Genome Biology

RESEARCH Open Access

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

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

* Correspondence: Juri.Reimand@utoronto.ca
†Christian A. Lee and Diala Abd-Rabbo contributed equally to this work.
[1]Computational Biology Program, Ontario Institute for Cancer Research, Toronto, ON, Canada
[2]Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada
Full list of author information is available at the end of the article

Abstract

Background: Cancer genomes are shaped by mutational processes with complex spatial variation at multiple scales. Entire classes of regulatory elements are affected by local variations in mutation frequency. However, the underlying mechanisms with functional and genetic determinants remain poorly understood.

Results: We characterise the mutational landscape of 1.3 million gene-regulatory and chromatin architectural elements in 2419 whole cancer genomes with transcriptional and pathway activity, functional conservation and recurrent driver events. We develop RM2, a statistical model that quantifies mutational enrichment or depletion in classes of genomic elements through genetic, trinucleotide and megabase-scale effects. We report a map of localised mutational processes affecting CTCF binding sites, transcription start sites (TSS) and tissue-specific open-chromatin regions. Increased mutation frequency in TSSs associates with mRNA abundance in most cancer types, while open-chromatin regions are generally enriched in mutations. We identify ~ 10,000 CTCF binding sites with core DNA motifs and constitutive binding in 66 cell types that represent focal points of mutagenesis. We detect site-specific mutational signature enrichments, such as SBS40 in open-chromatin regions in prostate cancer and SBS17b in CTCF binding sites in gastrointestinal cancers. Candidate drivers of localised mutagenesis are also apparent: *BRAF* mutations associate with mutational enrichments at CTCF binding sites in melanoma, and *ARID1A* mutations with TSS-specific mutagenesis in pancreatic cancer.

Conclusions: Our method and catalogue of localised mutational processes provide novel perspectives to cancer genome evolution, mutagenesis, DNA repair and driver gene discovery. The functional and genetic correlates of mutational processes suggest mechanistic hypotheses for future studies.

Introduction

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

Complex variation in mutation frequencies is also apparent across intermediate genomic resolutions spanning hundreds to thousands of nucleotides encapsulating diverse functional genomic elements such as exons, transcription factor binding sites (TFBS) and chromatin architectural elements [8, 20]. Genomic elements bound by nucleosomes and transcription factors (TFs) show increased mutation frequencies in cancer genomes [21–23]. Active promoters in melanoma are enriched in UV-induced C>T somatic mutations resulting from differential activity of nucleotide excision repair influenced by DNA binding of regulatory proteins [24, 25]. 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 harbour relatively fewer mutations due to increased DNA repair activity. While the majority of such non-coding mutations represent functionally neutral passengers, some regulatory elements at the high end of the mutation frequency spectrum may undergo positive selection due to their effects on cancer phenotypes. For example, the mutation 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 cells [31–33]. 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 elements [34–36] and shown that non-coding mutations converge on molecular pathways and regulatory networks involved in oncogenesis [37, 38]. Thus, we need to characterise localised mutational processes to deconvolute the role of carcinogens and endogenous mutational processes and the effects of positive selection in the non-coding genome. However, few dedicated computational methods exist to analyse variation in mutation frequencies at this resolution. As a result, there is a lack of large-scale analyses of the local mutation landscape in pan-cancer WGS datasets, leaving patterns of mutational enrichment and depletion undetected, and 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 burden and genetic covariates to control for variations at the trinucleotide and megabase resolution while isolating site-level effects. We performed a systematic analysis of local mutation frequency variation in three classes of gene-regulatory and chromatin architectural genomic elements across 2419 whole cancer genomes of the ICGC/TCGA Pan-cancer Analysis of Whole Genomes (PCAWG) project [3]. We found a pervasive mutational enrichment at these functional non-coding elements that was characterised by specific mutational signatures and transcriptional and pathway-level activities 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 reveal the diversity of mutational processes in functional non-coding elements of the cancer genome and their roles in somatic genome evolution, cancer phenotypes and molecular heterogeneity.

Results

A statistical framework for quantifying localised mutagenesis in cancer genomes

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

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

To determine whether the mutation frequencies of genomic elements differ from flanking sequences given trinucleotide-level and megabase-scale covariates, we evaluate the significance of the covariate *isSite* using a likelihood ratio test. Significant and

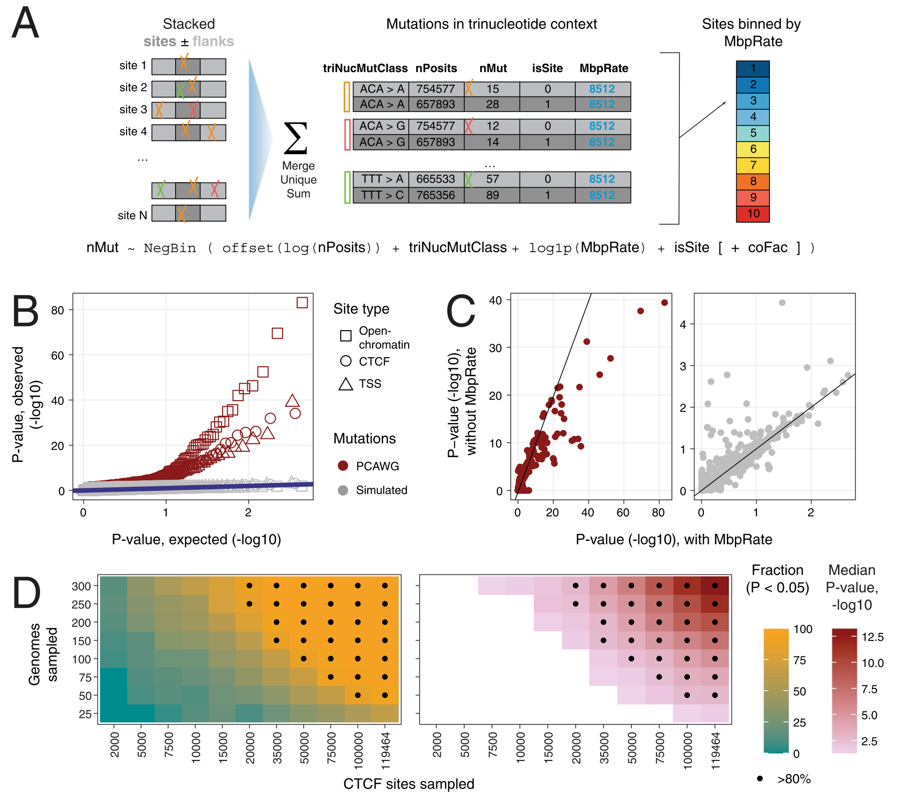

Fig. 1 Characterising local mutational processes with RM2. **a** Method overview. RM2 studies a set of genomic elements (i.e. sites) and somatic mutations in cancer genomes using a negative binomial regression model. Sites of constant genomic width (dark grey) and two control flanking sequences (light grey) are used (*isSite*). Sites and flanks are collapsed into unique nucleotides and grouped to ten bins using their megabase-scale mutation frequency (*MbpRate*). Mutations in sites and flanks (*nMut*) are grouped by trinucleotide type (triNucMutClass). Trinucleotide content corresponding to the potential genomic space for mutations is used as model offset (*nPosits*). Log-likehood tests are used to compare the mutation frequencies in sites and flanking regions by removing the model factor *isSite*. The optional factor *coFac* enables interaction analysis of genetic and clinical variables. **b** QQ-plot shows the observed and expected *P* values of true and simulated mutations from PCAWG. No significant signals were identified in simulated data (FDR < 0.05), indicating that our method is well-calibrated. **c** Comparison of model performance with and without *MbpRate* covariate. Analysis of true (left) and simulated mutations (right) shows the advantage of modelling megabase-scale mutation frequency. **d** Power analysis of RM2 using down-sampling of CTCF binding sites and liver cancer genomes. Fraction of significant results (left) and median *P* value (right) are shown. Panels **b** and **c** include total mutations, strands- and signature-specific mutations as in Fig. 3. Only total mutations were included for analyses in **c**, **d**

positive coefficients of this covariate indicate increased mutation frequencies 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 localised activity of mutational processes. Given a binary subgroup classification of tumours (*coFac*), we evaluate the significance of its interaction with local mutation frequencies (*isSite:coFac*). Positive coefficients of the interaction indicate that the mutation frequencies in a clinical or genetic tumour subgroup are elevated when accounting for the overall differences of the subgroups. We also extend the analysis to classes of mutations, such as those of COSMIC mutational signatures, by allowing only specific classes to be included in the mutation counts (*nMut*). We evaluated the performance of our method using simulated datasets, power analysis and parameter variations as described below (Fig. 1b–d).

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

To study localised mutation frequencies in gene-regulatory and chromatin architectural elements, we used the pan-cancer dataset of 2514 whole cancer genomes of the PCAWG project [3] with high-confidence SNVs and indels. We individually analysed 25/35 cancer types with at least 25 samples, as well as the pan-cancer set of all 35 cancer types (Additional file 1: Figure S1A). Hypermutated tumours (69 or 2.7%) were excluded to avoid confounding effects. To perform a conservative analysis, we excluded a few tumours as outliers (33 or 1.4%) where even single-sample RM2 analysis revealed highly significant mutational enrichments ($FDR < 0.001$) (Additional file 1: Figure S1B). This ensured that our findings were shared across most tumour genomes and did not represent isolated trends representative of few tumours. The final analysis considered 2419 cancer genomes of 35 cancer types with 22.7 million mutations including 1.62 million indels. In addition to all mutations combined, we grouped the mutations by COSMIC mutational signatures of single-base substitutions (SBS) inferred in the PCAWG study [14], reference and alternative nucleotide pairs, and DNA strand information. Indel mutations were pooled with SNVs and also analysed separately.

Three classes of genomic elements spanning 1,269,347 unique loci and 10% (337.0 Mbps) of the human genome were analysed. These included 119,464 CTCF binding sites conserved in at least two cell lines in the ENCODE project [39], 37,309 transcription start sites (TSS) of protein-coding genes from the Ensembl database (GRCh37) and 1,193,391 unique open-chromatin sites. Open-chromatin sites were analysed in tissue-specific subsets, ranging from 43,000 to 500,000 sites per cancer type (Additional file 2: Table S1A). For most cancer types (17/25), matching open-chromatin sites were adapted from the ATAC-seq profiles of primary tumours of The Cancer Genome Atlas (TCGA) project [40]. The pan-cancer analysis used 500,183 pan-cancer sites from TCGA. For eight cancer types, DNAse-seq profiles of the closest relevant normal tissues of the Roadmap Epigenomics project [41] were used. Open-chromatin sites were filtered to exclude TSSs and CTCF binding sites to allow direct comparison and reduce confounding effects of the three site classes.

The analysis revealed a landscape of localised mutational processes in gene-regulatory and chromatin architectural elements of cancer genomes (Fig. 2a) (Additional file 2: Table S1B). We found 307 significant differences in mutation frequencies in the three classes of genomic elements (RM2, $FDR < 0.05$). These affected 21 cancer types and included 23 unique mutational signatures. The majority of findings (281) indicated mutational enrichments at the sites compared to adjacent flanking regions, while a few reduced mutation frequencies were also observed (26 or 8%). The strongest cumulative signals were found in open-chromatin sites in prostate, liver and breast cancers and CTCF binding sites in liver and oesophageal cancers.

