# Chromatin accessibility of primary human cancers ties regional mutational processes with tissues of origin

Oliver Ocsenas[1,2], Jüri Reimand[1,2,3,@]

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

ABSTRACT:

Regional mutagenesis in cancer genomes associates with DNA replication timing (RT) and chromatin accessibility (CA) of normal cells, however human cancer epigenomes remain uncharacterized in this context. Here we model megabase-scale mutation frequencies in 2517 cancer genomes with 773 CA and RT profiles of cancers and normal cells. We find that CA profiles of matching cancers, rather than normal cells, predict regional mutagenesis and mutational signatures, indicating that most passenger mutations follow the epigenetic landscapes of transformed cells. Carcinogen-induced and unannotated signatures show the strongest associations with epigenomes. Associations with normal cells in melanomas, lymphomas and SBS1 signatures suggest earlier occurrence of mutations in cancer evolution. Frequently mutated regions unexplained by CA and RT are enriched in cancer genes and developmental pathways, reflecting contributions of localized mutagenesis and positive selection. These results underline the complex interplay of mutational processes, genome function and evolution in cancer and tissues of origin.

## INTRODUCTION

The cancer genome is a footprint of its evolution and molecular environment that is shaped by somatic mutations, such as single nucleotide variants (SNVs) and structural alterations (1,2). While a minority of mutations called drivers confer cells with selective advantages (3-5), most mutations are considered functionally neutral passengers that are caused by diverse mutational processes (6-8). Somatic mutagenesis and positive selection of known cancer genes also affects normal tissues (9,10). Characterizing the landscape of somatic mutations helps understand the underlying mutational processes and better evaluate the functional consequences of mutations and their roles in cancer etiology and evolution.

Processes of somatic mutagenesis act at different scales of the genome (11,12). At the trinucleotide resolution, mutational signatures of SNVs are associated with endogenous and exogeneous processes related to aging, carcinogen exposures, DNA repair deficiencies, and cancer therapies (6,13). At the local resolution of 100-1000 bps, non-coding genomic elements, such as transcription start sites and binding sites of CTCF, are enriched in mutations (14-16). However, the precise molecular mechanisms driving these mutational processes remain uncharacterized. At the regional, megabase-scale resolution of the genome, variation in mutation frequencies shows a complex interplay of DNA replication timing (RT), chromatin accessibility (CA) and transcriptional activity (17-19). Early-replicating, transcriptionally active regions of open chromatin have fewer mutations than late-replicating, passive regions of heterochromatin, potentially due to increased error rates and decreased mismatch repair later in DNA replication (20-23). Mutational signatures are distributed asymmetrically with respect to DNA replication origins and timing (24). Regional mutagenesis has been associated with epigenetic information of related normal cells, providing evidence of cells of cancer origin contributing to somatic variation (25) and allowing classification of cancers of unknown origin (26). However, CA and RT profiles of only common cell lines and normal tissues have been used to characterize regional mutational processes while the epigenetic landscapes of primary human cancers remain unexplored.

To decipher regional mutational processes in the context of cancer epigenomes, we analyzed a large and diverse collection of CA and RT profiles of cancers, normal tissues, and cell lines as predictors of regional mutagenesis in thousands of whole cancer genomes using machine learning. CA profiles of matching cancer types, rather than normal tissues, appear as determinants of regional mutagenesis and mutational signatures. We found tissue-of-origin effects of CA and RT in most predictions, bespoke deviations in specific cancer types and mutational signatures, and a convergence of excess mutations to developmental and cancer pathways. Together, these results underline the spatial and temporal complexity of regional mutagenesis in cancer genomes.

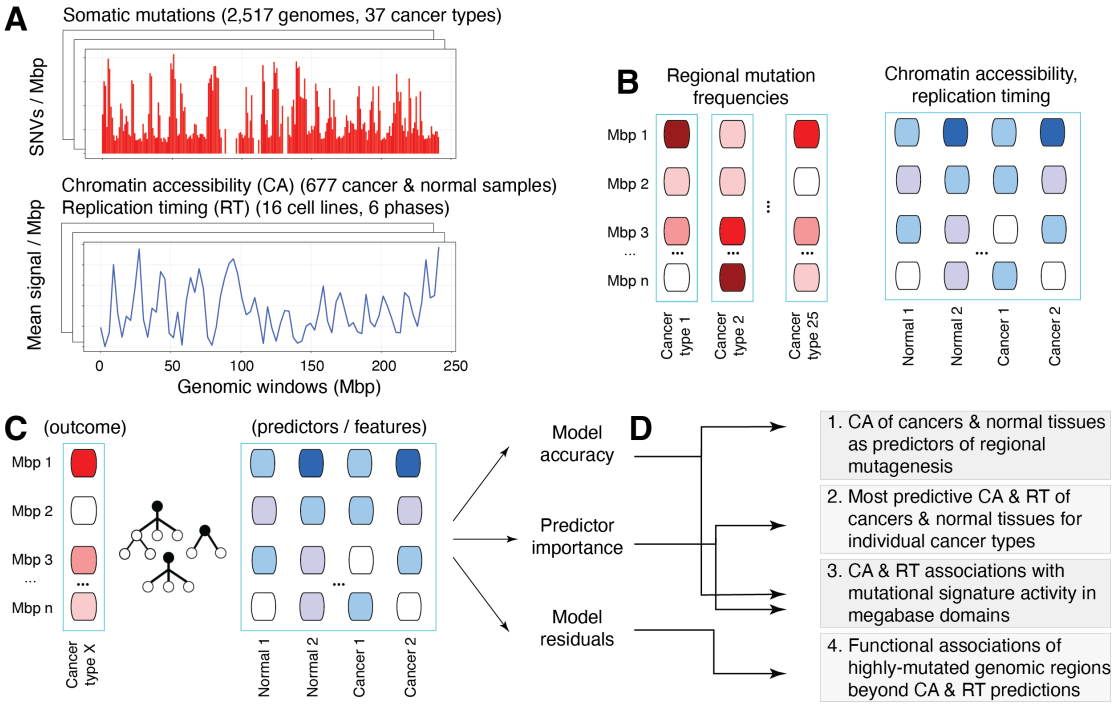


**Figure 1. Characterizing chromatin accessibility (CA) and replication timing (RT) as determinants of regional mutagenesis in cancer genomes. A.** Somatic mutations in cancer genomes (top) and CA and RT datasets of normal tissues and cancers (bottom) were integrated to study regional mutational processes. Somatic single nucleotide variants (SNVs) of 2,517 whole cancer genomes of the PCAWG project were analyzed with 677 genome-wide CA profiles of primary human cancers, normal tissues and cell lines, and 96 RT profiles of cell lines and cell cycle phases. **B.** SNVs were aggregated by summing in 2,465 high-confidence genomic regions of one megabase (Mbp) as a measure of regional mutagenesis. Mean CA and RT scores per Mbp were derived for all profiles. **C.** Random forest models were trained using regional mutagenesis data as the outcome and CA and RT profiles as predictors. 25 cancer types of the largest cohorts in PCAWG and the pan-cancer cohort were analyzed. **D.** To associate regional mutagenesis with CA and RT, mutational signatures and gene function, models were evaluated in terms of accuracy, predictor importance, and model residuals.

## RESULTS

### Chromatin accessibility of primary cancers is a major determinant of regional mutagenesis

To evaluate the associations of CA and RT with regional mutagenesis in cancer genomes, we analyzed somatic variant calls of whole cancer genomes, 677 CA profiles of primary human cancers, normal tissues and cell lines, and RT profiles of 16 cell lines in 6 cell cycle phases using the random forest framework (**Figure 1**). We integrated 23 million SNVs from 2,517 whole cancer genomes spanning 37 types of the ICGC/TCGA PCAWG project (1) with CA and RT profiles collected from The Cancer Genome Atlas (TCGA), Epigenomics Roadmap and ENCODE3 projects (27-30) (**Supplementary Figures 1-2**). Focusing on 2,465 mappable one-megabase regions, we derived somatic SNV counts for the pan-cancer dataset and 25 cancer types of the largest cohorts. We processed 773 CA and RT profiles of primary cancers, normal tissues, and cell lines as mean genomic signals per megabase (**Figure 1B**). To map the complex non-linear associations of CA and RT with regional mutagenesis, random forest regression

models were trained with megabase-scale mutation frequencies as outcomes and CA and RT profiles as predictors (*i.e.*, features) (**Figure 1C**). The most informative predictors were quantified statistically and using local prioritization methods (31) (**Figure 1D**). As expected, late RT profiles inversely correlated with CA profiles and regional mutagenesis profiles clustered according to cancer types (**Supplementary Figures 3-4**).

