1 **Modeling tissue-specific breakpoint proximity of structural variations from 2,382 whole-genomes** 2 **to identify cancer drivers**

3 Alexander Martinez-Fundichely^{1,2,3,4,*}, Austin Dixon⁵, Ekta Khurana^{1,2,3,4,*}

4 **Affiliations:**

5 ¹ Sandra and Edward Meyer Cancer Center, Weill Cornell Medicine, New York, NY, 10021, USA.

6 ² Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, 10021, USA.

7 ³ Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, 10065, USA.

8 ⁴ Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY, 10021, USA.

9 ⁵ Children's National Hospital, Washington, DC, 20010, USA

10 * Corresponding authors. Email: ekk2003@med.cornell.edu (E.K.); alm2069@med.cornell.edu (A.MF.)

11 **Abstract**

12 Structural variations (SVs) in cancer cells often impact large genomic regions with functional
 13 consequences. However, little is known about the genomic features related to the breakpoint distribution
 14 of SVs in different cancers, a prerequisite to distinguish loci under positive selection from those with
 15 neutral evolution. We developed a method that uses a generalized additive model to investigate the
 16 breakpoint proximity curves from 2,382 whole-genomes of 32 cancer types. We find that a multivariate
 17 model, which includes linear and nonlinear partial contributions of various tissue-specific features and
 18 their interaction terms, can explain up to 57% of the observed deviance of breakpoint proximity. In
 19 particular, three-dimensional genomic features such as topologically associating domains (TADs), TAD-
 20 boundaries and their interaction with other features show significant contributions. The model is
 21 validated by identification of known cancer genes and revealed putative drivers in novel cancers that
 22 have previous evidence of therapeutic relevance in other cancers.

23 Introduction

24 Whole-genome sequencing of cancer genomes has revealed that they contain a wide variety of DNA
 25 structural variations (SVs) that include deletions, duplications, translocations and other complex events
 26 [1]. The SVs in cancer cells arise from different mechanisms and vary in size from kilobases to whole
 27 chromosomal rearrangements [1-4]. Consequently, SVs usually span several genes and their associated
 28 regulatory elements. While it is well known that genomic rearrangements and copy number variations
 29 (CNVs) can lead to dysregulation of tumor-suppressors or oncogenes and act as drivers of cancer
 30 progression [1, 5-8], identification of SVs under positive selection in cancer remains a challenging task.
 31 This is because SVs are heterogeneously distributed across the genome leading to many genomic
 32 regions recurrently altered in multiple samples due to neutral background processes [4]. To identify the
 33 events under positive selection, the null background distribution of SV breakpoints must be
 34 characterized accounting for the genomic covariates [5, 9]. Additionally, the identification of the
 35 specific functional element that is the target of positive selection (i.e., the coding sequence of a gene, its
 36 cis-regulatory regions, or non-coding RNAs) constitutes another challenge due to the large genomic
 37 span of SVs.

38 While numerous computational methods have been developed to model the background distribution of
 39 single-nucleotide variants (SNVs) and identify drivers in a tissue-specific manner, similar methods for
 40 SVs are lacking [5, 10-13]. The Pan-Cancer Analysis of Whole Genomes (PCAWG) SV Working Group
 41 used a Gamma-Poisson fit to model the breakpoint density from 2,658 whole-genomes using eight
 42 covariates to identify the significant driver genes in several cancers [5]. However, this analysis was
 43 performed at the pan-cancer level without accounting for tissue-specific covariates. Since there is ample
 44 evidence of different SV distributions and putative mechanisms across cancer types [1, 2, 14], it is

critical to model SV breakpoint distribution in a tissue-specific manner to obtain the corresponding accurate null models. Importantly, the three-dimensional (3D) high-order genomic structure, such as the topologically associating domains (TADs), has not previously been considered a covariate of breakpoint distribution. Recent studies have analyzed the relationship between 3D topology and SVs in cancer [8, 15-17]. It was reported that while SVs can lead to changes in chromatin folding, only 14% of TAD-boundary deletions are associated with significant gene expression changes [16]. This finding indicates that many of these SV events may be related to neutral evolution in tumor cells and must be accounted for in the null model to identify true drivers. Furthermore, previous studies did not account for the non-linear contribution of covariates or their interactions in the modeling of SV breakpoint distribution.

