral CTCF binding sites, which upon disruption, can lead to changes in underlying chromatin architecture and gene regulation⁴⁸ and is associated with activation of proto-oncogenes²⁶.

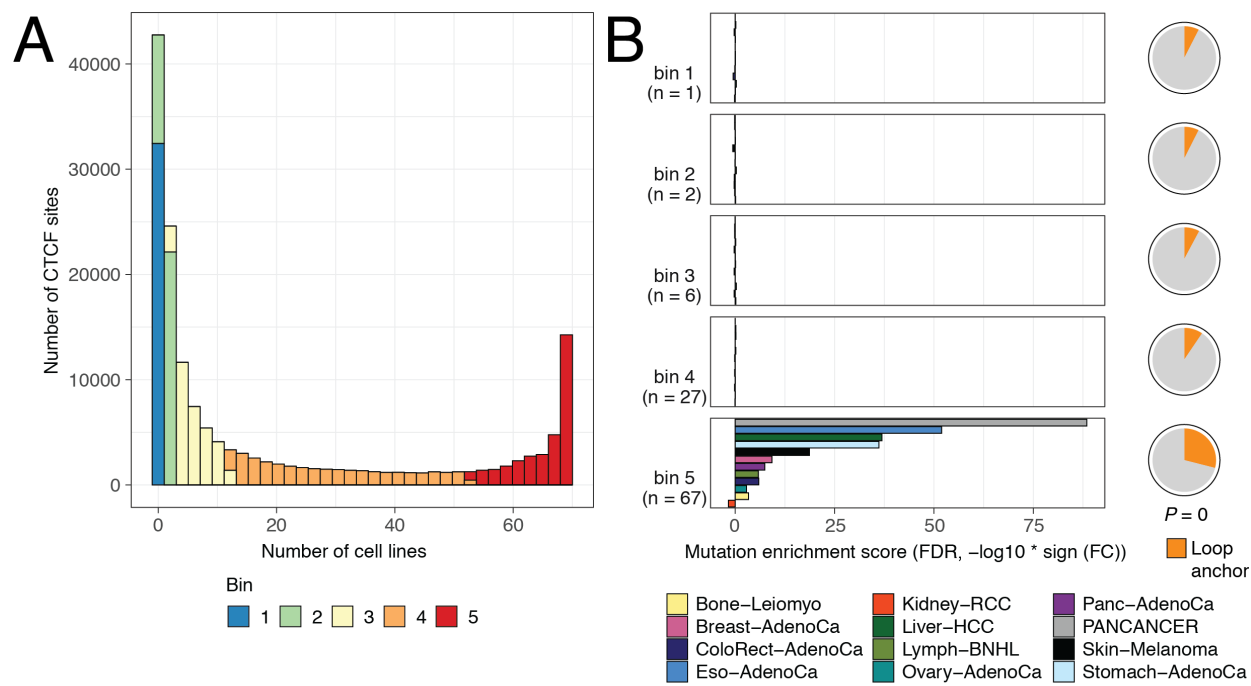


Figure 3. Localised mutation rates at constitutively active binding sites of CTCF. **A.** Histogram of CTCF binding sites with respect to the number of cell lines with sites observed. CTCF binding sites were grouped into five equal bins based on site activity across the cell lines in ENCODE. The bimodal 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