We first focused on the mutational profiles of CTCF DNA-binding sites. Mutational enrichments in CTCF binding sites were the strongest in liver hepatocellular carcinoma (RM2 $FDR = 1.22 \times 10^{-12}$, fold-change (FC) = 1.08), oesophageal adenocarcinoma ($FDR = 6.0 \times 10^{-20}$, FC = 1.17) and stomach adenocarcinoma ($FDR = 6.2 \times 10^{-11}$, FC = 1.16), and the pan-cancer cohort. Smaller enrichments were detected in melanoma, pancreatic and breast cancer ($FDR \leq 0.02$). Subgroups of mutations revealed further signals. Strong enrichments of thymine base substitutions (T>G, T>C, T>A) were found

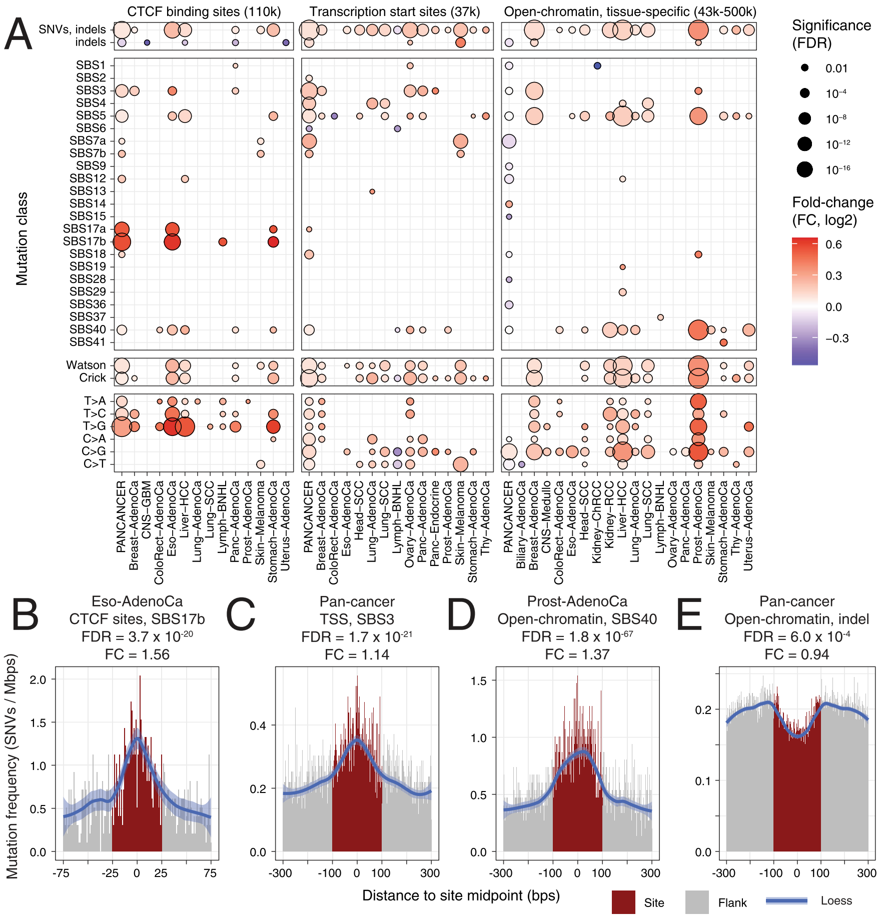

Fig. 2 Comprehensive map of mutational processes in gene-regulatory and chromatin architectural elements of cancer genomes. **a** Comparison of mutation frequencies in DNA-binding sites of the CTCF chromatin architectural factor (left), transcription start sites (TSS) and cancer-specific open-chromatin sites in 2419 whole cancer genomes (FDR < 0.05). Total mutations (SNVs, indels) and mutations grouped by single-base substitution (SBS) signatures, substitution types and DNA strand were analysed. Open-chromatin sites were filtered to exclude TSSs and CTCF sites. **b–e** Examples of localised mutation frequencies and signatures: **b** Enrichment of SBS17B in CTCF binding sites in oesophageal adenocarcinoma, **c** pan-cancer enrichment of SBS3 mutations in TSSs, **d** enrichment of SBS40 mutations in open-chromatin sites in prostate adenocarcinoma, **e** pan-cancer depletion of indel mutations in open-chromatin sites

in nine cancer types (e.g. T>G in Liver-HCC, FDR = 1.3 × 10^{-30}, FC = 1.46), while cytosine base substitutions were not enriched. We then asked whether CTCF binding sites were characterised by COSMIC SBS signatures. The strongest site-specific enrichments were found for SBS17 in oesophageal and stomach cancer and non-Hodgkin's lymphoma (SBS17b in oesophageal cancer: FDR = 3.7 × 10^{-20}, FC = 1.56) (Fig. 2b). Liver and other cancer types with overall frequent CTCF binding site mutations did not show a site-specific enrichment of SBS17. The aetiology of SBS17b is unknown; however, it has been linked to acid reflux and oxidative damage to DNA in gastro-oesophageal cancers [42, 43], and a similar mutational signature found in metastatic tumours 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. These results confirm earlier reports of elevated mutation frequencies in CTCF DNA-binding sites [21, 27, 28] in a large and diverse dataset of whole cancer genomes, validating our computational model and refining the annotation of the mutational processes associated with CTCF binding sites.

TSS of protein-coding genes were significantly enriched in mutations in the pan-cancer cohort ($FDR = 1.7 \times 10^{-37}$, FC = 1.07) and in cohorts of 13/25 cancer types, most prominently in melanoma ($FDR = 3.6 \times 10^{-14}$, FC = 1.15), breast, head, lung, ovary and pancreatic cancers ($FDR \leq 10^{-5}$). Enrichments of cytosine base substitutions were found in melanoma (C>T, $FDR = 1.4 \times 10^{-15}$, FC = 1.19) as well as lung, ovarian and head and neck cancers. In contrast to CTCF binding sites, thymine base substitution frequencies were not elevated. Mutational signature analysis highlighted an elevated activity of the ageing-associated signature SBS5 in the pan-cancer cohort ($FDR = 2.4 \times 10^{-13}$, FC = 1.07) and in eight cancer types. Signature SBS3, associated with defects of homologous recombination-based DNA damage repair, was enriched at TSSs in the pan-cancer cohort ($FDR = 1.7 \times 10^{-21}$, FC = 1.14) (Fig. 2c) as well as breast, pancreatic and ovarian cancers ($FDR \leq 10^{-5}$). Spontaneous formation of endogenous DNA double-strand breaks at promoters has been associated with the pause and release of RNA polymerase II and linked to chromosomal translocations in cancer [44], suggesting a mechanism of this TSS-specific mutagenesis in cancer genomes. TSSs were enriched in carcinogen-driven mutational signatures, such as the ultraviolet light signature in melanoma (SBS7a: $FDR = 1.1 \times 10^{-14}$, FC = 1.21) and the tobacco signature in two lung cancer cohorts (SBS4: $FDR \leq 10^{-5}$). These match the major mutagens and exposures of those cancer types, indicating an overall increased vulnerability of TSSs to mutational processes and carcinogens. This analysis extends previous reports of increased mutation frequencies in promoters in melanoma [24, 25] to additional cancer types, demonstrating that TSS-specific mutational processes are widely active in cancer genomes.

Tissue-specific open-chromatin regions were also enriched in mutations in 12/25 cancer types, with the strongest signals in prostate cancer ($FDR = 9.8 \times 10^{-81}$, FC = 1.29) as well as liver, lung, breast and kidney cancer ($FDR \leq 10^{-14}$) whereas C>G mutations were more frequently observed compared to other base substitutions. Mutational signature analysis revealed the enrichment of SBS40 mutations in open-chromatin regions in nine cancer types and the pan-cancer cohort, with the strongest signal apparent in prostate cancer ($FDR = 1.8 \times 10^{-67}$, FC = 1.33) (Fig. 2d). The ageing-associated signature SBS5 was enriched in open-chromatin sites in eleven cancer types, especially in liver, prostate, breast and lung cancer ($FDR \leq 10^{-9}$). In contrast to strong signals in individual cancer types, the analysis of pan-cancer open-chromatin sites and mutations revealed only a trend of elevated mutation frequencies at open-chromatin sites (FDR = 0.056, FC = 1.01). The pan-cancer sites were depleted of indel mutations ($FDR = 6.0 \times 10^{-4}$, FC = 0.94) (Fig. 2e) and several less-frequent mutational signatures, while those effects were not observed in individual cancer types. The limited findings of the pan-cancer analysis emphasise the importance of integrating open-chromatin profiles of matched cancer types to study the interactions of chromatin and mutagenesis. Overall, open-chromatin regions of individual cancer types are enriched in mutations through specific mutational processes acting at these local scales.

We used the generated map to benchmark our model and compared it to a PCAWG dataset of simulated variant calls designed to approximate neutral genome evolution [4]. First, analysis of simulated data expectedly revealed no significant differences in mutation frequencies in the three classes of elements (all FDR > 0.05). Quantile-quantile analysis of *P* values confirmed that the model was well calibrated for true and simulated mutations (Fig. 1b). Second, we compared the models with and without accounting for the megabase-scale mutation frequency covariate (MbpRate) and found that including the information led to higher significance in the true mutation dataset and lower significance in the simulated mutation dataset (Fig. 1c). Third, we evaluated the statistical power of our model by analysing total mutations in down-sampled subsets of liver cancer genomes and CTCF binding sites (Fig. 1d). For example, the mutational enrichment in CTCF sites was detectable 80% of the time when sampling 75 genomes and 75,000 CTCF binding sites. Fourth, we varied the parameter corresponding to the normalised width of genomic elements for the three classes (Additional file 1: Figure S2A). Site-specific differences in mutation frequencies were robustly detected for sites defined using a range of possible sizes. However, mutational enrichments in sites bound by CTCF were generally narrower (50 bps) compared to TSS and open-chromatin sites (200 bps), indicating differences in mutational processes. In summary, our method provides a versatile and well-calibrated framework for analysing localised mutational processes in cancer genomes.

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

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

Higher mutation frequencies in TSSs strongly associated with tissue-specific mRNA abundance of target genes (Fig. 3a) (Additional file 2: Table S1C). In the fifth bin of highly transcribed genes, TSSs were consistently enriched in mutations in 12/20 cancer types (FDR < 0.05) and the pan-cancer cohort (FDR = 6.2×10^{-38}, FC = 1.15). The strongest associations of transcriptional activity and TSS-specific mutagenesis were found in melanoma, ovarian, lung, pancreatic and breast cancer (FDR ≤ 10^{-4}, FC ≥ 1.15) (Fig. 3b, c). The TSSs of genes with intermediate transcript abundance (40–80 percentile) were also enriched in mutations in seven cancer types. In contrast, TSSs of silent and low-abundance genes were not differentially mutated compared to flanking controls. Interestingly, the overall mutation frequency was higher in and around silent TSSs, suggesting that closed chromatin is exposed to lower levels of DNA repair. Of note, the top gene bin was highly variable in terms of mRNA abundance (e.g. range 11–7100 FPKM-UQ in the pan-cancer cohort) (Fig. 3c). Therefore, the mutational enrichment in the top gene bin may be partially contributed to by genes with very high expression. TSS bins with higher transcript abundance were also enriched in mutational signatures commonly observed in their respective cancer types (Additional file 1: Figure S3A). This analysis highlights a pan-cancer mutational process localised in TSSs that appears to be driven by transcriptional activity.