Here, we describe a new method that models the breakpoint proximity of SVs in 2,382 cancer whole-genomes in a tissue-specific manner. We used nine different genomic covariates, including the recurrence of TADs and TAD-boundaries across cell lines and tissues as well as their functional classification based upon chromatin states. We implemented a generalized additive model (GAM) to describe the genome-wide SV breakpoint proximity. The use of a GAM allows us to include both the linear and non-linear contributions of variables as well as their interplay. Modeling breakpoint proximity has the advantage of allowing us to analyze the clustering of breakpoints over dynamic genomic lengths to capture different breakpoint trends, unlike breakpoint-density-based approaches that rely on fixed genomic bin sizes. Finally, we use this approach to identify loci that exhibit signals of positive selection and the functional elements that are the likely targets of selection. The method is implemented in Cancer Structural Variation Drivers (CSVDriver), a user-friendly tool that can be used by researchers to model the SVs from whole-genome sequencing data and identify putative cancer drivers.

66

67 **Results**

68 We analyzed a set of 324,838 high confidence somatic SVs derived from whole-genome sequencing of
69 2,382 patients of 32 distinct cancer types from 15 different organ systems (Supplementary Table 1). The
70 cancer types include those with a high SV burden from the PCAWG project [1, 18] and metastatic
71 prostate cancer samples [19, 20]. Based on the rationale that tissue-specific covariates can influence the
72 rearrangement landscape, the cancer types from different organ systems were modeled separately.
73 Furthermore, the prostatic primary and metastatic cohorts were analyzed separately since they can have
74 distinct drivers.

75 **Breakpoint proximity curve to model genome-wide SV distribution**

76 To describe the genome-wide distribution of SV breakpoints in a given cohort, we included all unique
77 breakpoint coordinates for each sample. Then we computed the breakpoint proximity curve ($BPpc$)
78 based upon the breakpoint neighbor reachability ($BPnr_i$) for each individual breakpoint (BP_i) in the
79 cohort. This metric captures the genomic regions with high or low proximity between breakpoints
80 (Methods). The $BPpc$ is defined as the smooth curve resulting from the nonparametric local polynomial
81 regression (locally estimated scatterplot smoothing, LOESS) fitted to the dataset of $BPnr_i$, after reverse
82 scale normalization $-\log_{10}(BPnr_i+1)$ (Fig. 1a). $BPpc$ allows the use of a peak-calling approach to
83 directly identify the regions with higher breakpoint clustering relative to the surrounding area (i.e., peak
84 summits), thereby overcoming the inherent issues associated with pre-defining a genomic bin size for
85 computing breakpoint density along the genome [5, 21]. This is important because functionally relevant
86 breakpoint clustering events may occur over a wide range of genomic lengths. Thus, $BPpc$ models the
87 underlying distribution of breakpoints in a more robust and unsupervised manner compared to the
88 computation of breakpoint densities.

89 Expected *BPpc* using genomic covariates in a GAM

90 The core of the genomic breakpoint distribution arises under neutral selection due to background
 91 processes likely corresponding to the tissue-specific functional and structural genome heterogeneity. We
 92 modeled the expected background *BPpc* per tissue type using a GAM with multiple variables
 93 (Supplementary Table 2), including the tissue-specific chromatin state annotations from the Roadmap
 94 Epigenomics project [22] [23], recurrence of TADs (TAD.recurr), and TAD-boundaries (TAD-B.recurr)
 95 from the 3D Genome Browser [24], replication timing (RT) from ENCODE Repli-seq data [25] and
 96 fragile sites (FS) from the HumCFS database [26]. Other variables include the genome complexity from
 97 UCSC genome browser repeat elements data [27], gene density and GC content (GC). Thus, the
 98 expected breakpoint proximity curve is defined as $\widehat{BPpc} = \beta X + \varepsilon$, a variation of the generalized linear
 99 model in which X is a matrix of covariates (breakpoint predictors) that can contribute linearly as well as
 100 non-linearly to the response variable *BPpc* (Methods) (Fig. 1b).