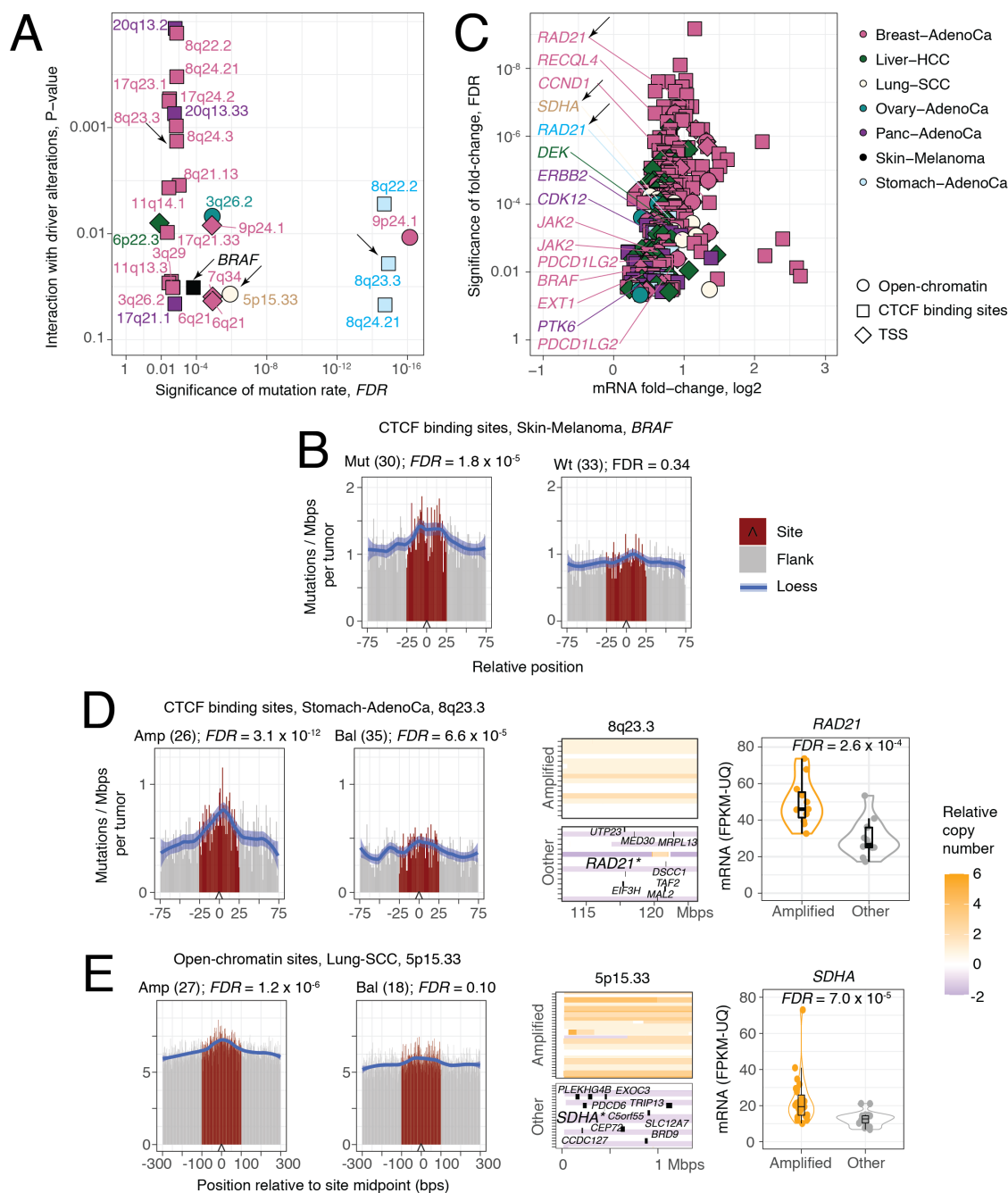


Figure 4. Driver mutations and amplifications associated with gene-regulatory and architectural mutation rates. **A.** Scatter plot shows the overall significance of localised mutations (X-axis; $FDR < 0.05$) and their statistical interactions with drivers and CNAs (Y-axis; $P < 0.05$). Arrows highlight 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*.