Given the diverse collection of epigenomic profiles, we asked whether CA profiles of cancers (n = 383) or normal cells and tissues (n = 244) were more informative of regional mutagenesis in cancer genomes. We predicted regional mutation frequencies using random forests in two Monte-Carlo cross-validation experiments with matched data splits where the predictors included either cancer CA profiles or normal CA profiles, respectively. RT profiles were also included in both models to focus on the relative contributions of CA profiles. We found that CA profiles of cancers were more accurate predictors of regional mutagenesis in 19/25 cancer types (empirical $P < 0.01$) (**Figure 2A**). The strongest signal was observed in breast cancer where the predictions informed by cancer CA profiles were nearly twice as accurate as the models informed by normal tissue CA (median adj.$R^2$ 0.69 *vs.* 0.36; $P < 0.001$) (**Figure 2B**). Stronger associations of cancer CA profiles and regional mutagenesis were also found in cancers of the prostate, ovary, uterus, kidney, and pancreas ($\Delta$ adj.$R^2 > 0.1$; $P < 0.001$) and in the pooled pan-cancer set of 37 cancer types (adj.$R^2$ 0.90 *vs.* 0.87; $P < 0.001$). A few exceptions were also apparent; in melanoma, models utilizing CA profiles of normal tissues were significantly more accurate (adj.$R^2$ 0.69 for normal CA *vs.* 0.65 cancer CA; $P = 0.004$). The high somatic mutation burden of normal skin cells due to long-term ultraviolet light exposure (9) appears consistent with the model predictions that a fraction of mutations in melanomas are distributed according to CA profiles of normal tissues. In medulloblastoma, CA profiles of normal tissues were also more predictive of regional mutagenesis (adj.$R^2 = 0.45$ for normal CA *vs*. adj.$R^2 = 0.36$ for cancer CA; $P = 0.001$), potentially explained by the developmental origin of this pediatric brain cancer (32). CA profiles of normal tissues also improved prediction accuracy in B-cell non-Hodgkin's lymphoma (BNHL) and chronic lymphocytic leukemia (CLL) ($P < 0.01$). Overall model accuracy was partially explained by genome-wide mutation burden of cancer types (Spearman rho = 0.61, $P = 0.0011$) but not cohort size (rho = 0.22, $P = 0.28$) (**Supplementary Figure 5**). In summary, regional mutagenesis is more strongly associated with CA of primary human cancers rather than normal tissues and cell lines in most cancer types, indicating that most somatic mutations occur after the cells have acquired the epigenetic characteristics of cancer cells.

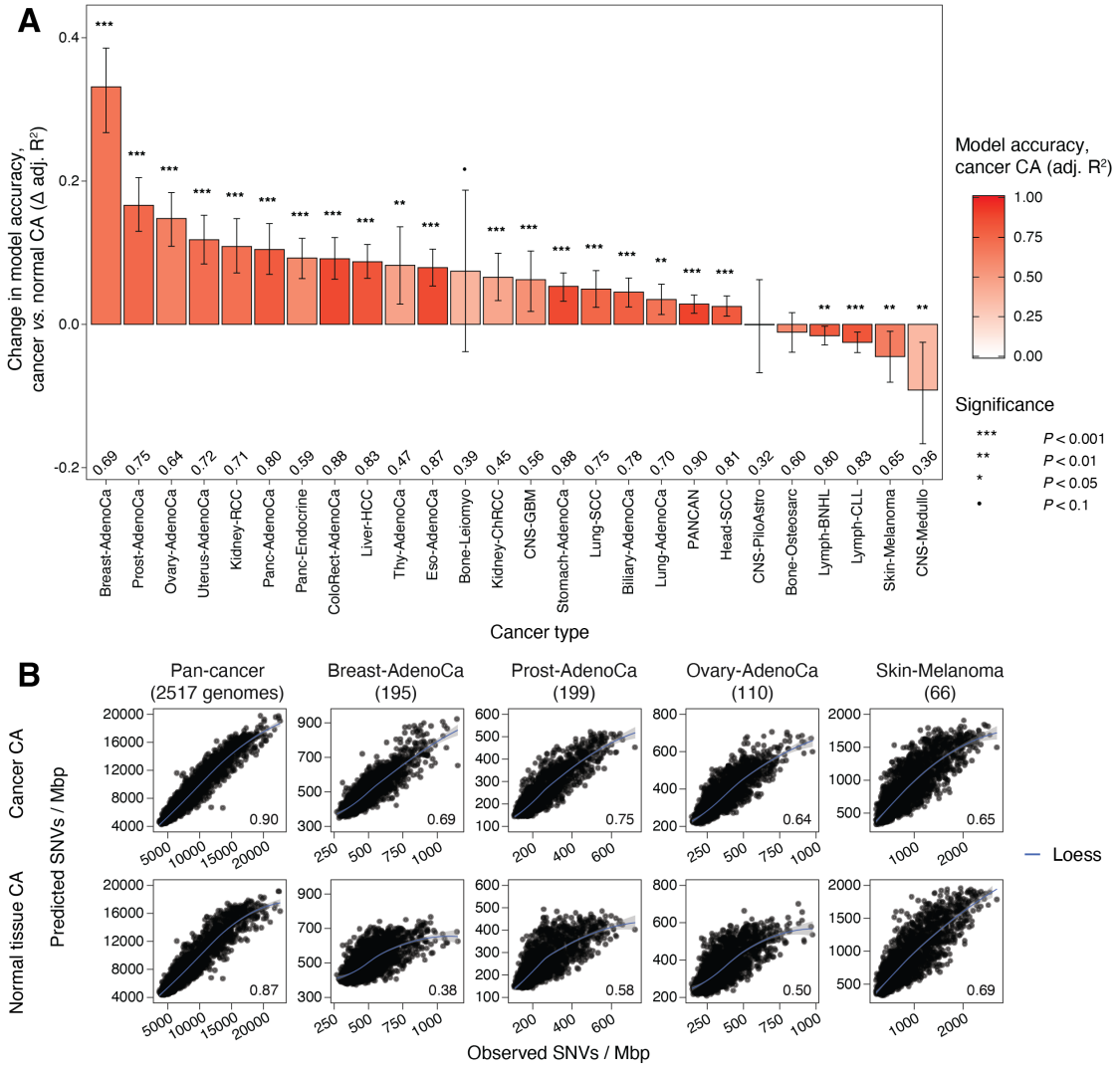


**Figure 2. Chromatin accessibility of primary cancers is a major determinant of regional mutagenesis in cancer genomes. A.** Random forest models informed by CA profiles of primary cancers are more accurate predictors of regional mutagenesis compared to models informed by normal tissues. Bar plot shows relative changes in accuracy (Δ adjusted $R^2$) of cancer CA-informed models in 25 cancer types in PCAWG. Replication timing (RT) profiles are included in both classes of models as reference. Permutation P-values and 95% bootstrap confidence intervals are shown. Accuracy of models informed by cancer CA profiles are shown below bars (adj.$R^2$). **B.** Examples of regional mutagenesis predicted using CA profiles of cancers (top) and normal tissues (bottom). Scatterplots show observed and model-predicted mutation frequencies. Model accuracy values are shown below.