Compared to TSSs, mutation frequencies in open-chromatin sites did not associate with transcript abundance of target genes (Fig. 3a,b). Mutational enrichments in multiple gene bins were observed in melanoma, liver cancer and prostate cancer; however, the effect sizes were smaller and most cancer types showed no significant changes in local mutation frequencies. In this unbiased comparison, we included equally sized bins of TSSs and open-chromatin sites using down-sampling, since open-chromatin sites considerably outnumbered TSSs and had the power to detect smaller effects. An extended analysis of all open-chromatin sites grouped by transcript abundance showed stronger mutational enrichments in most cancer types that is consistent with our analysis above (Additional file 1: Figure S2B); however, no associations with transcript abundance were apparent. This analysis suggests that transcription-associated local mutagenesis is more prominent in TSSs than in other regions of accessible chromatin.

To explore the functional associations of elevated mutation frequencies at TSSs, we performed a pathway enrichment analysis by adapting RM2 to gene sets of GO biological processes and Reactome molecular pathways [47]. This allowed us to test our hypothesis that the TSSs enriched for mutations are concentrated in specific biological processes. We found 336 unique pathways and processes with pronounced enrichments of mutations at TSSs in eleven cancer types (FDR < 0.05) (Fig. 3d) (Additional file 2: Table S1D). Half of the pathways were found in at least the melanoma cohort (51%). However, one third of the pathways were found in two or more cancer types, indicating that the pathway associations of elevated mutagenesis at TSSs often apply more generally to multiple cancer types. Translation, ribosome biogenesis and RNA processing were among the largest groups of pathways found. This is expected as the translational machinery is ubiquitously active in proliferating cells and includes many highly expressed genes. Processes related to gene regulation and chromatin organisation were also detected. Besides these general housekeeping processes, 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, and druggable signalling pathways (e.g. MAPK, Wnt, Notch) were identified in multiple cancer types. The pathway analysis supports our findings of frequent TSS mutations associated with increased transcription. It also highlights a variety of core cellular processes and cancer pathways where mutations accumulate in core promoters across multiple cancer types and may have functional consequences.

Mutational enrichment at CTCF binding sites is associated with constitutive DNA binding and core sequence motifs

We tested whether the elevated mutation frequencies in CTCF binding sites were associated with functional site characteristics. We used the extent of conservation of DNA binding across 70 cell lines as a proxy of site activity. We grouped 162,209 unique CTCF binding sites catalogued in ENCODE into five equal bins such that the first bin contained sites observed in a single cell line while the fifth bin included constitutively bound sites with the median site found in 67/70 cell lines (Fig. 4a). Doing so allowed us

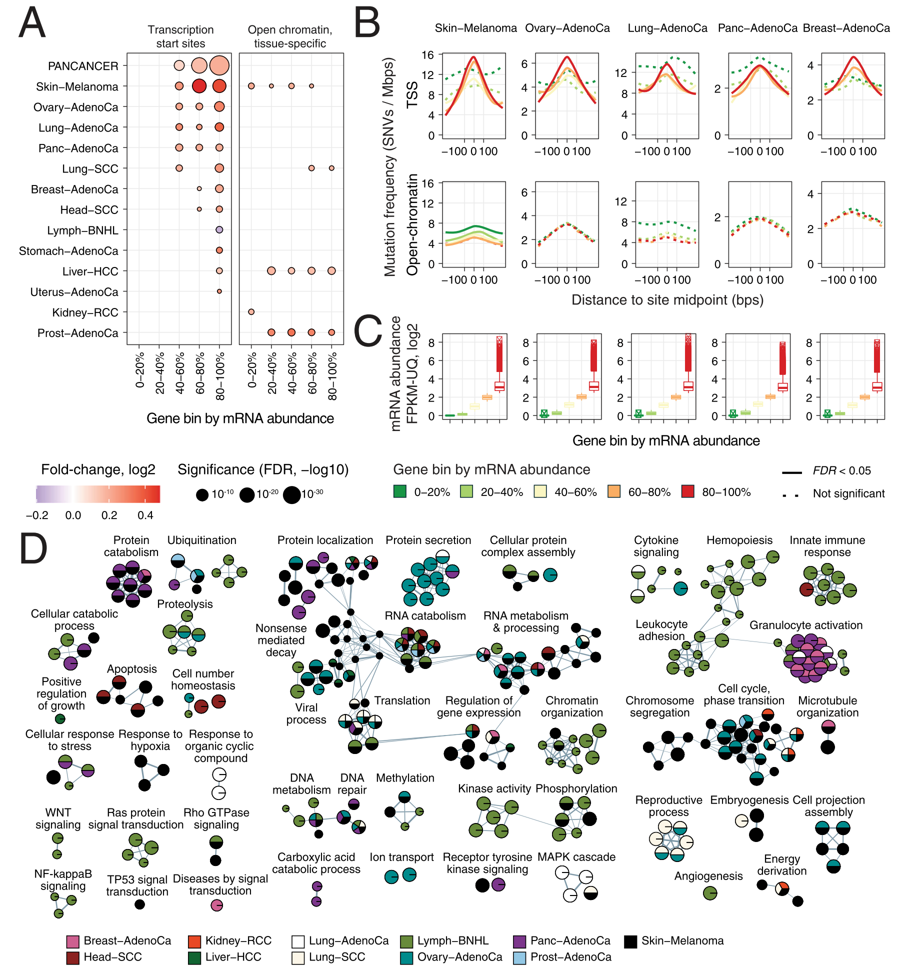

Fig. 3 Mutational enrichment at transcription start sites associates with mRNA abundance of target genes and diverse pathways. **a** Comparison of mutation frequencies in TSSs and tissue-specific open-chromatin sites grouped by mRNA abundance of target genes in matching tumours (FDR < 0.05). Equal numbers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

to test mutation frequency and genomic associations at different levels of site activity. The constitutively bound CTCF binding sites were enriched in chromatin loop anchors [48] (29% observed vs. 12% expected, Fisher's exact $P < 10^{-300}$) and consensus core CTCF DNA-binding motifs (48% observed vs. 29% expected, $P < 10^{-300}$) (Fig. 4b). Together, these features imply that the constitutive bin of sites represents functionally important CTCF binding sites.

The constitutive CTCF binding sites were characterised by elevated mutation frequency in the pan-cancer cohort (FDR = 1.4×10^{-71}, FC = 1.18) and in ten cancer types, especially stomach, liver and oesophageal cancers as well as melanoma (FDR ≤ 10^{-19}, FC ≥ 1.25) (Fig. 4c,d) (Additional file 2: Table S1E). In contrast, all other bins of CTCF sites showed no significant enrichment of mutations (FDR > 0.05), even in the sites of the fourth bin which still represent a high level of conservation of CTCF binding

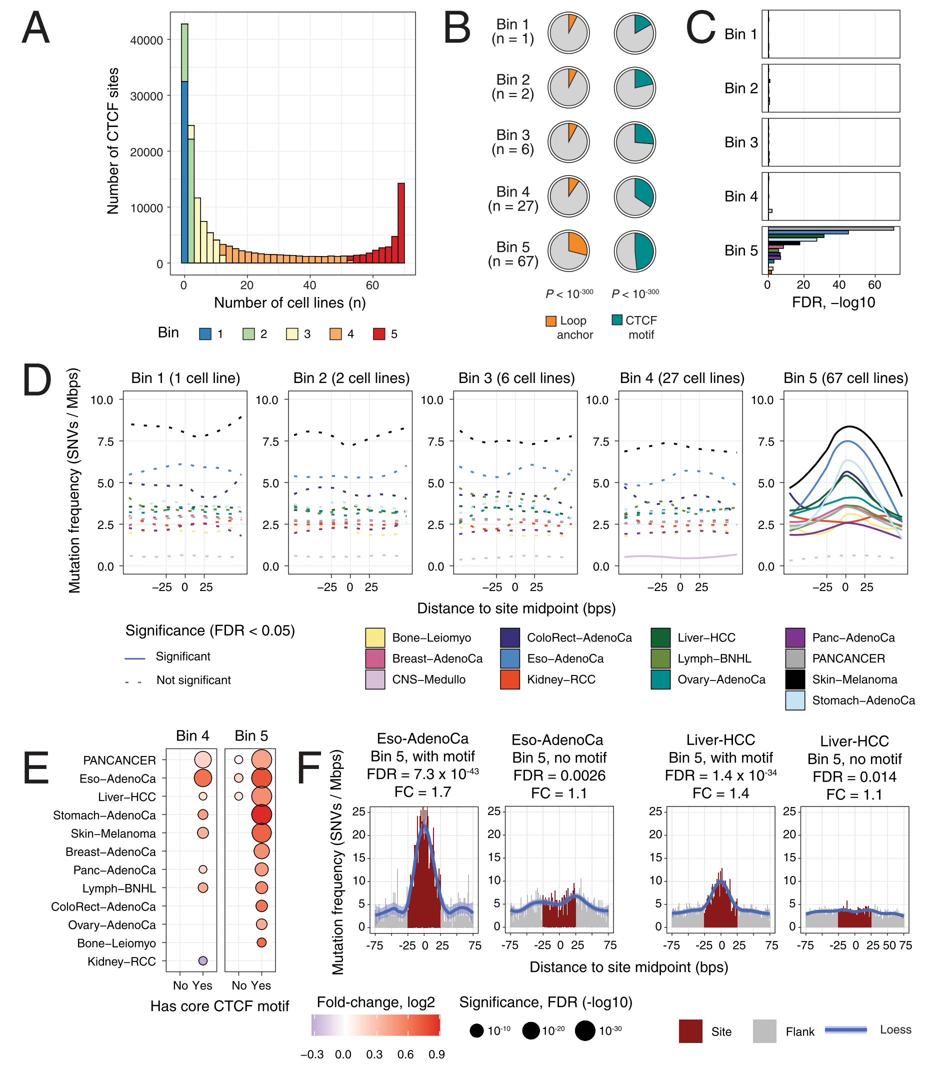

Fig. 4 Localised mutational processes at constitutively active binding sites of CTCF. **a** Histogram of CTCF binding sites with number of cell lines in ENCODE. Sites were grouped as five equal bins based on conservation across cell lines (colours). Bimodal distribution reveals a subset of sites detected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

(median 27/70 cell lines). Mutational signature analysis revealed over-represented signatures such as SBS3, SBS5, SBS7, SBS17 and SBS40 at these sites in various cancer types; however, a common signature for constitutive CTCF binding was not apparent (Additional file 1: Figure S3B). In summary, this analysis emphasises the potential role of constitutive CTCF binding activity in local mutagenesis.