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

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

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

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

Amplification and transcriptional up-regulation of *RAD21* located at 8q23.3 was associated with increased mutation rates in CTCF binding sites in stomach and breast cancers. *RAD21* encodes a subunit of the cohesin complex that co-binds DNA with CTCF to orchestrate transcriptional insulation and chromatin architectural interactions^{46,47}. Stomach cancer genomes with amplifications of 8q23.3 showed a strong increase in mutations in CTCF binding sites ($FDR = 3.1 \times 10^{-12}$, $FC = 1.24$) while the increase was less significant in tumors with no amplification ($FDR = 6.6 \times 10^{-5}$, $FC = 1.13$; interaction $P = 0.020$) (**Figure 4D**). mRNA abundance of *RAD21* in 8q23.3-amplified stomach cancers was significantly higher compared to non-amplified samples ($FDR = 2.6 \times 10^{-4}$, 46 vs. 27 median FPKM-UQ), suggesting that *RAD21* expression is driven by the genomic amplification. This association was also observed in breast cancer where the genomes with 8q23.3 amplifications showed an increased mutation rate in CTCF binding sites ($FDR = 5.3 \times 10^{-5}$, $FC = 1.08$) that was not apparent in tumors lacking 8q23.3 amplifications ($FDR = 0.36$, $FC = 0.98$; interaction $P = 5.0 \times 10^{-4}$). *RAD21* amplification was also associated with increased mRNA abundance in breast cancer ($FDR = 1.4 \times 10^{-7}$, 108 vs. 50 median FPKM-UQ), which is also indicative of poor prognosis⁵³. In the ENCODE dataset, 51% of CTCF binding sites were also co-bound by cohesin, representing 94% of the high-confidence *RAD21* binding sites in ENCODE (60,636 / 64,483), significantly more than expected by chance (666 expected, binomial $P = 0$). Mutations in another cohesin subunit, *STAG2*, have been associated with specific mutational signatures and altered transcription at double strand break

sites⁵⁴. In addition to *RAD21*, components of the general transcriptional machinery were amplified in 8q23.3 and showed increased mRNA abundance in breast and stomach cancers, potentially contributing to mutation rates in CTCF binding sites. These included *MED30*, which encodes a subunit of the Mediator transcriptional coactivation complex that interacts with cohesin to connect promoters and enhancers⁵⁵, as well as the transcription initiation factor subunit encoded by *TAF2*. Alternative explanations to CTCF binding site mutations were also apparent in our data. For example, interactions with the amplification of 8q24.21 encoding the *MYC* oncogene were identified in both breast and stomach cancer, however these amplification events did not associate with *MYC* up-regulation. This analysis suggests that the elevated mutagenesis at CTCF binding sites may be driven by the genomic amplification and up-regulation of core transcriptional and genome architectural machinery interacting with CTCF.

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

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

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

Discussion

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

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

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

Our study enables a number of future developments. Integrative analysis of whole cancer genome sequences and rich clinical and pathological profiles of tumors¹⁷ may highlight associations of clinical variables and localised mutagenesis and lead to the discovery of novel WGS-based biomarkers. Considering patient lifestyle information, environmental exposures and germline variation in the analysis may elucidate the impact of carcinogens and endogenous DNA repair deficiencies. Our catalogue of genetic associations provides hypotheses on the mutational mechanisms that can be tested experimentally using genome editing and mutagenesis assays. Rare germline variants in the human population⁶⁰, *de novo* variants detected in genetic disorders²¹ and the widespread somatic genome variation found in healthy tissues⁶¹ provide further 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

Methods

Regression models for Regionalised 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 non-coding 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 of equal width to sites such that the upstream and downstream regions combined were twice as wide as the sites. To account for megabase-scale variation in mutation rates, 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 sites and flanking sequences were distinguished by a binary cofactor (*isSite*). To avoid inflated counts, mutations affecting more than one site were counted once. Likewise, mutations affecting the flanking sequence of more than one site were also counted once. 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 covariates of localised mutation rates. To evaluate the significance of localised

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

$$H_{null}: \text{NegBin}(nMut \sim \text{offset}(\log(nPosits)) + triNucMutClass + \log1p(MbpRate) + coFac).$$