107

108

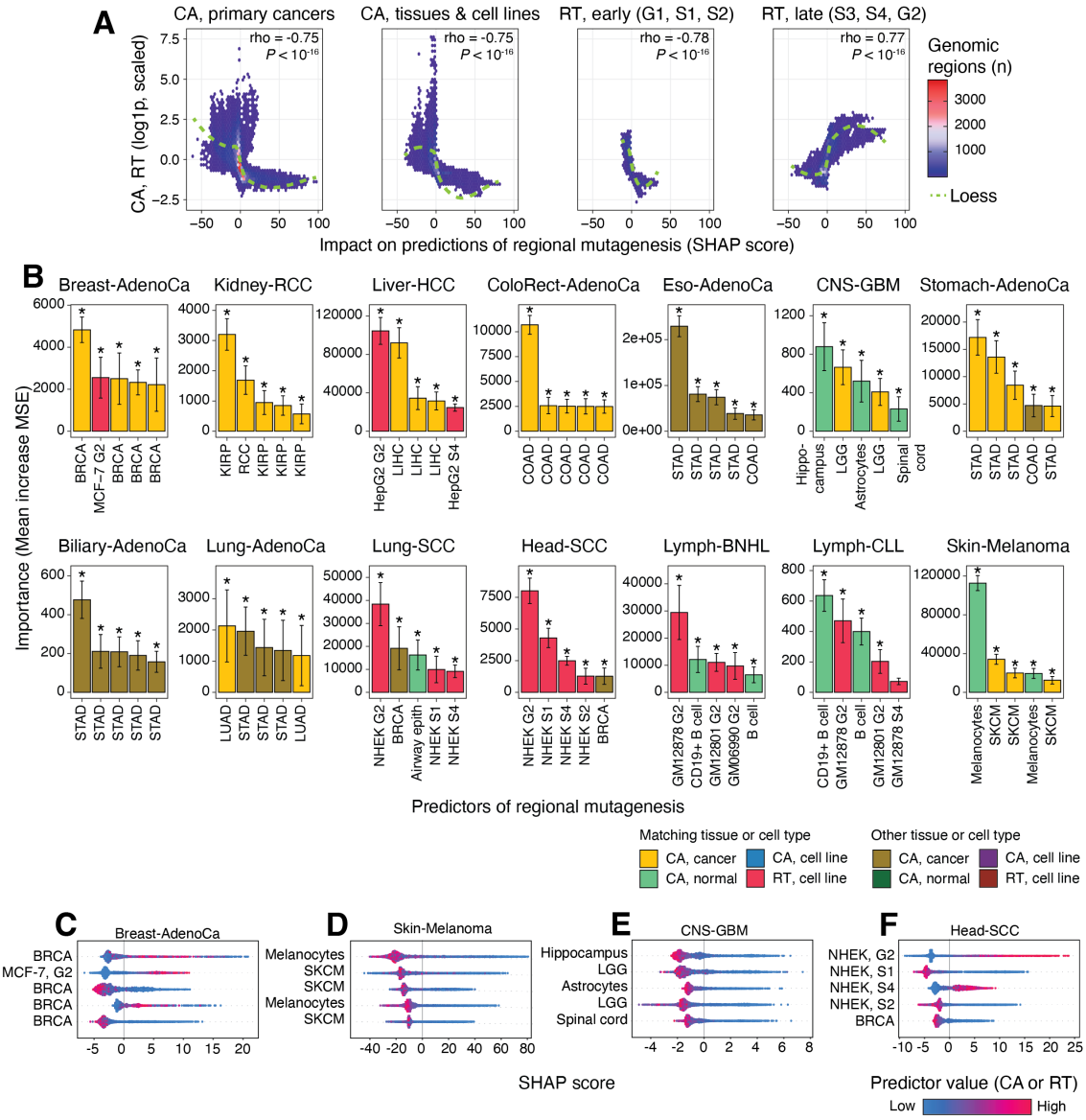


**Figure 3. Top predictors of regional mutagenesis tie chromatin accessibility and replication timing with sites of cancer origin. A.** Negative associations of CA and positive associations of late RT with regional mutagenesis are found in local predictor analysis of random forest models. 2D density plots summarize the effects of individual CA and RT profiles (Y-axis) on regional mutagenesis across 14 cancer types. The five most significant predictors of 14 cancer types are quantified using Shapley additive explanation (SHAP) scores (X-axis). Spearman correlation values are shown (top right). **B.** Major predictors of regional mutagenesis represent cancer tissues of origin. Barplot shows importance scores of top-5 predictors in random forest models of 14 cancer types ($P < 0.001$; ±1 s.d.). Colors indicate the predictor type (CA, RT) and its relationship to the cancer type where mutagenesis is predicted (matching or other). Brighter colors indicate predictors that match cancer type or tissues or cells of origin. **C-F.** Examples of predictors of regional mutagenesis. SHAP scores show the impact of a predictor on the predictions (X-axis) and corresponding predictor values (color gradient). **C.** Top predictors in breast cancer include four CA profiles of primary breast cancers (BRCA) and one RT profile of the breast cancer cell line MCF-7 in phase G2 of cell cycle. **D.** Top predictors in melanoma include the CA profiles of normal melanocytes and melanomas (SKCM). **E.** Top predictors in GBM include CA profiles of lower-grade gliomas (LGG) and normal neuronal tissues. **F.** Top predictors in head squamous cell carcinoma include RT profiles of the squamous cell line NHEK (primary normal human epidermal keratinocytes) and a CA profile of BRCA.

109

110 **Top predictors of regional mutagenesis match cancer types and sites of origin**

111 To interpret regional mutagenesis through cancer tissues of origin, we asked which CA and RT
112 profiles contributed the most to the predictive models. We included all 773 profiles as predictors
113 and analyzed 14 cancer types for which profiles of primary cancers and relevant normal tissues
114 were available. We selected five most significant predictors for each cancer type ($P < 0.001$) and
115 quantified these using Shapley Additive exPlanation (SHAP) scores (31) that reflect associations
116 with mutation burden accumulated across genomic regions. SHAP scores were negatively
117 correlated with CA of cancers and normal tissues (rho = -0.75; $P < 10^{-16}$) (**Figure 3A**). Late-
118 replicating regions were positively correlated with regional mutagenesis (rho = 0.77, $P < 10^{-16}$)
119 while early replicating regions showed a less-variable negative correlation (rho = -0.78, $P < 10^{-}$
120 $^{16}$). This inverse relationship of CA and RT with respect to regional mutagenesis is consistent
121 with previous studies (17-23), however it is extended to a diverse collection of CA and RT
122 profiles of primary cancers, normal tissues, and cell lines. Non-linear associations of regional
123 mutagenesis and its epigenomic predictors are apparent in individual cancer types
124 (**Supplementary Figure 6**). This analysis underlines the complex interactions of regional
125 mutagenesis with CA and RT in this pan-cancer cohort and warrants detailed analysis of
126 individual predictors.

127 We examined the top predictors of regional mutagenesis. CA profiles of matching cancers were
128 among the strongest predictors in eight of 14 cancer types including liver, breast, kidney,
129 stomach, and colorectal cancers ($P < 0.001$), emphasizing tissue-of-origin associations (**Figure**
130 **3B**). For example, regional mutagenesis in breast cancer showed positive and negative
131 associations with four CA profiles of breast cancers of the TCGA dataset (**Figure 3C**).
132 Additional associations appeared at the level of organ systems as CA profiles of stomach and
133 colorectal cancers were the top predictors of regional mutagenesis in colorectal, stomach, biliary
134 and esophageal cancers, suggesting similarities of mutational processes or epigenomes of the
135 gastrointestinal tract. Interestingly, regional mutagenesis in lung adenocarcinomas was also
136 explained by CA profiles of stomach and lung adenocarcinomas. Overall, matched cancer-
137 specific CA profiles showed stronger associations with regional mutagenesis than profiles of
138 normal cells.

139 CA profiles of matching normal tissues associated with regional mutagenesis in five cancer
140 types. Mutations in melanoma were predicted by two CA profiles of normal melanocytes and
141 three profiles of melanomas, whereas a three-fold higher feature importance score was assigned
142 to the normal tissue (incMSE $1.2 \times 10^5$ *vs.* $3.4 \times 10^4$) (**Figure 3B**). Accessible chromatin of
143 cancers and normal melanocytes was relatively depleted in mutations according to SHAP
144 analysis (**Figure 3D**). This is consistent with the shaping of melanoma genomes through the
145 chromatin landscape of normal melanocytes (9) earlier in cancer evolution. CA profiles of
146 normal B-cells were found as predictors in lymphoid cancers BNHL and CLL. Somatic
147 hypermutation (SHM) of immunoglobulin genes in normal B-cells and aberrant SHM in
148 lymphomas (33) potentially explains this association. In glioblastoma (GBM), CA profiles of

149 neuronal tissues (hippocampus, astrocytes, spinal cord) as well as lower-grade gliomas were
150 selected as top features (**Figure 3E**). The mixture of CA of cancers and normal neural cells
151 predictive of regional mutagenesis of GBM may reflect its extensive intratumoral heterogeneity
152 and proposed origin in stem-like cells (34).

153 In most cases, CA negatively correlated with regional mutagenesis according to SHAP scores,
154 both in CA profiles of primary cancers, such as liver and breast cancer, as well as related normal
155 tissues, such as melanocytes in melanoma and astrocytes in glioma (**Figure 3D-E**). However,
156 associations of high CA and increased mutation burden were also apparent. In breast cancer, the
157 two most predictive CA profiles showed positive SHAP scores in highly accessible genomic
158 regions, indicating the activity of a mutational process targeting open chromatin (**Figure 3C**)
159 (**Supplementary Figure 6**). Thus, deconvoluting the bulk profiles of megabase-scale mutation
160 burden helps map interactions of regional mutagenesis with CA and RT.