We then examined whether the presence of CTCF DNA-binding motifs in the ChIP-seq peaks played a role in localised mutagenesis. We categorised all the CTCF binding sites into two subsets based on the presence or absence of a consensus core DNA motif and repeated the conservation-based grouping of the two subsets of sites. The localised mutational process at CTCF binding sites appears to be driven by the combination of constitutive binding and core motif presence (Fig. 4e) (Additional file 2: Tables S1F, S1G). The constitutively bound CTCF sites matching the DNA motifs were strongly enriched in mutations in ten cancer types and the pan-cancer cohort. In contrast, the constitutively bound sites lacking the core DNA motif were not enriched. For example, oesophageal cancer showed a strong mutational enrichment in the constitutively bound CTCF sites with the core motif present ($FDR = 7.3 \times 10^{-43}$, $FC = 1.70$), while the signal was clearly attenuated in the constitutively-bound sites with no core motif ($FDR = 0.0026$, $FC = 1.14$) (Fig. 4f). This was confirmed in liver cancer (bin 5 with motifs: $FDR = 1.4 \times 10^{-34}$, $FC = 1.44$; bin 5 without motifs: $FDR = 0.014$, $FC = 1.08$). Highly conserved sites with median conservation in 27/70 cell lines also showed motif-driven mutational enrichments. Strikingly, no signal of mutational enrichment was detected in any other bins of sites regardless of motif presence (Additional file 1: Figure S4A). The findings were confirmed in a tissue-specific analysis integrating breast and liver cancer mutations with CTCF binding sites of the cancer cell lines MCF-7 and HepG2, respectively (Additional file 1: Figure S4B, S4C).

Our analysis highlights a minority of CTCF binding sites (9552 or 5.9%) with constitutive CTCF binding and core DNA motifs that are the primary target of a local mutational process in many cancer types. This analysis refines the established pattern of mutational enrichment in CTCF binding sites by defining a subset of frequently mutated sites with specific functional properties, such as involvement in chromatin architectural and gene-regulatory interactions [48, 49]. Our observations are consistent with previous findings of CTCF sites where a particularly strong mutational enrichment was reported in sites shared among a few cell lines [50, 51]. Conservation is a property of functionally integral CTCF binding sites, which upon disruption, can lead to changes in underlying chromatin architecture and gene regulation [52] and is associated with activation of proto-oncogenes [26]. Therefore, the accumulation of mutations at those sites may have functional consequences in cancer.

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

To find potential genetic mechanisms of localised mutagenesis, we tested whether the presence of specific features such as recurrent mutations in tumour genomes associated with higher mutation frequencies at gene-regulatory and architectural elements. To enable this analysis, we collected 37 driver genes with frequent SNVs and indels predicted using the ActiveDriverWGS method [34], 70 recurrent copy-number alterations (CNAs) detected in the PCAWG project using the GISTIC2 method [53] and two genome-wide measures of aneuploidy, whole-genome duplication (WGD) and percent genome altered (PGA) (Additional file 1: Figure S5). Integrative analysis with RM2 identified 41 significant interactions between genomic features and site-specific elevations in mutation frequencies in nine cancer types, including five driver genes

(*ARID1A*, *BRAF*, *CTNNB1*, *HIST1H1C*, *SETD2*), 25 recurrently amplified loci and one locus with genomic loss (RM2 FDR < 0.05, interaction $P < 0.05$) (Fig. 5a) (Additional file 2: Table S1H).

ARID1A mutations in pancreatic adenocarcinoma had one of the strongest interactions with localised mutational enrichment (interaction $P = 5.0 \times 10^{-4}$) (Fig. 5b). The 24 pancreatic cancer genomes with *ARID1A* mutations showed a significant increase in mutation frequency at TSSs (FDR = 1.7×10^{-7}, FC = 1.27). In contrast, the majority of tumours (206) lacking *ARID1A* mutations showed a weaker, albeit significant, mutational enrichment at TSSs, likely explained by the improved statistical power of the larger set of samples analysed (FDR = 1.2×10^{-6}, FC = 1.09). No significant interaction was observed at CTCF and open-chromatin sites (Additional file 1: Figure S6A). *ARID1A* is a tumour suppressor encoding a member of the SWI/SNF chromatin remodelling complex that regulates chromatin accessibility at regulatory elements and is involved in the maintenance of genomic stability and DNA repair [54]. *ARID1A* interacts with the topoisomerase TOP2A to facilitate its DNA binding and DNA decatenation, a process required for proper chromosome segregation in mitosis [55]. The SWI/SNF complex is also involved in DNA double-strand break repair [56] and nucleotide excision repair [57], and loss of *ARID1A* impairs polymerase pausing [58]. *ARID1A* is also one of the most frequently mutated genes in pancreatic adenocarcinoma [59] and in the PCAWG cohort, as 16/24 of tumours carried frameshift or stop-gain mutations that suggest loss of function. The inactivating *ARID1A* mutations may increase the mutation burden at TSSs as a result of impaired repair mechanisms and increased instability and, consequently, further contribute to transcriptional dysregulation.

In melanoma, driver mutations in *BRAF* were associated with mutational enrichment in CTCF binding sites (interaction $P = 0.022$) (Fig. 5c). The 28 tumours with *BRAF* mutations were enriched in CTCF binding site mutations (FDR = 5.8×10^{-5}, FC = 1.10), while 32 *BRAF*-wildtype tumours showed no enrichment (FDR = 0.31, FC = 1.03). BRAF serine/threonine kinase is a proto-oncogene and the activating V600E mutation defines a druggable subtype of melanoma [60, 61]. Previous work shows that ectopic expression of V600E-mutant *BRAF* in epithelial cell lines induces DNA double-stranded breaks and production of reactive oxygen species [62]. In this cohort, 22/28 melanomas carried V600E substitutions and V600K substitutions occurred in three additional tumours. No significant interactions of *BRAF* mutations and localised mutagenesis were found for TSSs and open-chromatin sites (Additional file 1: Figure S6B). This analysis suggests that melanomas defined by *BRAF* driver mutations may have increased activity of a mutational process acting specifically on CTCF binding sites.

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

To further decipher the interactions of recurrent genomic amplifications and local mutagenesis, we determined the genes located in the amplified loci that responded transcriptionally to amplifications. Using a cancer type-specific analysis, we found 282 unique genes in 25 amplified regions that were significantly upregulated in the tumours with these amplifications (Wilcoxon test, FDR < 0.05) (Fig. 5e) (Additional file 2: Table S1I). mRNA abundance associations were identified for all three types of genomic sites and six cancer types. Thirteen known cancer genes were found (*RAD21*, *CDK12*, *ERBB2*, *RECQL4*, *CCND1*, *DEK*, *BCL11A*, *MYC*, *JAK2*, *PDCD1LG2*, *BRAF*, *EXT1*, *PTK6*).

In breast cancer, *RAD21* mRNA abundance was significantly higher in 8q23.3-amplified tumours compared to non-amplified tumours (FDR $= 1.6 \times 10^{-7}$) (Fig. 5f), indicating that *RAD21* upregulation is driven in part by the genomic amplification. The 8q23.3 amplification was associated with mutational enrichments in CTCF binding sites in breast cancer ($P = 0.0089$) (Additional file 1: Figure S7A). *RAD21* encodes a subunit of the cohesin complex that co-binds DNA with CTCF to orchestrate transcriptional insulation and chromatin architectural interactions [48, 49], suggesting its role in CTCF-related mutagenesis.

As another example, *BRAF* was upregulated in the subset of breast cancers with 7q34 amplification (FDR $= 0.0081$) (Fig. 5g). The amplification was associated with increased mutagenesis at TSSs in breast cancer (interaction $P = 0.012$) (Additional file 1: Figure S7B). The 7q34 amplification was also significantly associated with CTCF binding site mutations in oesophageal cancer (interaction $P = 0.034$) (Additional file 1: Figure S7C); however, no matching RNA-seq data was available to confirm the transcriptional upregulation of *BRAF*. Thus, our analysis associates *BRAF* amplifications and point mutations to localised mutagenesis in multiple cancer types.

In summary, this integrative analysis of mutational processes with recurrent driver mutations, copy-number alterations and gene expression data provides a catalogue of somatic alterations that associate with localised mutational enrichments in gene-regulatory and chromatin architectural elements of the cancer genome. Since our evidence remains correlative, further computational and experimental approaches are needed to establish causal relationships. While a subset of these driver mutations and recurrent copy-number amplifications may be directly involved in mutagenesis and DNA repair, others may represent markers of tumour subtypes with specific exposures or endogenous factors. Deeper study of these associations may help decipher genetic mechanisms underlying mutational processes.

Discussion

The cancer genome is moulded by diverse mutational processes that continuously shape its broad megabase-scale domains and the fine context of single nucleotides. Here we focused on the mutational processes of an intermediate scale that affect