101 The multivariate model explained a larger proportion of the null deviance compared to the single
 102 predictor models, even without including the interplay of co-variables (Fig. 2a). The median explained
 103 deviance is 18% and ranges from 6% for lung cancer to 50% for lymph-nodes (Fig. 2a). We expected
 104 that modeling the interaction effects of genomic features may contribute substantially to the model due
 105 to the underlying complexity of SV breakpoints in cancer [21]. Since all possible combinations of
 106 features make the model computationally infeasible and intractable for a meaningful interpretation, only
 107 the interaction terms of features that contribute at least 10% in at least one cohort in single predictor
 108 models were included (Fig 2a). The resulting model accounts for the interaction of various features,
 109 including RT, GC, TAD.recurr, TAD-B.recurr and gene density, with lamina-associated domains
 110 (LADs) that correspond to the higher-order genome disposition within the nucleus [28]. The model also

accounts for the pairwise interplay between gene density and RT, gene density and TAD.recurr as well as RT and TAD.recurr, in each case the interactions are modeled separately for distinct TAD segment classes. The TAD segments are defined based on differential conservation across cell types and annotated into three classes (quiescent, low-active or active) using enrichment of tissue-specific chromatin annotations [16] (Methods). The multivariate GAM that accounts for the interaction of co-variates showed further substantial improvement over the model without the interaction terms. The median explained deviance is 29%, ranging from 10% in lung cancer to 57% in lymph-nodes (Fig. 2a). The model performs well in all cohorts (Fig. 2c) with fairly narrow and symmetric residual distributions (Supplementary Fig. 1). The explained deviance for each cohort is not biased by the sample size as shown by its lack of correlation with the average number of SVs per donor (Supplementary Fig. 2). Notably, we find a substantial decrease in the correlation of observed vs. predicted *BPpc* when using the covariates from unmatched tissues (Fig. 2b). This clearly highlights the importance of using tissue-specific co-variates.

Contribution of genomic covariates to the background *BPpc* distribution, nonlinearity and covariates interplay.

We find that the contribution of the chromatin state annotations and repeat elements is limited while the 3D genome features of TAD-recurrence and TAD-B recurrence contribute significantly to the model for most cancer types (Fig. 2d). This result is consistent for both the genomic loci close to the membrane of the nucleus, i.e. -within LAD-, as well as the inner-nucleus genomic regions, i.e. -outside LAD- (Fig. 2d). This suggests that the 3D chromosomal conformation plays an important role in the expectation of the genome-wide distribution of SVs. The results also show that fragile sites, replication timing, GC content and gene density contribute significantly to many cancer types, though the behavior can vary for

133 within vs. outside of LADs for some cohorts (Fig. 2d). The two-dimensional (2D) contour plots allow us
 134 to visualize the partial contributions of features to $BPpc$. Most cohorts exhibit peaks and valleys on the
 135 smooth function for the partial effects of most covariates demonstrating the importance of non-linear
 136 modeling (Figs. 3a, 3b, Supplementary Fig. 3). For instance, the partial contribution of TAD-recurrence
 137 shows non-linear behavior in brain, colorectal, kidney and prostate cancers although it is linear in breast
 138 cancer regardless of the LAD status (Fig. 3a). The behavior of some features may also vary within vs.
 139 outside LADs. For example, the contribution of GC content in breast cancer is linear inside LADs but
 140 non-linear outside LADs (Fig. 3b).

141 We find a statistically significant contribution for the interaction terms of covariates across all the three
 142 functional classes of TAD segments for most cohorts (Fig. 2d). However, their effect sizes show distinct
 143 behavior for each cancer type and often across TAD segment classes as evident from the 2D contour
 144 plots (Fig. 3c, Supplementary Fig. 3). Interestingly, the distinct behavior is also apparent for primary vs.
 145 metastatic prostate cancers, likely pointing to different processes contributing to early vs. late genome-
 146 wide SV distribution [29, 30].

147 In general, there is wide variability in the partial contributions of different features across cancer types
 148 demonstrating the importance of tissue-sepcific modeling.