161 RT profiles were the major predictors of regional mutagenesis in six cancer types. Mutations in
162 lung and head squamous cell carcinomas (SCC) associated with RT profiles of the squamous cell
163 line of normal human epidermal keratinocytes (NHEK) (**Figure 3B,F**). The squamous cell
164 association indicates cell-of-origin patterns of regional mutagenesis, while the association with
165 normal cells may reflect mutagenesis earlier in cancer evolution, potentially through the tobacco
166 signature SBS4 that represents 44% and 12% of SNVs in the Lung-SCC and Head-SCC cohorts,
167 respectively. Similarly, RT profiles of lymphoblastoid cell lines were among the top predictors in
168 CLL and BNHL. Tissue-specific RT profiles of the cancer cell lines MCF-7 and HepG2 were
169 found in breast and liver cancers, respectively. Most RT predictors (13/16) represented late-
170 replicating cell cycle phases G2 and S4. Individual RT profiles positively associated with
171 mutagenesis in late-replicating regions (*e.g.*, phase G2 of MCF-7 in breast cancer) and negatively
172 in early-replicating regions (*e.g.,* phase S1 of HNEK in head SCC) (**Figure 3C,F**), consistent
173 with earlier observations that elevated regional mutagenesis is an effect of increased DNA
174 damage and decreased repair in late replication (20). However, RT profiles were generally
175 underrepresented among top predictors compared to CA profiles. This is likely because fewer
176 and less-diverse RT profiles of cell lines offer only a limited representation of mutational
177 processes in diverse cancer genomes, while CA profiles of primary human cancers provide
178 complementary information. Together, this analysis extends our findings of tissue-specific CA
179 and RT profiles as the principal predictors of regional mutagenesis and underlines cell-of-origin
180 effects and cancer heterogeneity.

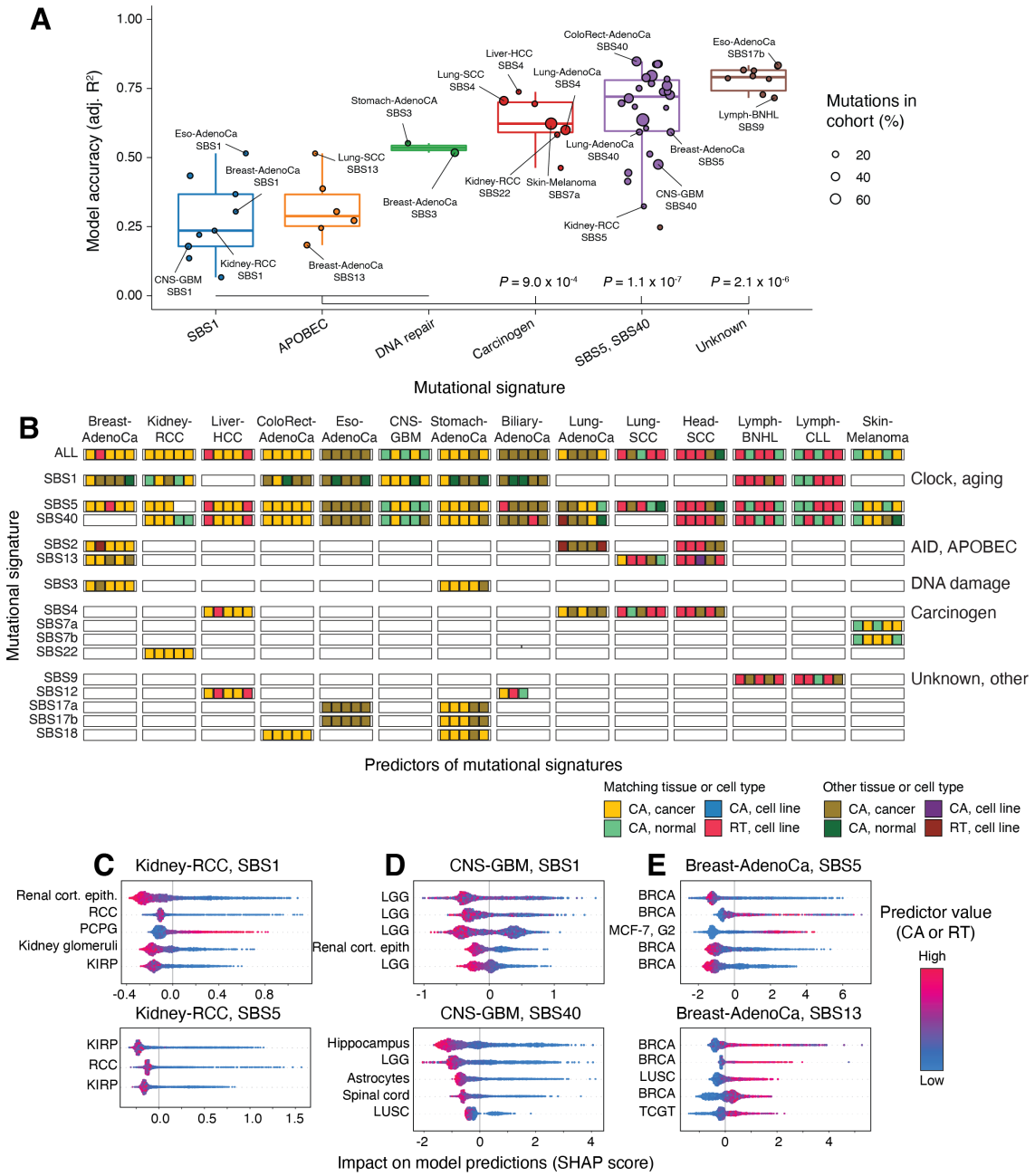


**Figure 4. Associations of mutational signatures with chromatin accessibility and replication timing. A.** Comparison of prediction accuracy of regional distribution of mutational signatures informed by CA and RT profiles. Signatures of carcinogens, unkown origin and SBS5,40 are more accurately predicted by CA and RT profiles than endogenous signatures. P-values adjust for genome-wide signature burden as covariate (F-test). **B.** CA and RT profiles associated with mutational signatures represent tissues of origin. Colored tiles show top-5 predictors of regional distribution of mutational signatures (permutation $P < 0.001$). Colors indicate the predictor type (CA, RT) and its relationship to the cancer type where mutagenesis is predicted (matching or other). Brighter colors indicate predictors that match cancer type or tissues or cells of origin. **C-F.** Examples of predictors of regional mutagenesis. Signatures are grouped vertically by etiology. Bulk mutation profiles are shown in the first row. SBS1 (2nd row) reveals a a diversity of predictors compared to other signatures. **C-E**. Examples of CA and RT profiles as predictors of mutational signatures. SHAP scores show the impact of a predictor on the predictions (X-axis) and corresponding predictor values (color gradient). **C.** In kidney cancer, SBS1 mutations (top) are best predicted by CA profiles of normal kidney cells and kidney cancers (RCC, KIRP) while only CA profiles of kidney cancers are the strongest predictors of SBS5 (bottom). **D.** In glioblastoma (GBM), SBS1 mutations (top) are best predicted by CA profiles of low-grade gliomas (LGG) while the strongest predictors of SBS40 (bottom) include CA profiles of normal neuronal tissues. **E.** In breast cancer, higher CA of BRCA primarily associates with fewer SBS5 mutations (top), while for SBS13 mutations, higher CA associates with higher burden (bottom).

182 **Associations of mutational signatures with chromatin accessibility and replication timing**

183 We asked whether the associations of regional mutagenesis with CA and RT can be further
184 explained by mutational signatures. We assigned each SNV to its most probable single base
185 substitution (SBS) signature (6) and predicted the regional distributions of signatures using 773
186 CA and RT profiles. First, we compared the accuracy values of random forest models in
187 predicting six classes of mutational signatures based on etiology: two age-related classes (SBS1
188 and SBS5/40), APOBEC/AID, DNA-repair, and carcinogen signatures, and signatures of
189 unknown cause. Three classes of signatures were more informative of CA and RT profiles across
190 14 cancer types (**Figure 4A**): predictions of carcinogenic signatures, signatures of unknown
191 cause, and aging-associated signatures (SBS5 and SBS40) were significantly more accurate than
192 predictions of endogenous signatures of DNA repair, APOBEC/AID, and SBS1 (median adj.$R^2 \geq$
193 $0.62$ *vs*. adj.$R^2 = 0.30$; F-test $P \leq 10^{-3}$), when accounting for total signature burden as a covariate
194 of model accuracy. Thus, the mutational processes of carcinogen exposures, aging and unknown
195 signatures show stronger interactions with CA and RT in cancer genomes.

196 We identified the top CA and RT profiles predictive of individual mutational signatures ($P <$
197 $0.001$) (**Figure 4B**). Top predictors of signatures were often in agreement with those of bulk
198 regional mutation burden: matching CA profiles of cancers were the top predictors of mutational
199 signatures in breast, kidney, colorectal and stomach cancers, while RT profiles of normal cells
200 associated with mutations in SCCs and lymphoid cancers. Top predictors of endogenous and
201 exogeneous signatures were also mostly consistent, indicating that various mutational processes
202 are affected by the epigenetic landscapes of cancers or cells of origin.