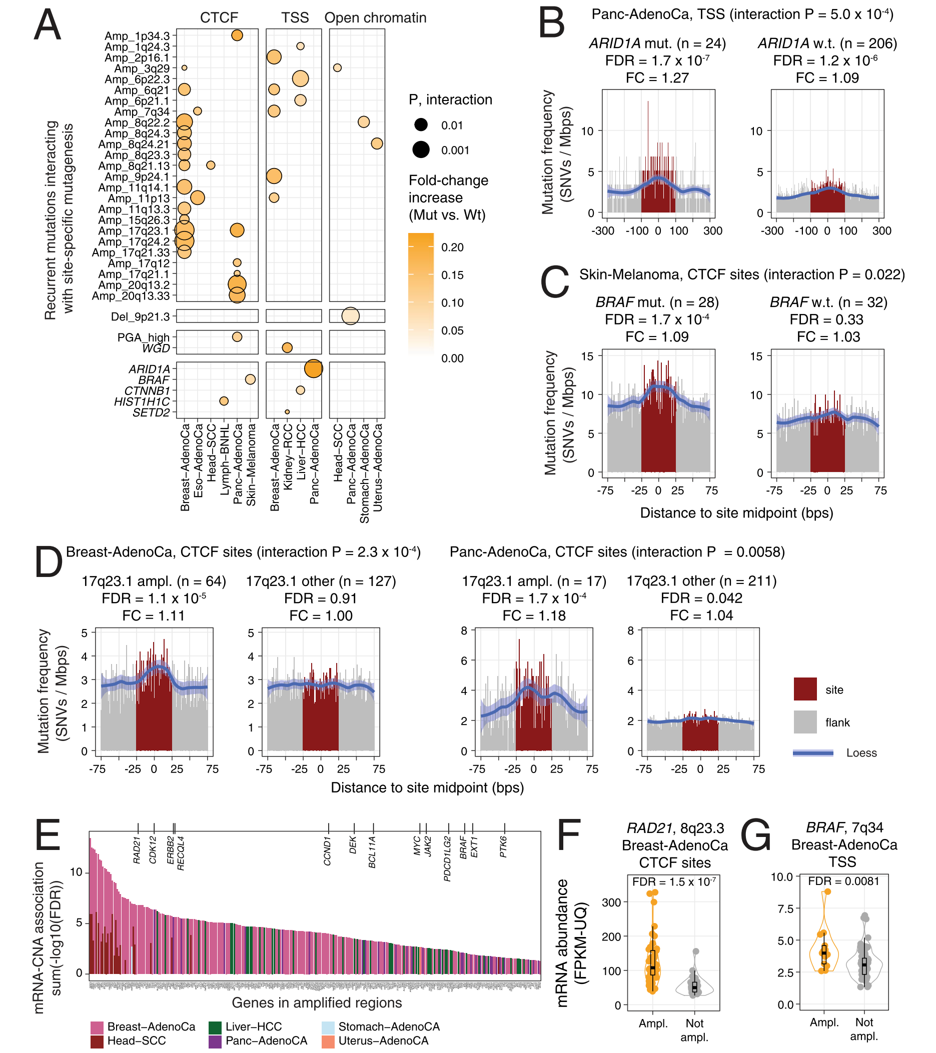

Fig. 5 Recurrent driver mutations and copy-number alterations (CNA) associate with localised mutagenesis. **a** Dotplot of the statistical interactions of recurrent mutations with increased mutation frequencies at sites (RM2 FDR < 0.05: interaction *P* < 0.05). Whole-genome duplication (WGD) and CNA burden (median-dichotomised percent genome altered, PGA) are also shown. **b**–**d** Examples of increased mutation frequency associating with recurrent driver mutations and CNAs. Tumours with and without recurrent mutations are shown (left vs right). **b** *ARID1A* mutations in pancreatic cancer associate with enriched mutations at TSSs. **c** *BRAF* mutations in melanoma associate with enriched mutations at CTCF binding sites. **d** 17q23.1 amplifications associate with enriched mutations at CTCF binding sites in breast and pancreatic cancer. **e** Copy-number amplified genes with amplification-driven increases in mRNA abundance. Known cancer genes are shown at the top. **f** *RAD21* is upregulated in the set of 8q23.3-amplified breast cancers associated with enriched mutations in CTCF binding sites. **g** *BRAF* is upregulated in the set of 7q34-amplified breast cancers associated with enriched mutations in TSSs

thousands of genomic elements, each spanning tens to hundreds of nucleotides. Such functional elements are widespread in the non-coding genome and play roles in gene expression control and chromatin architecture. Using our novel computational framework, we characterised the mutational landscape of gene-regulatory and chromatin architectural elements in diverse cancer types and identified their putative functional and genetic determinants. We show that the ubiquitous mutational enrichment in TSSs is associated with high transcription and biological processes involved in cellular housekeeping, but importantly with a variety of pathways and processes implicated in cancer. In contrast, transcriptionally silent genes showed no local deviations in mutation frequencies. Open-chromatin regions were generally enriched in mutations and showed no association with transcript abundance. CTCF binding sites exhibited a particular pattern of site-specific mutagenesis: a small fraction of hypermutated sites defined by a core CTCF DNA motif and constitutive binding across tens of human cell types dominated over the majority of sites that lacked mutational enrichment. Lastly, an integrative analysis of localised mutation frequencies with recurrent alterations in cancer genomes allowed us to predict genetic drivers of underlying mutational processes. The association of driver mutations in *ARID1A* and *BRAF* with localised mutational processes corroborate earlier functional and mechanistic evidence of mutagenesis and DNA repair disruption, lending confidence to our other identified interactions with driver genes and genomic amplifications and providing mechanistic hypotheses for future studies.

Open-chromatin regions were broadly enriched in mutations in most cancer types and the effect appeared to be independent of transcript abundance of predicted target genes. These findings contrast an earlier report that indicated decreased mutation rates in open-chromatin sites of cell lines at a similar resolution of the genome [29]. The use of non-matched chromatin accessibility data of cell lines and fewer cancer genomes may explain the depletion signals observed earlier that may be confounded by tissue-specific properties of chromatin state and mutagenesis. Our analysis of the open-chromatin regions of matching cancer types and a larger set of whole cancer genomes likely provides a more accurate view of mutational processes. In our study, the pan-cancer analysis of open-chromatin sites also revealed only attenuated effects of total mutational enrichments. Several mutational signatures as well as indels appeared depleted in sites compared to flanking regions. Contrarily, analysis of individual cancer types using cancer-specific maps of chromatin accessibility revealed robust enrichments of mutations at open-chromatin sites. This highlights the importance of using matched genomic, epigenomic and transcriptomic profiles to study localised mutational processes.

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

Our analysis has certain caveats and limitations. We analysed a broad catalogue of genomic elements that provides a limited representation of the heterogeneous dataset of tumour genomes. We used several strategies to prioritise functionally active elements in a tissue-specific manner: TSSs were grouped by target gene transcription in the matching tumour samples, open-chromatin sites were selected from genome-wide profiles of cancers of matching types or related normal tissues and CTCF binding sites were grouped by their binding conservation across a large panel of human cell types. To better address tumour heterogeneity, future analyses will benefit from detailed multi-omics cohorts where matching genomic, transcriptomic and epigenomic profiles of individual tumours are available. Also, the current method is designed for genomic elements of uniform width and it is not directly applicable to elements of variable width, such as exons or non-coding RNAs. 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 the relevant input parameter of the method when analysing new classes of genomic elements. Our method is designed to quantify localised differences of mutation frequencies 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. That said, our method can be used to evaluate localised mutation rates affecting regulatory elements of biological processes and pathways with hundreds to thousands of genes.

Our study opens new avenues for future developments. Integrative analysis of whole cancer genome sequences and rich clinical and pathological profiles of tumours [17] may highlight associations of clinical variables and localised mutagenesis and thus 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 mutational mechanisms that can be tested experimentally using genome editing and mutagenesis assays. Rare germline variants in the human population [64], de novo variants detected in genetic disorders [21] and the widespread somatic genome variation found in healthy tissues [65] provide further avenues to study mutational processes acting on functional non-coding elements. Our study provides a detailed annotation of localised mutational processes in whole genomes and enables future work to decipher the interplay of local mutagenesis and cancer driver mechanisms, molecular heterogeneity and genome evolution.

Methods

Regression models for localised mutations (RM2)

Local differences in mutation frequencies in functional genomic elements (i.e. sites) were evaluated using a negative binomial regression model we refer to as RM2. Single-nucleotide variants (SNVs) and small insertions-deletions (indels) were analysed. The model simultaneously considers a collection of sites, such as regulatory elements that are commonly ~ 10–1000 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 upstream and downstream sequences of fixed width (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 frequencies. Control regions were defined to be of equal width to sites such that the upstream and downstream regions combined were twice as wide as the sites. Next, site and flank sequences were pooled, duplicate sequences were removed, and overlapping coordinates of sites were removed from the flanking set such that every nucleotide and every mutation was considered as either as part of a site or a flanking control sequence but never both. The pooling and deduplication steps were used to account for groups of adjacent sites. To account for megabase-scale variation in mutation frequencies, we computed the total log-transformed mutation count for each site within its one-megabase window (i.e. ± 0.5 Mbps around site midpoint). Based on this estimate, all sites were distributed into ten equal bins (*MbpRate*). The value of ten bins worked well in our benchmarks and captured variation in smaller and larger cohorts of individual cancer types. However, custom values of this parameter can be used. Mutation counts for sites and flanking sequences for each bin were defined separately and were distinguished by a binary covariate (*isSite*). Sequence positions were counted separately by their trinucleotide context (*nPosits*) and expanded to three alternative nucleotides to account for the potential sequence space where such single-nucleotide variants could occur (*nPosits*). The observed mutations in these contexts were also counted (*nMuts*) and a covariate 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 covariate (*coFac*) was included to allow the consideration of genetic or clinical covariates of localised mutation frequencies. To evaluate the significance of localised mutation frequencies in sites compared to flanking control regions, we first constructed a null model that excluded the term *isSite*:

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

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

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

We extended our model to evaluate whether localised mutation frequencies differ between two subtypes of tumours, such as those defined by clinical annotations or genetic features, using the term *coFac.* Trinucleotide sequence content, trinucleotide-annotated mutations and megabase-scale covariations of mutation frequencies were computed separately for the two sets of tumours. To establish the associations of localised mutation frequencies and tumour subtypes (or presence of driver mutations), we added to the initial model the term *isSite:coFac* mapping the interaction of the tumour subtype and the covariate distinguishing sites and flanking sequences, as follows:

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

We used likelihood ratio tests to compare the models and evaluate the significance of localised mutation frequencies (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 frequencies 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 FDR procedure. As an exception, unadjusted *P* values were used to quantify the interactions of local mutation frequencies and tumour subtype (e.g. presence of driver mutation) since these analyses were conditional on the significance of main site-specific effects. Besides modelling total mutations in sites and flanking sequences, we evaluated mutations of multiple subclasses, such as mutations stratified by COSMIC mutational signatures or DNA strands. Mutation subclass analysis was conducted as described above. The same megabase-scale mutation frequencies estimated for all mutations were used rather than those of specific subclasses. The RM2 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 2583 primary tumours 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 tumours with at least 90,000 mutations, resulting in 2514 tumours and 24.7 million mutations. We also removed 33 tumours for which mutational signature predictions were not available in PCAWG. We analysed 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 tumours in which localised mutation frequencies were exceptionally strong even when analysing one tumour genome at a time (FDR < 0.001, RM2). The small subset of filtered tumours included 33 tumours (1.4% of the cohort) that provided 38 findings of elevated site-specific mutation frequencies of the three site classes. The outlier tumours showed overall higher mutation burden compared to the remaining pan-cancer cohort (mean 50,238 vs. 9404 mutations). We performed this filtering because based on our initial analyses, we found that the individual contribution of these tumours to the overall analysis would have caused overestimates of localised mutation frequencies. The filtering led to a conservative analysis that only revealed associations shared across multiple tumour genomes. To enable this filtering, we performed tumour-specific analyses for the three classes of sites (open-chromatin sites, CTCF binding sites and TSSs). We analysed each cohort of a cancer type separately and grouped the mutations according to tumour ID, allowing the model to learn expected background mutation frequencies in the respective cohort and then test each tumour genome separately for localised mutation frequencies. To perform this single-tumour analysis in smaller cohorts within the PCAWG dataset (< 25 tumours of a given type), we created a meta-cohort by pooling these smaller cohorts. After filtering hypermutated tumours, tumours without PCAWG signatures and tumours with exceptionally strong signals of localised mutations, we derived a conservative final set of 2419 genomes of 35 cancer types with 22.7 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 tumours derived from the PCAWG project (i.e. the Broad dataset) [4]. The true and simulated datasets were compared for method benchmarking (see below). 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]). Mutations were also 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. We also classified mutations by the trinucleotide signatures of single-base substitutions (SBS) that were derived earlier using the SigProfiler software in the PCAWG project [14]. We assigned each mutation to its most probable signature in the given patient tumour based on its trinucleotide context. For model evaluation, mutations were also annotated for the simulated dataset.