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

$$H_{alt}: \text{NegBin}(nMut \sim \text{offset}(\log(nPosits)) + triNucMutClass + \log1p(MbpRate) + coFac + isSite).$$

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

$$H_{cof}: \text{NegBin}(nMut \sim \text{offset}(\log(nPosits)) + triNucMutClass + \log1p(MbpRate) + coFac + isSite + isSite:coFac).$$

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

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

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

Mutation features and signatures. In addition to evaluating total mutations, several classes of mutations were analysed separately. Mutations were mapped to C and T nucleotides and grouped by reference and alternative nucleotides (C>[A,G,T], T>[A,C,G]). Four additional classifications were developed. First, mutations were classified as located either on the Watson (w) strand if the original reference nucleotide was C or T, or the Crick (c) strand if the original reference nucleotide was A or G. Second, transcriptional activity and orientation of the mutated nucleotides was mapped based on the coordinates of protein-coding genes defined in the Ensembl database (GRCh37) using 500 bps flanking sequence at both ends of genes to account for transcriptional initiation and termination. We then classified mutations as forward-transcribed (F), reverse-transcribed (R), bidirectionally transcribed (B), or not transcribed (O). This initial classification did not include information on tissue-specific transcription and was augmented using matching tumor-specific mRNA abundance data, as described below. Third, mutation strand and transcription status were combined into eight categories (w_[F,R,B,O] and c_[F,R,B,O]). Fourth, we classified mutations by the trinucleotide signatures of single base substitutions (SBS) that were derived earlier using the SigProfiler software in the PCAWG project¹⁴. We assigned each mutation to its most probable signature in the given patient tumor based on its trinucleotide context. For model evaluation, these five major categories of mutations were also derived for the dataset of simulated variant calls.

Chromatin architectural and gene-regulatory genomic elements. We performed a systematic analysis of three classes of genomic elements: DNA-binding sites of CTCF (CCCTC-binding factor) detected in multiple human cell lines, transcription start sites (TSS) of protein-coding genes, and open-chromatin sites (ATAC-seq sites) detected in human primary tumors. CTCF binding sites were retrieved from the ENCODE project³⁹. Sites observed in only one cell line were removed, resulting in 119,464 sites across 70 cell lines. TSS loci of protein-coding genes were retrieved from Ensembl Biomart (GRCh37) and filtered based on location of standard chromosomes (1-22, X, Y), resulting in 37,309 TSSs of 18,710 genes. Open-chromatin sites of 410 primary tumors defined by ATAC-seq were retrieved from the TCGA study⁴⁰. We used the pan-cancer set of sites in the GRCh37 genome as defined in the study and filtered sites on non-standard chromosomes and those lacking defined coordinates in GRCh37. This resulted in 561,057 open-chromatin sites. For the mRNA-based analysis of mRNA abundance described

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

Grouping gene-regulatory sites by mRNA abundance. TSS and open-chromatin sites were analysed in groups based on the mRNA abundance of associated genes in matching tumors. TSS target genes were retrieved from the Ensembl database and target genes of open-chromatin sites were retrieved from the original TCGA study. This analysis was carried out in 19 cohorts of cancer types with at least 20 tumor samples with mRNA and WGS data, as well as the pan-cancer cohort of all cancer types. We used the uniformly processed PCAWG RNA-seq dataset⁴⁴ (RPKM-UQ) and applied the same filtering of tumor samples described previously and excluded non-coding genes. Additionally, we discarded a subset of genes with duplicated HGNC symbols as well as the genes for which TSS or open-chromatin sites were not mapped. This resulted in mRNA measurements for 20,042 protein-coding genes in 1,267 tumor transcriptomes. Next we

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

Grouping CTCF binding sites by cell type specificity. To analyse CTCF binding sites by their tissue and cell type specificity, we grouped all 162,209 CTCF binding sites of the ENCODE dataset into five equally sized bins based on the number of cell lines where the sites were detected. To interpret these CTCF sites, we retrieved chromatin loops in eight cell lines from a Hi-C study⁴⁶, used a $\pm 1,000$ bps window around loop anchor midpoints to define narrower versions of loop anchors, and counted the number of CTCF binding sites in each bin overlapping these loop anchors. We used a Fisher's exact test to evaluate the enrichment of CTCF binding sites at loop anchors among the CTCF binding sites with constitutive activity across cell types (i.e., the 5th bin of CTCF sites).