203 Signature SBS1 deviated from broad CA-driven patterns of regional mutagenesis in several
204 cancer types. In kidney cancer, SBS1 mutations associated with CA profiles of normal tissues
205 such as renal cortex epithelium and kidney glomerulus as well as extra-adrenal
206 pheochromocytoma (PCPG), a rare endocrine cancer. In contrast, SBS5 mutations and others
207 predominantly associated with CA of kidney cancers (**Figure 4C**). Similar effects were observed
208 in breast, colorectal and stomach cancers: SBS1 mutations associated with CA profiles of normal
209 tissues and unrelated cancers, while other signatures associated with CA profiles of matching
210 cancers. However, SBS1 mutations were predicted less accurately than other signatures,
211 potentially due to their overall lower frequency (**Supplementary Figure 7**). Interestingly, an
212 inverse relationship was observed in GBM that may reflect its intratumoral heterogeneity and
213 stem cell origins: the top predictors of SBS1 included CA profiles of four lower-grade gliomas
214 (LGG) while the normal tissue profiles of hippocampus, astrocytes, and spinal cord primarily
215 associated with clock-like signatures SBS5 and SBS40 (**Figure 4D**). The clock-like SBS1
216 signature of 5-methylcytosine deamination is associated with cancer patient age and stem cell
217 division rate and affects the somatic genomes of normal tissues and adult stem cells (10,35,36).
218 The association of SBS1 with the epigenomes of normal tissues suggests that SBS1 mutations in
219 cancer genomes represent a footprint of earlier cancer evolution or somatic mutagenesis in
220 normal cells of cancer origin.

221 We asked whether the mutations of specific signatures were enriched or under-represented in
222 regions of open chromatin. While mutational signatures were generally negatively associated
223 with CA in accordance with bulk mutations, positive associations were also apparent. In breast
224 cancer, SBS13 mutations of APOBEC/AID activity positively associated with high CA scores
225 (**Figure 4E**), in agreement with the observations that AID targeting of epigenetically active
226 elements results in kataegis and clustered mutational signatures (6,37,38). As another example,
227 SBS1 mutations in kidney cancer positively associated with the CA profile of the PCPG cancer
228 while negative associations with CA were apparent in other signatures and CA profiles (**Figure**
229 **4C**). In summary, this analysis highlights the complex interactions of CA and RT with regional
230 mutagenesis and cancer heterogeneity and helps characterize the mechanisms of mutational
231 processes.

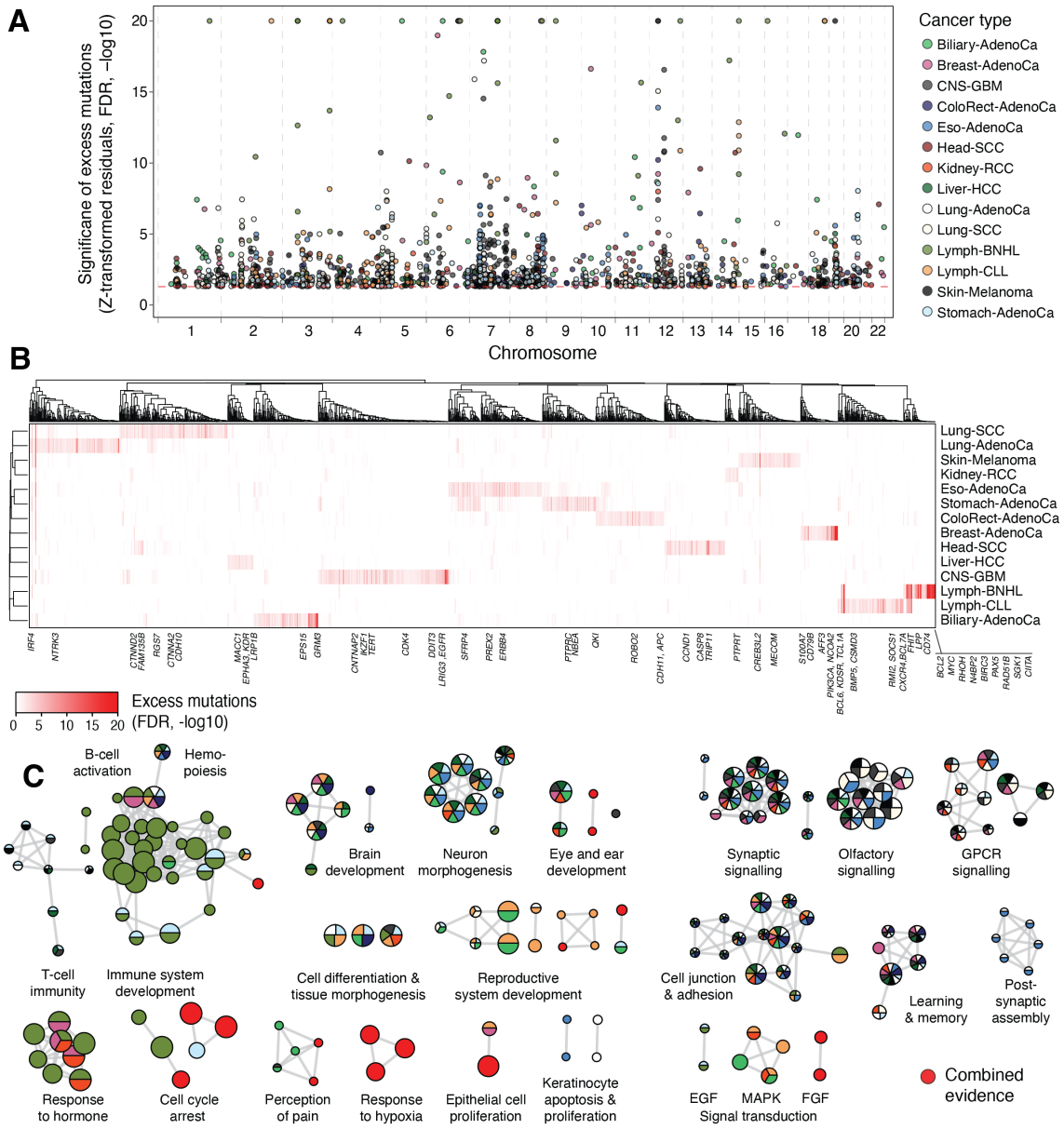


**Figure 5. Excess mutations unexplained by epigenomes converge to cancer genes and developmental pathways. A.** Manhattan plot shows the genomic regions (100 kbps) where the observed mutation frequencies exceed the predictions of CA and RT profiles in 14 cancer types. The significance of increased model residuals is shown on the Y-axis (FDR < 0.05; one-sided tests). **B.** Hierarchical clustering of frequently-mutated genomic regions shows the grouping of related cancer types and enrichment of known cancer genes (listed below). **C.** Enrichment map of biological processes and pathways enriched in the genes of frequently-mutated regions (*FDR* < 0.05). Nodes represent pathways and edges connect pathways that share many genes. Nodes are grouped as subnetworks representing common biological themes. Colors show the cancer types where the pathway enrichments were detected (color legend in (A)). Node size corresponds to the number of genes in the pathway. Red nodes represent pathways that were only detected in the joint analysis of multiple cancer types.

234 **Excess mutations unexplained by epigenomes converge to cancer genes and developmental**
235 **pathways**

236 To quantify the regional mutagenesis unexplained by CA and RT, we investigated the genomic
237 regions that were enriched in mutations above the levels expected from epigenomes. To enable a
238 gene-level functional analysis, we repeated the predictions of regional mutagenesis at a finer
239 genomic resolution (100 kbps) and selected 1,330 regions in 14 cancer types that were
240 significantly enriched in mutations based on the CA- and RT-informed model residuals (*FDR* <
241 0.05) (**Figure 5A**). While the mutation-enriched regions were largely tissue-specific with 86%
242 detected in only one cancer type, hierarchical clustering of the regions by residuals was
243 consistent with cancer types (**Figure 5B**). For example, lymphoid cancers, lung cancers and
244 gastrointestinal cancers made up the three most distinct clusters. Thus, the regional mutational
245 processes independent of CA and RT affect similar genomic regions in related cancer types.