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 detected in human primary tumours or related normal tissues (ATAC-seq and DNAse-seq sites). CTCF binding sites were retrieved from the ENCODE project [39] and sites observed in only one cell line were removed for the majority of the study, resulting in 119,464 sites across 70 cell lines. As an exception, the full set of 162,209 CTCF binding sites in ENCODE was used for the analysis of CTCF binding sites that considered site conservation across cell lines and motifs (Fig. 4), as discussed below. 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 protein-coding genes. Open-chromatin sites were collected from ATAC-seq profiles of a TCGA study [40] and DNAse-seq profiles of the Roadmap Epigenomics project [41]. We used tissue-specific open-chromatin profiles for every cancer type, first by matching cancer types of the PCAWG project and the TCGA project (17/25 cohorts), and second, by selecting the closest normal tissues or cell lines of the Epigenomics Roadmap project for cancer types in which matching open-chromatin profiles were not available (8/25 cohorts) (Additional file 2: Table S1A). The pan-cancer analysis considered pan-cancer open-chromatin sites of the TCGA ATAC-seq study. To better compare the mutation frequencies of TSSs and open-chromatin sites, we excluded the open-chromatin sites that overlapped with the TSSs or CTCF binding sites. Sites that were originally mapped to the GRCh38 reference genome were re-mapped to GRCh37 using the LiftOver method. Throughout the study, the three classes of sites were normalised to uniform widths 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. We systematically explored various values of the site width parameters and the final selection was based on the strength of signal and consistency (Additional file 1: Figure S2A).

Defining target genes of TSSs and open-chromatin sites

TSS and open-chromatin sites were analysed in groups based on the mRNA abundance of target genes in matching tumours. TSS target genes were retrieved from the Ensembl database and target genes of open-chromatin sites were predicted using two custom strategies. First, direct target genes of open-chromatin sites were defined based on their location in gene promoters, 5′ and 3′ UTRs, and exons (open-chromatin sites directly overlapping TSSs were excluded, as specified above). Second, a promoter-capture Hi-C dataset from a recent multi-tissue study [46] was used to identify long-range chromatin interactions of open-chromatin sites and target genes. We selected high-confidence interactions (i.e. min_freq≥10) that were observed in at least 5/28 cell types to define putatively constitutive interactions of open-chromatin sites and target genes.

Grouping TSSs and open-chromatin sites by tissue-specific mRNA abundance

We used mRNA abundance in matching tumours as a proxy of site functionality of TSSs and open-chromatin sites. To group sites by mRNA abundance, we used the uniformly processed PCAWG RNA-seq dataset [45] (RPKM-UQ) that covered ~ 50% of the whole cancer genome cohort. We limited the mRNA analysis to the tumours in the WGS dataset and excluded non-coding genes. Cancer type-specific analyses were carried out in 20 cohorts of cancer types for which at least 15 tumour samples with mRNA and WGS data were available, and also the pooled pan-cancer group as the 21st cohort. We discarded genes with duplicated HGNC symbols and genes for which TSS or open-chromatin sites were not mapped. Together, this resulted in mRNA measurements for 20,042 protein-coding genes in 1267 tumour transcriptomes. Next, for each cancer type, we grouped all the genes into five bins based on their median mRNA abundance and analysed the bins using RM2 separately for TSSs and open-chromatin sites. For the pan-cancer analysis, we binned genes using median mRNA abundance in the entire RNA-seq dataset.

Grouping CTCF binding sites by the conservation of CTCF binding across human cell lines

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 a binding event of CTCF to the sites was detected. To interpret these CTCF binding sites, we used information on the three-dimensional genome and CTCF sequence motifs. First, we retrieved chromatin loops in eight cell lines from a Hi-C study [48], used a ± 1000 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). Second, using the motif MA0139.1 from the JASPAR database [66], we detected the presence of the core CTCF motif in CTCF binding sites using the FIMO method [67] and selected significant motif instances ($FDR < 0.01$). Enrichment of CTCF motifs in the 5th bin of sites was also identified using a Fisher's exact test.

Analysis of localised mutation frequencies 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 for the pan-cancer cohort and all cohorts of selected cancer types. Total mutations and mutations grouped by COSMIC signatures 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 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.

Associating mutations in TSSs and open-chromatin sites with transcript abundance

To study the associations of mRNA abundance and mutations in TSSs and open-chromatin sites, we grouped the sites by mRNA abundance of genes in matching tumour types, and then performed the RM2 analysis separately for each bin and cancer type. Since RM2 is better powered to discover localised mutation frequencies in larger groups of sites, we used down-sampling of TSSs and open-chromatin sites in the mRNA-defined gene bins to provide better comparisons. In each bin and cancer type, we randomly sampled 4500 sites for RM2 analysis, repeated this procedure over 100 iterations and selected the RM2 result with the median P value for reporting. The number for down-sampling was chosen as the closest smallest common value for bin sizes for TSSs and open-chromatin sites. The results were adjusted for multiple testing and filtered (FDR < 0.05).

Identifying pathways with frequent mutations in transcription start sites

We asked whether the localised changes in mutation frequencies of TSSs affected specific biological processes and pathways. We repurposed the RM2 model to analyse TSSs of gene sets corresponding to biological processes of Gene Ontology [68] and molecular pathways of the Reactome database [69]. Gene sets were derived from the g:Profiler [70] web server (March 3rd, 2020) and subsequently filtered to include 1871 gene sets with 100 to 1000 genes while smaller and larger gene sets were excluded. Results were corrected for multiple testing for every cancer type and filtered for statistical significance (FDR < 0.05). The pathways with significantly higher TSS-specific mutation frequencies were visualised as an enrichment map [71] in Cytoscape and major biological themes were manually curated as described earlier [47]. Nodes in the enrichment map were painted to reflect cancer types where these pathway enrichments were detected according to the custom colour scheme of the PCAWG project.

Associating elevated mutation frequencies in CTCF binding sites with constitutive CTCF binding and core sequence motifs

To study the functional associations of localised mutation frequencies at CTCF binding sites, we first analysed five groups of CTCF binding sites defined based on the counts of cell lines where the sites were observed. These bins, ranging from sites of single cell lines to sites constitutively active in all or most cell lines, were analysed in all cohorts of individual cancer types and the pan-cancer cohort. Findings were corrected for multiple testing and filtered to select significant results (FDR < 0.05). To perform the motif-based analysis, we first split sites into two groups based on a significant match to the core CTCF DNA-binding motif using the FIMO method [67] (FDR < 0.01). Then, we repeated the binning of sites by shared activity, as described above, to ensure all bins of sites within each group (i.e. sites with or without the motif) were of equal size. The ten bins of CTCF sites were also analysed using RM2. Lastly, to perform a tissue-specific analysis of matching cancer types and cell lines, we selected CTCF binding sites for two cancer cell lines in the ENCODE dataset: MCF-7, a breast cancer cell line, and HepG2, a liver cancer cell line. In this analysis, the CTCF binding sites found in at least the given cell line were also split into ten bins based on conservation of CTCF binding and presence of the core sequence motif, and then analysed with RM2 using mutations of the PCAWG dataset for the Breast-AdenoCa cohort (with MCF-7) and Liver-HCC cohort (with HepG2).

Associating elevated mutation frequencies in sites with driver mutations and recurrent copy-number alterations

We tested whether the localised changes in mutation frequencies in CTCF binding sites, TSSs and open-chromatin sites were associated with driver mutations (i.e. SNVs, indels) or recurrent copy-number alterations (CNAs). 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 method [34]. For driver analysis, we used the PCAWG variant calls after filtering tumours as described above, corrected the results for multiple testing 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. Tumours with and without SNVs or indels in predicted driver genes were used for the localised mutation frequency analysis. Predictions of recurrent CNAs were derived from the pan-cancer dataset of GISTIC2 calls of the PCAWG project [53]. All lesions at 95% confidence scores were considered and amplifications and deletions were analysed separately. High-confidence CNA events were used (GISTIC2 score = 2). Tumours with and without CNAs in the recurrently altered regions were compared in the analysis using RM2. Two additional chromosomal features were used. First, we grouped the tumours by presence or absence of Whole Genome Duplication (WGD) as determined in the PCAWG project [53]. Second, we computed the Percent Genome Altered (PGA) metric for every tumour as a proxy of aneuploidy. PGA was computed as the percentage of the autosomal genome affected by CNA segments whose total copy number deviated from the global reference (two copies for non-WGD genomes; four copies for WGD genomes). PGA metrics were median-dichotomised separately for every cancer type, resulting in two subgroups of tumours: the PGA-high group and the PGA-low group corresponding to above-median and below median CNA burden of tumours. Following the construction of these genetic features (i.e. driver gene mutations, recurrent CNAs, WGD, PGA-high), we filtered overly frequent (more than 2/3 of the cohort) and infrequent features (less than 15 tumours) to improve the power of the RM2 analysis. For each cancer type, the relevant features were then analysed for localised elevations in mutation frequencies in the three categories of genomic elements (open-chromatin sites, CTCF binding sites, TSSs). The binary covariate in RM2 was used to indicate the presence or absence of the given feature in the given tumour genome. We first computed the significance of site-specific localised mutation frequencies given the presence or absence of the genetic feature. To validate this joint model, subgroups of tumours with and without the defining feature were also analysed separately using RM2. All combined RM2 results concerning individual genetic features, cancer types and genomic sites were then adjusted for multiple testing correction and significant results were selected (FDR < 0.05). We filtered the results to only include positive and significant interactions of genetic features and elevated localised mutation frequencies (interaction $P < 0.05$, main and interaction coefficients > 0) and compared the FDR values, fold-change values and visualisations of localised mutation frequencies to validate these findings.

Associating mutagenesis-related copy-number amplifications with mRNA abundance

To evaluate the functional role of recurrent copy-number amplifications associated with increased mutation frequencies in the gene-regulatory and chromatin architectural sites, we studied the genes located in the amplified regions and retrieved these from the PCAWG GISTIC2 dataset. We compared the mRNA abundance levels of the genes between groups of tumours defined by the presence or absence of the amplifications, using matching RNA-seq data available in PCAWG [45]. Genes with low mRNA abundance were removed from the analysis (median FPKM-UQ < 1). mRNA abundance levels of genes in amplified and non-amplified tumours were compared using one-sided non-parametric Wilcoxon tests, assuming that an increase in mRNA abundance would match the underlying copy-number amplification. The non-amplified group included tumours with balanced and deleted status of the CNA region. Results were adjusted for multiple testing and significant results were selected (FDR < 0.05). We ranked the resulting genes based on the total significance of CNA/mRNA associations across all cancer types and site types by summing the negative log10-transformed significant FDR values of each gene. Known cancer genes of the COSMIC Cancer Gene Census database [72] (v91, downloaded 14.05.2020) were highlighted.

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 tumour genomes using trinucleotide-informed shuffling of mutations [4]. Simulated variant calls were analysed similarly to true variant calls for total mutation counts, reference and alternative nucleotide combinations and predicted mutational signatures. Results from RM2 were adjusted for multiple testing separately for results derived from true and simulated variant calls. As expected, the simulated variant calls revealed no statistically significant results of localised mutation frequencies 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 systematically down-sampled the input data for RM2 by randomly selecting subsets of sites and tumours. 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, …, 250, 300 genomes sampled), as well as the full datasets of sites and genomes. Each configuration was tested 100 times with different random subsets of data. As 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 frequencies were consistently identified at multiple values of the width parameter for each class of site (open-chromatin sites, CTCF binding sites and 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 differences in the underlying mutational processes. 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. We evaluated the effect of grouping sites by their megabase-scale mutation frequencies (i.e. the MbpRate covariate of RM2). We repeated the systematic analysis using a simplified RM2 model that excluded the MbpRate covariate and compared the resulting *P* values and FDR values derived from the original and simplified models. The original model (including MbpRate) outperformed the simplified model (excluding MbpRate) by assigning stronger significance to findings in the dataset of true mutations and reduced significance to findings in the simulated dataset, indicating that accounting for megabase-scale mutation frequency as covariate improves sensitivity and specificity of the model.