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

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

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

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

separately for each cancer type, results were corrected for multiple testing using the Benjamini-Hochberg FDR procedure separately for every cancer type and filtered for statistical significance ($FDR < 0.1$). We chose the less stringent significance filter since the mutation rate analysis of functional gene sets was relatively less powered given that fewer sites were considered. The pathways with significant TSS-specific mutation rates were visualised as an enrichment map⁶⁵ in Cytoscape and major biological themes were manually curated as described earlier⁴⁵. Nodes in the enrichment map were painted to reflect cancer types where these pathway enrichments were detected following the custom color scheme of the PCAWG project.

Associating regional mutation rates with driver mutations and recurrent copy-number alterations. We asked whether the localised mutation rates in CTCF binding sites, TSSs and open-chromatin sites were associated with driver mutations (*i.e.*, SNVs, indels) or recurrent copy-number alterations (CNAs) in cancer genomes. First we collected a high-confidence set of driver mutations and CNAs in the PCAWG cohort. Driver mutations in exons of protein-coding genes were predicted for each selected cancer type using the ActiveDriverWGS method³⁴. We used the PCAWG variant calls after filtering tumors as described above, corrected the results for multiple testing using the Benjamini-Hochberg FDR procedure and selected significant driver genes ($FDR < 0.05$). FDR correction was conducted separately for each cancer type across the pooled set of protein-coding and non-coding genes. Tumors with and without SNVs or indels in predicted driver genes were used for localised mutation rate analysis. Predictions of recurrent CNAs were derived from the pan-cancer dataset of GISTIC2 calls of the PCAWG project⁴⁹. All CNA lesions at 95% confidence scores were considered and amplifications and deletions were analysed separately. High-confidence CNA events were used (GISTIC2 score = 2). Tumors with and without CNAs in the recurrently altered regions as defined by GISTIC2 were used for localised mutation rate analysis. Each cancer type was considered separately. Next we filtered very frequent and infrequent drivers and CNA events to improve the power of the RM2 analysis. We selected the driver genes and CNA regions with at least 25 tumors in the mutated (or copy-number altered) group of tumors and filtered very frequent drivers and CNAs affecting more than 2/3 of the cohort. Each driver gene and recurrent CNA locus in each cancer type was then analysed for associations with localised mutation rates in the three categories of genomic elements (open-chromatin sites, CTCF binding sites, TSSs). The binary co-factor in RM2 was

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

Associating CNAs with mRNA abundance. To evaluate the functional role of CNAs associated with localised mutation rates, we compared mRNA abundance levels of genes in the CNA loci in groups of tumors defined by the presence or absence of the CNA events, using matching RNA-seq data available in PCAWG⁴⁴. Genes in CNA loci were retrieved from the PCAWG GISTIC2 dataset and genes with low mRNA abundance were removed (median FPKM-UQ < 1). mRNA abundance levels of genes in CN amplified and non-amplified (*i.e.*, balanced and deleted) tumors were compared using the nonparametric Wilcoxon test. One-sided tests were used, assuming that change in mRNA abundance would match the underlying copy-number change (*i.e.*, copy number amplifications were tested for increase in mRNA abundance of matching genes). Results were adjusted for multiple testing correction using the Benjamini-Hochberg FDR procedure and significant results were selected ($FDR < 0.05$). To confirm the CNAs, we retrieved the consensus dataset of CNA calls in each tumor from the PCAWG study⁴⁹ and visualised the detected CNA segments normalised by tumor ploidy predictions in subsets of tumors defined by the presence or absence of these CNA events. Known cancer genes of the COSMIC Cancer Gene Census database⁶⁶ (v91, downloaded 14.05.2020) were identified among the genes with mRNA/CNA associations.