246 We performed a functional analysis of the frequently mutated genomic regions, hypothesizing
247 that these could be characterized by pathways and genes involved in cancer. The regions
248 encoded 730 protein-coding genes including 61 known cancer genes (39), significantly more
249 than expected by chance (27 expected, Fisher's exact $P = 1.9 \times 10^{-9}$) (**Figure 5B**). Most driver
250 genes were only found in single cancer types and represented key disease-specific drivers such as
251 *EGFR* and *TERT* in glioma, *MYC* in BNHL, *PIK3CA* in breast cancer and *APC* in colorectal
252 cancer (**Supplementary Figure 8**). As an exception, one genomic window recurrently mutated
253 in 11 cancer types includes the interferon regulatory factor and oncogene *IRF4 (40)*, the
254 phosphatase *DUSP22* recently suggested as a network-implicated driver gene due to non-coding
255 mutations (41), and super-enhancers of immune cells (42), indicating a potential pan-cancer
256 region of interest (**Supplementary Figure 9**).

257 We then asked whether the frequently-mutated regions were associated with common biological
258 functions by prioritizing pan-cancer signals of mutation enrichment using the integrative
259 ActivePathways method (43). The analysis revealed 220 significantly enriched pathways (*FDR* <
260 0.05), of which 162 (74%) were detected in more than one cancer type (**Figure 5C**).
261 Developmental processes including the nervous system, heart and kidney, stem cell development
262 and morphogenesis were prominently represented together with cancer hallmark processes such
263 as cell cycle, apoptosis, cell adhesion, hypoxia response, and MAPK, EGF and FGF signalling
264 pathways. Processes of the immune system, stress response, reproduction and hormone
265 regulation were also apparent. Enriched mutations converged to similar pathways and processes
266 across multiple cancer types although most genomic regions were only detected in one or few
267 cancer types. Convergence of these excess mutations to developmental and cancer pathways is
268 potentially explained by further mutational processes targeting active regions of the genome,
269 while the enrichment of known cancer driver genes suggests that positive selection of functional
270 mutations may also contribute to this additional mutation burden. This analysis exemplifies the
271 complex interplay of multi-scale mutational processes and genome function.

272

## DISCUSSION

Our analysis highlights chromatin accessibility of primary human cancers as a major covariate of regional mutational processes that is supported by tissue of origin associations of whole cancer genomes and epigenomes of matching cancer types. These observations are apparent in several common cancer types of the largest global burden. Cancers such as melanoma and lymphoma where normal tissue epigenomes are highly predictive of regional mutagenesis have an etiology consistent with early somatic mutagenesis in normal tissues of origin. Combined associations with normal and cancer epigenomes as observed in GBM may also reflect intratumoral heterogeneity of cell populations and regional mutagenesis. These findings extend earlier studies that used the epigenetic profiles of cell lines and normal tissues to characterize mutational processes. Overall, this analysis suggests that in most cancer types, the megabase-scale landscape of passenger mutations is primarily shaped later in cancer evolution following the epigenetic transformation to cancer cells.

Replication timing information also associated with regional mutagenesis and confirmed strong effects with cell types related to cancer origin. However, CA profiles of primary human cancers evidently captured a larger fraction of variation of regional mutagenesis compared to RT profiles, apart from squamous cell cancers that strongly associated with relevant cell lines. RT profiles make up a smaller subset of epigenomic predictors in our dataset and include mitotic cell lines that offer only limited representation of the diverse disease types in the pan-cancer cohort. Interestingly, DNA replication has been shown to determine chromatin state (44). Thus, the informative CA profiles of human cancers may represent a proxy of cancer-specific replication dynamics.

Mutational signature analysis revealed interactions of mutational processes with CA and tissues of origin. Carcinogen signatures, as well as signatures of unknown etiology, were overall better predicted by CA and RT, in contrast to signatures of aging and DNA damage where the genome-wide predictions were less accurate. The stronger association of carcinogen signatures suggests that the chromatin environment interacts with DNA damage or repair processes of carcinogen exposure, for example through elevated mutational processes targeting active genes that are otherwise protected from mutations through error-free mismatch repair (38). Early replicating regions in cells exposed to tobacco mutagens show elevated mutagenesis in transcribed strands due to differential nucleotide excision repair activity (45). Based on their stronger interactions with RT and CA profiles, we extrapolate that some mutational signatures of currently unknown etiology may relate to carcinogens. SBS17a/b mutations show some of the strongest interactions with CA and RT in stomach and esophageal cancers in our analysis. This signature is currently of unknown cause, however it has been linked to gastric acid reflux and reactive oxygen species (24). Further integrative analysis of clinical and lifestyle information with patterns of regional mutagenesis may shed light to these mutational processes.

Mutations of signature SBS1 associated with CA profiles of relevant normal tissues in multiple cancer types, in contrast to other signatures that were often associated with cancer epigenomes.

312 SBS1 mutations follow a clock-like pattern whose frequency in cancer genomes correlates with
313 patient age and stem cell replication rates (35). SBS1 mutations contribute to the somatic
314 variation landscapes of normal adult tissues (46) and in embryonic development (47). Based on
315 our data, we speculate that SBS1 mutations in cancer genomes represent a footprint of early
316 cancer evolution or somatic variation of normal cells that occurs prior to the acquisition of
317 cancer-specific epigenetic profiles.

318 We observed a functional convergence to developmental processes and cancer-related pathways
319 in the genomic regions where CA and RT profiles insufficiently captured elevated mutation
320 burden. These data suggest that additional mutational processes affect lineage-specific
321 developmental genes and open-chromatin regions that are distinct in individual cancer types,
322 however map to the same molecular pathways across cancer types. For example, transcription
323 start sites of highly expressed genes and constitutively-bound binding sites of CTCF are subject
324 to elevated local mutagenesis in multiple cancer types (16). Lineage-specific genes are enriched
325 in indel mutations in solid cancers (48). Such local mutational processes confound the
326 observations at the megabase resolution of the genome where open chromatin is generally
327 associated with a lower mutation frequency. On the other hand, the enrichment of cancer genes
328 and pathways in our data suggests that some mutations unexplained by CA and RT are functional
329 in cancer and their frequent occurrence at specific genes, non-coding elements and molecular
330 pathways is explained by positive selection (3-5,41). Further study of these regions may deepen
331 our understanding of mutational processes and refine the catalogues of driver mutations.

332 This approach enables future studies to decipher the mechanisms and phenotypic associations of
333 mutational processes. Clinical, genetic, and epigenetic profiles of cancer patients can be
334 integrated to understand how regional mutational processes and the chromatin landscape are
335 modulated by clinical variables such as stage, grade or the therapies applied, genetic features
336 such as somatic driver mutations or inherited cancer risk variants, or lifestyle choices such as
337 tobacco or alcohol consumption. Complementary insights from sub-clonal reconstruction
338 analysis of cancer genomes (2,49), as well as single-cell sequencing of genomes and epigenomes
339 will allow mapping of regional mutagenesis at the level of distinct cell populations contributing
340 to temporal and spatial variation in mutational processes. As such multimodal datasets grow, we
341 can learn about early cancer evolution by comparing regional mutagenesis in the genomes of
342 cancers and normal cells. Understanding the molecular and genetic determinants of regional
343 mutagenesis and signatures in cancer genomes may help characterize carcinogen exposures and
344 genetic predisposition, ultimately enhancing early cancer detection and prevention in the future.

## Methods

**Somatic mutations in whole cancer genomes.** Somatic single nucleotide variants (SNVs) of 2,583 whole cancer genomes of were derived from the Pan-cancer Analysis of Whole Genomes (PCAWG) project (1) and hypermutated tumors (66) were removed, resulting in a dataset of 23.2 million SNVs in 2,517 cancer genomes. Indel mutations and all variants in sex chromosomes were excluded. We analyzed 25 cancer types with at least 30 genomes per cohort and the pan-cancer cohort of 37 cancer types (**Supplementary Figure 1)**. Mutations were mapped to GRCh38 coordinates using LiftOver (50).

**Chromatin accessibility (CA) and replication timing (RT).** 381 CA profiles of primary human cancers were retrieved from the TCGA study (27). 296 CA profiles of normal human tissues and cell lines were derived from ENCODE3 (28) and Epigenomics Roadmap (29) (**Supplementary Figure 2**). RT (RepliSeq) profiles of 16 human cell lines and 6 cell cycle phases were derived from the ENCODE study (30). CA and RT profiles were mapped to GRCh38 where needed.

**Regional variation in mutation burden, CA, and RT.** The genome was segmented into 2,465 distinct regions of one megabase (Mbps) after excluding sex chromosomes and filtering lowly mappable regions using the UMAP software (51). For each window, bulk SNV counts and SNV counts grouped by SBS signatures were derived for every cancer cohort. To create CA and RT profiles, mean values of each track were derived for every genomic region.

**Random forest regression.** Regional mutation burden was modeled as a function of CA and RT profiles with random forest regression (52). Monte-Carlo cross-validation was used evaluate model performance over 1,000 80/20% data splits for training and validation. We used the adjusted $R^2$ (adj.$R^2$) metric of accuracy that measures the complexity-adjusted fraction of variance explained by the model.