Supplementary Information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-021-02318-x.

Additional file 1: Figure S1. Dataset of whole cancer genomes and filtering of outliers. **Figure S2.** Evaluating the site width parameter of RM2 and analysis of all TSSs and open-chromatin sites grouped by mRNA abundance of target genes. **Figure S3.** Enrichments of specific mutational signatures localised in active TSSs and CTCF binding sites. **Figure S4.** Comparison of local mutation frequencies in CTCF binding sites that include or lack a core CTCF DNA-binding sequence motif. **Figure S5.** Driver gene mutations and recurrent CNAs used in the RM2 analysis. **Figure S6.** Extended analysis of local mutation frequencies and mutations in *ARID1A* and *BRAF*. **Figure S7.** Additional interactions with CNAs and local mutation frequencies.

Additional file 2: Table S1A. Description of the open-chromatin sites of primary tumours and related normal tissues used in the analysis. **Table S1B.** Localised mutation frequencies in three classes of genomic sites across total mutations and subsets of mutations (total mutations, SBS signatures, ref./alt pairs). **Table S1C.** Localised mutation frequencies in TSSs and open-chromatin sites binned by mRNA abundance of target genes in matching tumour transcriptomes.**Table S1D.** Biological processes and molecular pathways with elevated mutation frequencies in TSSs. **Table S1E.** Localised mutation frequencies in CTCF sites grouped by conservation of CTCF binding in multiple cell lines. **Table S1F.** Localised mutation frequencies in CTCF sites grouped by motif presence and conservation of CTCF binding in multiple cell lines. **Table S1G.** List of hypermutated CTCF sites with constitutive binding and motifs (GRCh37). Table S1H. Interactions of genetic features (driver mutations, CNAs, genomic instability) and localised mutation frequencies. **Table S1I.** Genes in the identified CNA regions for which mRNA abundance increase is associated with copy number amplifications.

Additional file 3. Review history.

Acknowledgements

We would like to thank Dr. Liis Uusküla-Reimand and the reviewers of the manuscript for constructive comments. We acknowledge the contributions of the many clinical networks across ICGC and TCGA who provided samples and data to the PCAWG Consortium, and the contributions of the Technical Working Group and the Germline Working Group of the PCAWG Consortium for collation, realignment and harmonised variant calling of the cancer genomes used in this study. We thank the patients and their families for their participation in the individual ICGC and TCGA projects. The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Review history

The review history is available as Additional file 3.

Peer review information

Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' 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.

Funding

This work was supported by a Project Grant from the Canadian Institutes of Health Research (CIHR) to J.R., an Operating Grant from the Cancer Research Society to J.R., New Investigator Award from the Terry Fox Research Institute (TFRI) to J.R. and the Investigator Award to J.R. from the Ontario Institute for Cancer Research (OICR). C.A.L. was partially supported by a Graduate Student Fellowship from the Department of Medical Biophysics, University of Toronto. Funding to OICR is provided by the Government of Ontario.

Availability of data and materials

The RM2 method is available as an open-source software package on the GitHub repository (https://github.com/reimandlab/RM2) [74]. The original version of RM2 used for this publication is available in the Zenodo repository under a Creative Commons Attribution 4.0 International licence (https://doi.org/10.5281/zenodo.4530813) [75]. Somatic variant calls, transcript abundance and other core data generated by the ICGC/TCGA Pan-cancer Analysis of Whole Genomes Consortium are described in the marker paper [3] and available for download at the ICGC Data Portal (https://dcc.icgc.org/releases/PCAWG) [76]. Additional information on accessing the data, including raw read files, can be found at https://docs.icgc.org/pcawg/data/. In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier which does not require access approval. To access potentially identifiable information, such as germline alleles and underlying sequencing data, researchers will need to apply to the TCGA Data Access Committee (DAC) via dbGaP (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login) for access to the TCGA portion of the dataset, and to the ICGC Data Access Compliance Office (DACO; https://icgc.org/daco) for the ICGC portion. In addition, to access somatic single-nucleotide variants derived from TCGA donors, researchers will also need to obtain dbGaP authorization.

Declarations

Ethics approval and consent to participate

Sequencing of human subjects' tissue was performed by ICGC and TCGA consortium members under a series of locally approved Institutional Review Board (IRB) protocols as described earlier [73]. Informed consent was obtained from all human participants. Ethical review of the current data analysis project was granted by the University of Toronto Research Ethics Board (REB) under protocol #37521.

Competing interests

The authors declare that they have no competing interests.

Author details

[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.

Received: 5 August 2020 Accepted: 19 March 2021
Published online: 03 May 2021

References

1. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009;458(7239):719–24. https://doi.org/10.1038/nature07943.
2. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science. 2013; 339(6127):1546–58. https://doi.org/10.1126/science.1235122.
3. ICGC-TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020; 578:82–93. https://doi.org/10.1038/s41586-020-1969-6.
4. Rheinbay E, et al. Analyses of non-coding somatic drivers in 2,693 cancer whole genomes. Nature. 2020;578(7793):102–11. https://doi.org/10.1038/s41586-020-1965-x.
5. Bailey MH, et al. Comprehensive characterization of Cancer driver genes and mutations. Cell. 2018;174(4):1034–5. https://doi.org/10.1016/j.cell.2018.07.034.
6. Kumar S, et al. Passenger mutations in more than 2,500 cancer genomes: overall molecular functional impact and consequences. Cell. 2020;180:915–927 e916. https://doi.org/10.1016/j.cell.2020.01.032.
7. Gerstung M, Jolly C, Leshchiner I, et al. The evolutionary history of 2,658 cancers. Nature. 2020;578:122–8. https://doi.org/10.1038/s41586-019-1907-7.
8. Supek F, Lehner B. Scales and mechanisms of somatic mutation rate variation across the human genome. DNA Repair (Amst). 2019:102647. https://doi.org/10.1016/j.dnarep.2019.102647.
9. Lawrence MS, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013; 499(7457):214–8. https://doi.org/10.1038/nature12213.
10. Reijns MAM, Kemp H, Ding J, Marion de Procé S, Jackson AP, Taylor MS. Lagging-strand replication shapes the mutational landscape of the genome. Nature. 2015;518(7540):502–6. https://doi.org/10.1038/nature14183.
11. Schuster-Bockler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature. 2012;488(7412):504–7. https://doi.org/10.1038/nature11273.
12. Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence MS, Reynolds A, Rynes E, Vlahoviček K, Stamatoyannopoulos JA, Sunyaev SR. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature. 2015;518(7539):360–4. https://doi.org/10.1038/nature14221.
13. Alexandrov LB, et al. Signatures of mutational processes in human cancer. Nature. 2013;500(7463):415–21. https://doi.org/10.1038/nature12477.
14. Alexandrov LB, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578(7793):94–101. https://doi.org/10.1038/s41586-020-1943-3.
15. Kucab JE, et al. A Compendium of Mutational Signatures of Environmental Agents. Cell. 2019;177:821–836 e816. https://doi.org/10.1016/j.cell.2019.03.001.
16. Pich O, Muiños F, Lolkema MP, Steeghs N, Gonzalez-Perez A, Lopez-Bigas N. The mutational footprints of cancer therapies. Nat Genet. 2019;51(12):1732–40. https://doi.org/10.1038/s41588-019-0525-5.
17. Priestley P, Baber J, Lolkema MP, Steeghs N, de Bruijn E, Shale C, Duyvesteyn K, Haidari S, van Hoeck A, Onstenk W, Roepman P, Voda M, Bloemendal HJ, Tjan-Heijnen VCG, van Herpen CML, Labots M, Witteveen PO, Smit EF, Sleijfer S,

Voest EE, Cuppen E. Pan-cancer whole-genome analyses of metastatic solid tumours. Nature. 2019;575(7781):210–6. https://doi.org/10.1038/s41586-019-1689-y.