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

systematic analysis of open-chromatin sites, TSSs, and CTCF binding sites in a comparable set of simulated variant calls from PCAWG. This simulated variant set was derived earlier from the same set of tumor genomes using trinucleotide-informed shuffling of mutations⁴. Simulated variant calls were analysed similarly to true variant calls for total mutation counts, reference and alternative nucleotide combinations, predicted mutational signatures and transcription and mutation strand properties. Results from RM2 were adjusted for multiple testing using the Benjamini-Hochberg FDR procedure separately for results derived from true and simulated variant calls. As expected, simulated variant calls revealed no statistically significant results of localised mutation rates in any cancer type, site type or mutation subset ($FDR < 0.05$). We then visualised the distribution of log-transformed P -values derived from true and simulated variant calls using quantile-quantile (QQ) plots and found that highly significant P -values were detected in true datasets while the P -values derived from simulated variant calls were uniformly distributed as expected. These analyses show that our model is well-calibrated and is not subject to inflated false-positive findings. Second, to evaluate the statistical power of RM2, we performed systematic down-sampling by randomly selecting subsets of sites and tumors for localised mutation rate analysis. We focused on the PCAWG liver hepatocellular carcinoma (Liver-HCC) cohort of 300 samples and CTCF sites. A series of down-sampling configurations were used (2000, 5000, ..., 100,000 sites sampled; 25, 50, ..., 300 genomes sampled). Each configuration was tested 100 times in different subsets of data. For a power analysis, we evaluated the fraction of runs that revealed a significant enrichment of somatic mutations at CTCF sites ($P < 0.05$) and the median P -value of these 100 runs. Third, we evaluated the parameter values of RM2 that determine the genomic width of each site and the control regions of upstream and downstream flanking sequences. As expected, site-specific mutation rates were consistently identified at multiple values of the width parameter for each class of site (open-chromatin sites, CTCF binding sites, TSSs), indicating robustness of our analysis to different parameter values. However, different site classes showed preferences towards shorter sites (CTCF binding sites: 20-100 bps) or longer sites (open-chromatin sites and TSS: 200-800 bps), likely due to different underlying mutational processes. For the entire study, the optimal genomic size of every site class was selected based on the strongest effect size and significance across multiple cancer types. The value of 50 bps (± 25 bps) was selected for CTCF sites. For open-chromatin sites and TSSs, we selected the common site width of 200 bps (± 100 bps) that showed

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

Acknowledgments. This work was supported by the Canadian Institutes of Health Research (CIHR) Project Grant to J.R. and the Investigator Award to J.R. from the Ontario Institute for Cancer Research (OICR). C.A.L. partially was 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.

References

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

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

hypoxia-inducible factor activation and tumorigenesis. *Mol Cell Biol* **28**, 718-731, doi:10.1128/MCB.01338-07 (2008).

59 Ishii, T. *et al.* A mutation in the SDHC gene of complex II increases oxidative stress, resulting in apoptosis and tumorigenesis. *Cancer Res* **65**, 203-209 (2005).

60 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443, doi:10.1038/s41586-020-2308-7 (2020).

61 Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880-886, doi:10.1126/science.aaa6806 (2015).

62 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25-29, doi:10.1038/75556 (2000).

63 Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res* **46**, D649-D655, doi:10.1093/nar/gkx1132 (2018).

64 Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic acids research* **35**, W193-200, doi:10.1093/nar/gkm226 (2007).

65 Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PloS one* **5**, e13984, doi:10.1371/journal.pone.0013984 (2010).

66 Futreal, P. A. *et al.* A census of human cancer genes. *Nat Rev Cancer* **4**, 177-183, doi:10.1038/nrc1299 (2004).