**CA of normal cells and cancers in regional mutagenesis.** Two sets of random forest regression models with CA profiles of cancers and normal cells, respectively, were run in 1,000 Monte-Carlo 80/20% cross-validations using matched genomic regions for training and validation. Differences in adj.$R^2$ values of models informed by cancer CA profiles relative to models informed by normal tissue CA profiles were computed ($\Delta$adj.$R^2$) with 95% confidence intervals. Scatterplots of mutation counts were derived from models trained with full data.

**Feature analysis of individual CA and RT profiles.** The increase in mean-squared-error (incMSE) metric was used to evaluate CA and RT profiles as predictors of regional mutagenesis. Significance and variation of incMSE was evaluated. Permutation tests were used to detect the profiles where the incMSE values significantly exceeded those derived from randomly shuffled model responses. Empirical p-values were computed for every profile and cancer type across 1,000 iterations and significant profiles were selected ($P < 0.001$). Bootstrap analysis of genomic regions over 1,000 iterations revealed the variation in incMSE values.

**Local effects of CA and RT profiles to regional mutagenesis.** The Shapley Additive exPlanation (SHAP) method (31) was used to evaluate the effects of profiles to mutagenesis in

specific cancers and genomic regions. SHAP scores represent the importance and direction of each feature (*i.e.*, a CA or RT profile) in predicting an observation (*i.e.*, a genomic window). SHAP scores were computed separately for cancer types on models trained on full datasets.

**Mutational signature analysis.** SNVs were annotated to single base substitution (SBS) signatures from PCAWG (6) using top probabilities. In each cancer type, signatures with at least 10,000 mutations and at least 5% of total mutation burden were selected (**Supplementary Figure 7**). Random forest regression was conducted with evaluation of accuracy and feature analysis as described above. Signatures were grouped based on etiology according to the COSMIC database (v 3.2, downloaded March 2021). Adj.$R^2$ values for different groups of signatures were compared using ANOVA analysis and F-tests where the average signature exposures in cancer types were used as covariates.

**Selecting genomic regions with excess mutations beyond CA and RT predictions.** Regional mutagenesis was predicted for 100-kbps genomic regions to enable more granular functional interpretation. Genomic regions were selected based on model residuals, *i.e.*, where the observed mutation counts significantly exceeded model predictions. Residuals were Z-transformed and the one-tailed P-values were adjusted for multiple testing ($FDR < 0.05$). Regions with excess mutations were visualized as a heatmap with hierarchical clustering and correlation distance. Known cancer genes were derived from the Cancer Gene Census database (39) (downloaded Nov 26$^{th}$ 2020). Enrichment of cancer genes was evaluated with a Fisher's exact test.

**Pathway enrichment analysis of highly-mutated genomic regions.** Integrative pathway enrichment analysis with the ActivePathways method (43) was used to find overrepresented pathways and prioritize genes. Genes were assigned the P-values of their genomic regions and were scored in ActivePathways such that genes in regions with mutation enrichments in multiple cancer types ranked higher. Significantly over-represented pathways ($FDR < 0.05$) were visualized as an enrichment map using standard protocols (53).

A detailed description of methods is available in **Supplementary Materials**.

**Acknowledgments.** We thank Christian A. Lee, Kevin Cheng, Dr. Phedias Diamandis and Dr. Anne Martel for constructive comments on this study. This work was supported by the Canadian Institutes of Health Research (CIHR) Project Grant to J.R., A New Investigator Award of the Terry Fox Research Institute (TFRI) to J.R., and the Investigator Award to J.R. from the Ontario Institute for Cancer Research (OICR). Funding to OICR is provided by the Government of Ontario. The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. We acknowledge the contributions of the many clinical networks of ICGC and TCGA who provided samples and data to PCAWG. We thank the patients and their families for their participation in ICGC and TCGA projects.

**Author contributions.** O.O. analyzed the data and prepared the figures. J.R. and O.O. interpreted the data and wrote the manuscript. J.R. conceived and supervised the project. The authors reviewed and edited the manuscript and approved the final version.

421 **References**


422 1. ICGC-TCGA Pan-Cancer Analysis of Whole Genomes Consortium. (2020) Pan-cancer
423 analysis of whole genomes. *Nature*, **578**, 82-93.
424 2. Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S.C., Gonzalez, S., Rosebrock, D.,
425 Mitchell, T.J., Rubanova, Y., Anur, P., Yu, K. *et al.* (2020) The evolutionary history of
426 2,658 cancers. *Nature*, **578**, 122-128.
427 3. Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P.,
428 Davies, H., Stratton, M.R. and Campbell, P.J. (2017) Universal Patterns of Selection in
429 Cancer and Somatic Tissues. *Cell*, **171**, 1029-1041 e1021.
430 4. Rheinbay, E., Nielsen, M.M., Abascal, F., Wala, J.A., Shapira, O., Tiao, G., Hornshøj,
431 H., Hess, J.M., Juul, R.I., Lin, Z. *et al.* (2020) Analyses of non-coding somatic drivers in
432 2,693 cancer whole genomes. *Nature*, **578**, 102–111.
433 5. Zhu, H., Uuskula-Reimand, L., Isaev, K., Wadi, L., Alizada, A., Shuai, S., Huang, V.,
434 Aduluso-Nwaobasi, D., Paczkowska, M., Abd-Rabbo, D. *et al.* (2020) Candidate Cancer
435 Driver Mutations in Distal Regulatory Elements and Long-Range Chromatin Interaction
436 Networks. *Mol Cell*.
437 6. Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot,
438 A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N. *et al.* (2020) The repertoire of
439 mutational signatures in human cancer. *Nature*, **578**, 94-101.
440 7. Li, Y., Roberts, N.D., Wala, J.A., Shapira, O., Schumacher, S.E., Kumar, K., Khurana,
441 E., Waszak, S., Korbel, J.O., Haber, J.E. *et al.* (2020) Patterns of somatic structural
442 variation in human cancer genomes. *Nature*, **578**, 112-121.
443 8. Kumar, S., Warrell, J., Li, S., McGillivray, P.D., Meyerson, W., Salichos, L., Harmanci,
444 A., Martinez-Fundichely, A., Chan, C.W.Y., Nielsen, M.M. *et al.* (2020) Passenger
445 Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact
446 and Consequences. *Cell*, **180**, 915-927 e916.
447 9. Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge,
448 D.C., Fullam, A., Alexandrov, L.B., Tubio, J.M. *et al.* (2015) Tumor evolution. High
449 burden and pervasive positive selection of somatic mutations in normal human skin.
450 *Science*, **348**, 880-886.
451 10. Blokzijl, F., de Ligt, J., Jager, M., Sasselli, V., Roerink, S., Sasaki, N., Huch, M.,
452 Boymans, S., Kuijk, E., Prins, P. *et al.* (2016) Tissue-specific mutation accumulation in
453 human adult stem cells during life. *Nature*, **538**, 260-264.
454 11. Supek, F. and Lehner, B. (2019) Scales and mechanisms of somatic mutation rate
455 variation across the human genome. *DNA Repair (Amst)*, 102647.
456 12. Gonzalez-Perez, A., Sabarinathan, R. and Lopez-Bigas, N. (2019) Local Determinants of
457 the Mutational Landscape of the Human Genome. *Cell*, **177**, 101-114.
458 13. Pich, O., Muinos, F., Lolkema, M.P., Steeghs, N., Gonzalez-Perez, A. and Lopez-Bigas,
459 N. (2019) The mutational footprints of cancer therapies. *Nature genetics*, **51**, 1732-1740.
460 14. Katainen, R., Dave, K., Pitkanen, E., Palin, K., Kivioja, T., Valimaki, N., Gylfe, A.E.,
461 Ristolainen, H., Hanninen, U.A., Cajuso, T. *et al.* (2015) CTCF/cohesin-binding sites are
462 frequently mutated in cancer. *Nature genetics*, **47**, 818-821.
463 15. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. and Lopez-Bigas, N.
464 (2016) Nucleotide excision repair is impaired by binding of transcription factors to DNA.
465 *Nature*, **532**, 264-267.