149 **Putative driver candidates identified at significantly recurrent peaks in $BPpc$**

150 We obtained the adjusted $B\hat{P}pc = BPpc - \widetilde{BPpc}$ by correcting the observed curve with the expected
 151 model (Fig. 1C). Values around zero in the adjusted curve correspond to the observed $BPpc$ close to the
 152 expected one from the GAM. Peaks in positive values of $B\hat{P}pc$ correspond to regions where the
 153 breakpoints are closer than expected while the valleys in negative values are loci where the breakpoints

are sparser than expected (Fig. 4a and Supplementary Fig. 4). As expected, there is a wide variability of the peaks and valleys representing the differential landscape of rearrangements for each cohort (Fig. 4a and Supplementary Fig. 4)

To identify the peaks that potentially correspond to positively selected loci across the *BPpc* rearrangement landscape, we computed the peak recurrence score (*PRs*) (Methods).

$$PRs = \frac{N_{smp}}{N_{sv}} \times \frac{Peak_A}{Peak_{GR}}$$

where *Nsmp* is the number of unique samples in the peak, *Nsv* is the number of unique SVs, *Peak_A* is the area under the peak and *Peak_{GR}* is the genome range of the peak. Thus, *PRs* is highest for peaks where many samples (high *Nsmp*) contribute same SVs (low *Nsv*) to create tight clusters of breakpoints (high *Peak_A*) over narrow regions (low *Peak_{GR}*). Next, for each cohort, we identified peaks with *PRs* significantly higher than the fitted theoretical density using a gamma distribution (FDR < 0.2) (Fig. 1d, Fig. 4b and Supplementary Fig. 5).

We identify 79 significantly recurrent peaks that potentially correspond to positively selected loci (Supplementary Table 3). The peak summits corresponding to putative driver candidates range from 179 bp to 10.71 Mb, with a median of 822.96 kb, highlighting the strength of our approach to capture breakpoint clustering over varied genomic lengths (Fig. 4c). The number of significant peaks ranges from 0 in bone cancer to a maximum of 13 in colon cancer (Supplementary Fig. 6 and Fig. 4c). Upon comparison with PCAWG results [5], we found that the novel peaks identified as significant in our analysis tend to be more cancer-type specific. Among the 42 peaks that overlap with regions of PCAWG candidates (Supplementary Table 4), 16 are cancer-type specific, while 26 were found in multiple cancer types. Conversely, 32 peaks have no overlap with PCAWG results: 25 of those peaks are cancer-type

specific and 7 peaks were found in multiple cohorts. Thus, there is a significant enrichment of cancer-type specific candidate peaks identified by our approach relative to PCAWG (chi-squared test p -value=0.0006) and further demonstrates the power of tissue-specific analysis.

Next, we predicted the functional elements that are the most likely targets of positive selection by computing the element rearrangement scores (RS_E) for all protein-coding exons, long non-coding RNAs (lncRNAs), enhancers and CTCF-insulators in the 79 significantly recurrent peaks (Figs. 1d, 1e, Supplementary Fig. 6 and Supplementary Table 5) (Methods). The highest RS_E values point to elements impacted by the largest number of SVs in the maximum number of samples within the entire peak. Across all the analyzed cancer cohorts, we identified 53 coding genes, 24 enhancers of 17 other genes and 3 lncRNAs as the most likely targets (Fig. 5) (Methods). Our results are validated by known cancer genes in different cohorts. For example, *TMPRSS2-ERG* fusion [31, 32], and *PTEN* and *TP53* deletions [33, 34] in prostate cancer. *AR* enhancer is significantly affected only in the metastatic prostate cancer cohort, which is consistent with the development of the disease [33-35]. We also find *EGFR*, *MYCN* and *MYC* in brain cancers [36, 37]; *BCL2*, *MYC* and the loci of *IGH* translocations in lymph nodes cancer [38-40]; *CDK12* in breast cancer; *CCND1* in liver cancer; and *RUNX1*, *GATA6* and *PDE4D* in esophageal cancer (Fig. 5). Besides these known cancer genes in specific cohorts, we also identified several others with known roles in multiple cancers (Fig. 5). Interestingly, most of these common candidates are large cancer genes within fragile sites, such as *FHIT*, *WWOX*, *CCSER1*, *IMMP2L*, *CDKN2A* and *CDKN2B* [26, 41-44].