18. Kübler K, et al. Tumor mutational landscape is a record of the pre-malignant state. bioRxiv. 2019:517565. https://doi.org/10.1101/517565.
19. Jiao W, et al. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. Nat Commun. 2020;11(1):728. https://doi.org/10.1038/s41467-019-13825-8.
20. Gonzalez-Perez A, Sabarinathan R, Lopez-Bigas N. Local determinants of the mutational landscape of the human genome. Cell. 2019;177(1):101–14. https://doi.org/10.1016/j.cell.2019.02.051.
21. Kaiser VB, Taylor MS, Semple CA. Mutational biases drive elevated rates of substitution at regulatory sites across cancer types. PLoS Genet. 2016;12(8):e1006207. https://doi.org/10.1371/journal.pgen.1006207.
22. Yazdi PG, Pedersen BA, Taylor JF, Khattab OS, Chen YH, Chen Y, Jacobsen SE, Wang PH. Increasing nucleosome occupancy is correlated with an increasing mutation rate so long as DNA repair machinery is intact. PLoS One. 2015; 10(8):e0136574. https://doi.org/10.1371/journal.pone.0136574.
23. Hara R, Mo J, Sancar A. DNA damage in the nucleosome core is refractory to repair by human excision nuclease. Mol Cell Biol. 2000;20(24):9173–81. https://doi.org/10.1128/mcb.20.24.9173-9181.2000.
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. 2016;532(7598):264–7. https://doi.org/10.1038/nature17661.
25. Perera D, Poulos RC, Shah A, Beck D, Pimanda JE, Wong JWH. Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. Nature. 2016;532(7598):259–63. https://doi.org/10.1038/nature17437.
26. Hnisz D, Weintraub AS, Day DS, Valton AL, Bak RO, Li CH, Goldmann J, Lajoie BR, Fan ZP, Sigova AA, Reddy J, Borges-Rivera D, Lee TI, Jaenisch R, Porteus MH, Dekker J, Young RA. Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science. 2016;351(6280):1454–8. https://doi.org/10.1126/science.aad9024.
27. Katainen R, Dave K, Pitkänen E, Palin K, Kivioja T, Välimäki N, Gylfe AE, Ristolainen H, Hänninen UA, Cajuso T, Kondelin J, Tanskanen T, Mecklin JP, Järvinen H, Renkonen-Sinisalo L, Lepistö A, Kaasinen E, Kilpivaara O, Tuupanen S, Enge M, Taipale J, Aaltonen LA. CTCF/cohesin-binding sites are frequently mutated in cancer. Nat Genet. 2015;47(7):818–21. https://doi.org/10.1038/ng.3335.
28. Guo YA, Chang MM, Huang W, Ooi WF, Xing M, Tan P, Skanderup AJ. Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. Nat Commun. 2018;9(1):1520. https://doi.org/10.1038/s41467-018-03828-2.
29. Polak P, Lawrence MS, Haugen E, Stoletzki N, Stojanov P, Thurman RE, Garraway LA, Mirkin S, Getz G, Stamatoyannopoulos JA, Sunyaev SR. Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. Nat Biotechnol. 2014;32(1):71–5. https://doi.org/10.1038/nbt.2778.
30. Frigola J, Sabarinathan R, Mularoni L, Muiños F, Gonzalez-Perez A, López-Bigas N. Reduced mutation rate in exons due to differential mismatch repair. Nat Genet. 2017;49(12):1684–92. https://doi.org/10.1038/ng.3991.
31. Bell RJ, et al. The transcription factor GABP selectively binds and activates the mutant TERT promoter in cancer. Science. 2015;348(6238):1036–9. https://doi.org/10.1126/science.aab0015.
32. Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA. Highly recurrent TERT promoter mutations in human melanoma. Science. 2013;339(6122):957–9. https://doi.org/10.1126/science.1229259.
33. Fredriksson NJ, Ny L, Nilsson JA, Larsson E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. Nat Genet. 2014;46(12):1258–63. https://doi.org/10.1038/ng.3141.
34. Zhu H, Uusküla-Reimand L, Isaev K, Wadi L, Alizada A, Shuai S, Huang V, Aduluso-Nwaobasi D, Paczkowska M, Abd-Rabbo D, Ocsenas O, Liang M, Thompson JD, Li Y, Ruan L, Krassowski M, Dzneladze I, Simpson JT, Lupien M, Stein LD, Boutros PC, Wilson MD, Reimand J. Candidate cancer driver mutations in distal regulatory elements and long-range chromatin interaction networks. Mol Cell. 2020;77(6):1307–1321.e10. https://doi.org/10.1016/j.molcel.2019.12.027.
35. Corona RI, Seo JH, Lin X, Hazelett DJ, Reddy J, Fonseca MAS, Abassi F, Lin YG, Mhawech-Fauceglia PY, Shah SP, Huntsman DG, Gusev A, Karlan BY, Berman BP, Freedman ML, Gayther SA, Lawrenson K. Non-coding somatic mutations converge on the PAX8 pathway in ovarian cancer. Nat Commun. 2020;11(1):2020. https://doi.org/10.1038/s41467-020-15951-0.
36. Liu EM, et al. Identification of Cancer Drivers at CTCF Insulators in 1,962 Whole Genomes. Cell Syst. 2019;8:446–455 e448. https://doi.org/10.1016/j.cels.2019.04.001.
37. Reyna MA, et al. Pathway and network analysis of more than 2,500 whole cancer genomes. Nat Commun. 2020;11(1): 729. https://doi.org/10.1038/s41467-020-14367-0.
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. 2019;36:674–689 e676. https://doi.org/10.1016/j.ccell.2019.10.005.
39. Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74. https://doi.org/10.1038/nature11247.
40. Corces MR, et al. The chromatin accessibility landscape of primary human cancers. Science. 2018;362 https://doi.org/10.1126/science.aav1898.
41. Roadmap Epigenomics Consortium, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518: 317–30. https://doi.org/10.1038/nature14248.
42. Dvorak K, Payne CM, Chavarria M, Ramsey L, Dvorakova B, Bernstein H, Holubec H, Sampliner RE, Guy N, Condon A, Bernstein C, Green SB, Prasad A, Garewal HS. Bile acids in combination with low pH induce oxidative stress and oxidative DNA damage: relevance to the pathogenesis of Barrett's oesophagus. Gut. 2007;56(6):763–71. https://doi.org/10.1136/gut.2006.103697.
43. Tomkova M, Tomek J, Kriaucionis S, Schuster-Bockler B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. Genome Biol. 2018;19(1):129. https://doi.org/10.1186/s13059-018-1509-y.
44. Dellino GI, Palluzzi F, Chiariello AM, Piccioni R, Bianco S, Furia L, de Conti G, Bouwman BAM, Melloni G, Guido D, Giacò L, Luzi L, Cittaro D, Faretta M, Nicodemi M, Crosetto N, Pelicci PG. Release of paused RNA polymerase II at specific loci favors DNA double-strand-break formation and promotes cancer translocations. Nat Genet. 2019;51(6):1011–23. https://doi.org/10.1038/s41588-019-0421-z.

45. PCAWG Transcriptome Core Group, Calabrese C, Davidson NR, et al. Genomic basis for RNA alterations in cancer. Nature. 2020;578:129–36. https://doi.org/10.1038/s41586-020-1970-0.
46. Jung I, Schmitt A, Diao Y, Lee AJ, Liu T, Yang D, Tan C, Eom J, Chan M, Chee S, Chiang Z, Kim C, Masliah E, Barr CL, Li B, Kuan S, Kim D, Ren B. A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat Genet. 2019;51(10):1442–9. https://doi.org/10.1038/s41588-019-0494-8.
47. Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, Wadi L, Meyer M, Wong J, Xu C, Merico D, Bader GD. Pathway enrichment analysis and visualization of omics data using g:profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc. 2019;14(2):482–517. https://doi.org/10.1038/s41596-018-0103-9.
48. Rao SS, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014; 159(7):1665–80. https://doi.org/10.1016/j.cell.2014.11.021.
49. Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, Tsutsumi S, Nagae G, Ishihara K, Mishiro T, Yahata K, Imamoto F, Aburatani H, Nakao M, Imamoto N, Maeshima K, Shirahige K, Peters JM. Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature. 2008;451(7180):796–801. https://doi.org/10.1038/nature06634.
50. Kaiser VB, Semple CA. Chromatin loop anchors are associated with genome instability in cancer and recombination hotspots in the germline. Genome Biol. 2018;19(1):101. https://doi.org/10.1186/s13059-018-1483-4.
51. Pinoli P, Stamoulakatou E, Nguyen AP, Rodriguez Martinez M, Ceri S. Pan-cancer analysis of somatic mutations and epigenetic alterations in insulated neighbourhood boundaries. PLoS One. 2020;15(1):e0227180. https://doi.org/10.1371/journal.pone.0227180.
52. Khoury A, Achinger-Kawecka J, Bert SA, Smith GC, French HJ, Luu PL, Peters TJ, du Q, Parry AJ, Valdes-Mora F, Taberlay PC, Stirzaker C, Statham AL, Clark SJ. Constitutively bound CTCF sites maintain 3D chromatin architecture and long-range epigenetically regulated domains. Nat Commun. 2020;11(1):54. https://doi.org/10.1038/s41467-019-13753-7.
53. Li Y, Roberts ND, Wala JA, et al. Patterns of somatic structural variation in human cancer genomes. Nature. 2020;578: 112–21. https://doi.org/10.1038/s41586-019-1913-9.
54. Wu RC, Wang TL, Shih Ie M. The emerging roles of ARID1A in tumor suppression. Cancer Biol Ther. 2014;15(6):655–64. https://doi.org/10.4161/cbt.28411.
55. Dykhuizen EC, Hargreaves DC, Miller EL, Cui K, Korshunov A, Kool M, Pfister S, Cho YJ, Zhao K, Crabtree GR. BAF complexes facilitate decatenation of DNA by topoisomerase IIalpha. Nature. 2013;497(7451):624–7. https://doi.org/10.1038/nature12146.
56. Park JH, Park EJ, Lee HS, Kim SJ, Hur SK, Imbalzano AN, Kwon J. Mammalian SWI/SNF complexes facilitate DNA double-strand break repair by promoting gamma-H2AX induction. EMBO J. 2006;25(17):3986–97. https://doi.org/10.1038/sj.emboj.7601291.
57. Zhao Q, et al. Modulation of nucleotide excision repair by mammalian SWI/SNF chromatin-remodeling complex. J Biol Chem. 2009;284:30424–32. https://doi.org/10.1074/jbc.M109.044982.
58. Trizzino M, Barbieri E, Petracovici A, Wu S, Welsh SA, Owens TA, Licciulli S, Zhang R, Gardini A. The tumor suppressor ARID1A controls global transcription via pausing of RNA polymerase II. Cell Rep. 2018;23(13):3933–45. https://doi.org/10.1016/j.celrep.2018.05.097.
59. The Cancer Geome Atlas Network. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. Cancer Cell. 2017;32:185–203 e113. https://doi.org/10.1016/j.ccell.2017.07.007.
60. The Cancer Genome Atlas Network. Genomic classification of cutaneous melanoma. Cell. 2015;161:1681–96. https://doi.org/10.1016/j.cell.2015.05.044.
61. Chapman PB, Hauschild A, Robert C, Haanen JB, Ascierto P, Larkin J, Dummer R, Garbe C, Testori A, Maio M, Hogg D, Lorigan P, Lebbe C, Jouary T, Schadendorf D, Ribas A, O'Day SJ, Sosman JA, Kirkwood JM, Eggermont AMM, Dreno B, Nolop K, Li J, Nelson B, Hou J, Lee RJ, Flaherty KT, McArthur GA. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. N Engl J Med. 2011;364(26):2507–16. https://doi.org/10.1056/NEJMoa1103782.
62. Sheu JJ, et al. Mutant BRAF induces DNA strand breaks, activates DNA damage response pathway, and up-regulates glucose transporter-1 in nontransformed epithelial cells. Am J Pathol. 2012;180(3):1179–88. https://doi.org/10.1016/j.ajpath.2011.11.026.
63. Fredriksson NJ, Elliott K, Filges S, van den Eynden J, Ståhlberg A, Larsson E. Recurrent promoter mutations in melanoma are defined by an extended context-specific mutational signature. PLoS Genet. 2017;13(5):e1006773. https://doi.org/10.1371/journal.pgen.1006773.
64. Karczewski KJ, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020; 581(7809):434–43. https://doi.org/10.1038/s41586-020-2308-7.
65. Martincorena I, Roshan A, Gerstung M, Ellis P, van Loo P, McLaren S, Wedge DC, Fullam A, Alexandrov LB, Tubio JM, Stebbings L, Menzies A, Widaa S, Stratton MR, Jones PH, Campbell PJ. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science. 2015;348(6237):880–6. https://doi.org/10.1126/science.aaa6806.
66. Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D, Santana-Garcia W, Tan G, Chèneby J, Ballester B, Parcy F, Sandelin A, Lenhard B, Wasserman WW, Mathelier A. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2020;48: D87–92. https://doi.org/10.1093/nar/gkz1001.
67. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27(7):1017–8. https://doi.org/10.1093/bioinformatics/btr064.
68. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9. https://doi.org/10.1038/75556.
69. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, Milacic M, Roca CD, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Viteri G, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018;46(D1):D649–55. https://doi.org/10.1093/nar/gkx1132.

70. 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 Res. 2007;35(suppl_2):W193–200. https://doi.org/10.1093/nar/gkm226.
71. Merico D, Isserlin R, Stueker O, Emili A, Bader GD. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PLoS One. 2010;5(11):e13984. https://doi.org/10.1371/journal.pone.0013984.
72. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. Nat Rev Cancer. 2004;4(3):177–83. https://doi.org/10.1038/nrc1299.
73. Hudson TJ, et al. International network of cancer genome projects. Nature. 2010;464(7291):993–8. https://doi.org/10.1038/nature08987.
74. Reimand J. RM2 - regression models for localised mutations. GitHub, https://github.com/reimandlab/RM2 (2021).
75. Reimand J. RM2 - regression models for localised mutations. Zenodo, https://doi.org/10.5281/zenodo.4530813 (2021).
76. ICGC/TCGA Pan-cancer Analysis of Whole Genomes Consortium. Pan-Cancer Analysis of Whole Genomes. ICGC Data Portal, https://dcc.icgc.org/releases/PCAWG (2020).

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.