466 16. Lee, C.A., Abd-Rabbo, D. and Reimand, J. (2021) Functional and genetic determinants of
467 mutation rate variability in regulatory elements of cancer genomes. *Genome Biol*, **22**,
468 133.
469 17. Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A.,
470 Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A. *et al.* (2013) Mutational
471 heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**,
472 214-218.
473 18. Schuster-Bockler, B. and Lehner, B. (2012) Chromatin organization is a major influence
474 on regional mutation rates in human cancer cells. *Nature*, **488**, 504-507.
475 19. Stamatoyannopoulos, J.A., Adzhubei, I., Thurman, R.E., Kryukov, G.V., Mirkin, S.M.
476 and Sunyaev, S.R. (2009) Human mutation rate associated with DNA replication timing.
477 *Nature genetics*, **41**, 393-395.
478 20. Supek, F. and Lehner, B. (2015) Differential DNA mismatch repair underlies mutation
479 rate variation across the human genome. *Nature*, **521**, 81-84.
480 21. Zheng, C.L., Wang, N.J., Chung, J., Moslehi, H., Sanborn, J.Z., Hur, J.S., Collisson,
481 E.A., Vemula, S.S., Naujokas, A., Chiotti, K.E. *et al.* (2014) Transcription restores DNA
482 repair to heterochromatin, determining regional mutation rates in cancer genomes. *Cell*
483 *Rep*, **9**, 1228-1234.
484 22. Woo, Y.H. and Li, W.H. (2012) DNA replication timing and selection shape the
485 landscape of nucleotide variation in cancer genomes. *Nat Commun*, **3**, 1004.
486 23. Liu, L., De, S. and Michor, F. (2013) DNA replication timing and higher-order nuclear
487 organization determine single-nucleotide substitution patterns in cancer genomes. *Nat*
488 *Commun*, **4**, 1502.
489 24. Tomkova, M., Tomek, J., Kriaucionis, S. and Schuster-Bockler, B. (2018) Mutational
490 signature distribution varies with DNA replication timing and strand asymmetry. *Genome*
491 *Biol*, **19**, 129.
492 25. Jiao, W., Atwal, G., Polak, P., Karlic, R., Cuppen, E., Subtypes, P.T., Clinical Translation
493 Working, G., Danyi, A., de Ridder, J., van Herpen, C. *et al.* (2020) A deep learning
494 system accurately classifies primary and metastatic cancers using passenger mutation
495 patterns. *Nat Commun*, **11**, 728.
496 26. Polak, P., Karlic, R., Koren, A., Thurman, R., Sandstrom, R., Lawrence, M.S., Reynolds,
497 A., Rynes, E., Vlahovicek, K., Stamatoyannopoulos, J.A. *et al.* (2015) Cell-of-origin
498 chromatin organization shapes the mutational landscape of cancer. *Nature*, **518**, 360-364.
499 27. Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C.,
500 Groeneveld, C., Wong, C.K., Cho, S.W. *et al.* (2018) The chromatin accessibility
501 landscape of primary human cancers. *Science*, **362**.
502 28. Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D.,
503 Neri, F., Haugen, E. *et al.* (2020) Global reference mapping of human transcription factor
504 footprints. *Nature*, **583**, 729-736.
505 29. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M.,
506 Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J. *et al.* (2015)
507 Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317-330.
508 30. Hansen, R.S., Thomas, S., Sandstrom, R., Canfield, T.K., Thurman, R.E., Weaver, M.,
509 Dorschner, M.O., Gartler, S.M. and Stamatoyannopoulos, J.A. (2010) Sequencing newly
510 replicated DNA reveals widespread plasticity in human replication timing. *Proc Natl*
511 *Acad Sci U S A*, **107**, 139-144.

31. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.I. (2020) From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell*, **2**, 56-67.
32. Marino, S. (2005) Medulloblastoma: developmental mechanisms out of control. *Trends Mol Med*, **11**, 17-22.
33. Odegard, V.H. and Schatz, D.G. (2006) Targeting of somatic hypermutation. *Nat Rev Immunol*, **6**, 573-583.
34. Singh, S.K., Hawkins, C., Clarke, I.D., Squire, J.A., Bayani, J., Hide, T., Henkelman, R.M., Cusimano, M.D. and Dirks, P.B. (2004) Identification of human brain tumour initiating cells. *Nature*, **432**, 396-401.
35. Alexandrov, L.B., Jones, P.H., Wedge, D.C., Sale, J.E., Campbell, P.J., Nik-Zainal, S. and Stratton, M.R. (2015) Clock-like mutational processes in human somatic cells. *Nature genetics*, **47**, 1402-1407.
36. Abascal, F., Harvey, L.M.R., Mitchell, E., Lawson, A.R.J., Lensing, S.V., Ellis, P., Russell, A.J.C., Alcantara, R.E., Baez-Ortega, A., Wang, Y. *et al.* (2021) Somatic mutation landscapes at single-molecule resolution. *Nature*.
37. Wang, Q., Oliveira, T., Jankovic, M., Silva, I.T., Hakim, O., Yao, K., Gazumyan, A., Mayer, C.T., Pavri, R., Casellas, R. *et al.* (2014) Epigenetic targeting of activation-induced cytidine deaminase. *Proc Natl Acad Sci U S A*, **111**, 18667-18672.
38. Supek, F. and Lehner, B. (2017) Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. *Cell*, **170**, 534-547 e523.
39. Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. and Stratton, M.R. (2004) A census of human cancer genes. *Nat Rev Cancer*, **4**, 177-183.
40. Iida, S., Rao, P.H., Butler, M., Corradini, P., Boccadoro, M., Klein, B., Chaganti, R.S. and Dalla-Favera, R. (1997) Deregulation of MUM1/IRF4 by chromosomal translocation in multiple myeloma. *Nature genetics*, **17**, 226-230.
41. Reyna, M.A., Haan, D., Paczkowska, M., Verbeke, L.P.C., Vazquez, M., Kahraman, A., Pulido-Tamayo, S., Barenboim, J., Wadi, L., Dhingra, P. *et al.* (2020) Pathway and network analysis of more than 2,500 whole cancer genomes. *Nature Communications*, **11**, 729.
42. Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-Andre, V., Sigova, A.A., Hoke, H.A. and Young, R.A. (2013) Super-enhancers in the control of cell identity and disease. *Cell*, **155**, 934-947.
43. Paczkowska, M., Barenboim, J., Sintupisut, N., Fox, N.S., Zhu, H., Abd-Rabbo, D., Mee, M.W., Boutros, P.C., PCAWG Drivers and Functional Interpretation Working Group, Reimand, J. *et al.* (2020) Integrative pathway enrichment analysis of multivariate omics data. *Nature Communications*, **11**, 735.
44. Klein, K.N., Zhao, P.A., Lyu, X., Sasaki, T., Bartlett, D.A., Singh, A.M., Tasan, I., Zhang, M., Watts, L.P., Hiraga, S.I. *et al.* (2021) Replication timing maintains the global epigenetic state in human cells. *Science*, **372**, 371-378.
45. Kucab, J.E., Zou, X., Morganella, S., Joel, M., Nanda, A.S., Nagy, E., Gomez, C., Degasperi, A., Harris, R., Jackson, S.P. *et al.* (2019) A Compendium of Mutational Signatures of Environmental Agents. *Cell*, **177**, 821-836 e816.
46. Moore, L., Cagan, A., Coorens, T., Neville, M.D., Sanghvi, R., Sanders, M.A., Oliver, T.R., Leongamornlert, D., Ellis, P. and Noorani, A. (2020) The mutational landscape of human somatic and germline cells. *bioRxiv*.

558 47. Bizzotto, S., Dou, Y., Ganz, J., Doan, R.N., Kwon, M., Bohrson, C.L., Kim, S.N., Bae,
559 T., Abyzov, A., Network, N.B.S.M. *et al.* (2021) Landmarks of human embryonic
560 development inscribed in somatic mutations. *Science*, **371**, 1249-1253.
561 48. Imielinski, M., Guo, G. and Meyerson, M. (2017) Insertions and Deletions Target
562 Lineage-Defining Genes in Human Cancers. *Cell*, **168**, 460-472 e414.
563 49. Dentro, S.C., Leshchiner, I., Haase, K., Tarabichi, M., Wintersinger, J., Deshwar, A.G.,
564 Yu, K., Rubanova, Y., Macintyre, G., Demeulemeester, J. *et al.* (2021) Characterizing
565 genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell*.
566 50. Lawrence, M., Gentleman, R. and Carey, V. (2009) rtracklayer: an R package for
567 interfacing with genome browsers. *Bioinformatics*, **25**, 1841-1842.
568 51. Karimzadeh, M., Ernst, C., Kundaje, A. and Hoffman, M.M. (2018) Umap and Bismap:
569 quantifying genome and methylome mappability. *Nucleic Acids Res*, **46**, e120.
570 52. Ho, T.K. (1995) Random decision forests. *IEEE Proceedings of 3rd international*
571 *conference on document analysis and recognition*, **1**, 278-282.
572 53. Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A.,
573 Wadi, L., Meyer, M., Wong, J., Xu, C. *et al.* (2019) Pathway enrichment analysis and
574 visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat*
575 *Protoc*, **14**, 482-517.

576