Overall, out of the 73 genes whose exons or enhancers are identified as putative drivers, 47 are known cancer genes (Fig. 5). Genes identified as potential drivers in our analysis that are known to be oncogenic in another cancer type [45, 46] are of high interest since they are often therapeutic targets

under investigation (Fig. 6). For example, *DMD* is known to be a cancer gene [45] in human myogenic tumors, such as gastrointestinal stromal tumors [47], rhabdomyosarcoma [48] and leiomyosarcoma [49], and we find it is a putative driver gene in esophageal cancer where 54% of the cohort carries SVs in this region (Fig. 6a). We also identified *LRRN3* as a driver candidate in esophageal cancer with SVs in 20.7% of the samples (Fig. 6b). Other members of the *LRRN* gene family have been found as drivers in multiple cancer types, including neuroblastoma and gastric cancer [50], and the role of *LRRN3*, in particular, has been studied in fibrosarcoma [51]. As shown in Fig. 6, clear peaks in *BPpc* correspond to tight clustering of breakpoints allowing us to confidently pinpoint these genes as putative drivers.

Another interesting result is our finding in ovarian cancer of a region affected in 25% of the cohort where our method points to *NF1* as a potential driver (Fig. 6c). Previous studies have shown the importance of *NF1* in several other cancers including glioblastoma, melanoma, breast and lung cancers [36, 52-54]. The gene *CDK12* is a candidate within a region affected in 14.9% of the uterine cancer cohort (Fig. 6d). *CDK12* is a well-studied target in other female reproductive cancers, such as ovarian and breast [55], and it has also been studied in stomach and prostate cancers [20, 56]. In colorectal cancer, we predict *WWOX* as a putative driver in a region that shows SVs in 25% of the samples (Fig. 6e). *WWOX* is within a known fragile site and has been reported in several studies of different tumor types including breast, prostate, lung, esophagus, cervical, ovarian and bladder cancers [57-61].

Limited availability of RNA-seq data prevented analysis of impact on gene expression for many candidates (Supplementary Table 6). Among the cohorts with sufficient sample sizes for RNA-seq analysis, we find that an enhancer of *CTDSP2* is a putative driver in brain cancer (glioblastoma multiforme) and the SVs are associated with its differential expression (Fig. 6f). *CTDSP2* is known to be important in other cancer types [62, 63]. The genes *MACROD2* in pancreatic cancer (Fig. 6h) and

219 *CDKN2B* in liver cancer (Fig. 6g) also show significant differential expression between samples with
220 SVs at these loci relative to those without SVs.

221 These findings highlight the importance of both the tissue-specific and clustering-based approaches used
222 in our method to capture the significantly rearranged regions in different cancers. Notably, while
223 breakpoint clustering clearly points to these specific candidates as seen in Fig. 6, breakpoint density over
224 fixed bins would require different bin sizes to capture the ones with maximum density. This is because
225 using a fixed-size window for all these instances will likely generate several bins with high density
226 without clearly pointing to the most probable target element of selection.

227 **CSVDriver: Computational tool to identify SV drivers from whole-genome sequences**

228 The computational approach developed in this work to identify SVs and functional elements under
229 positive selection in cancer whole-genomes is implemented in CSVDriver, Cancer Structural Variation
230 Driver, a user-friendly shiny app tool (<https://github.com/khuranalab/CSVDriver>). The input for the tool
231 is SV calls and researchers can provide the tissue-specific co-variates data or use existing datasets to run
232 the tool on their cancer cohort/s. Besides the list of functional elements that are putative drivers in a
233 given cohort, the app provides the graphical visualization of the GAM results including the analysis of
234 non-linearity and covariates interaction.

235 **Discussion**

236 One of the major challenges in cancer genomics is the accurate estimation of the expected
237 heterogeneous distribution of passenger variants. This distribution represents the null background, which
238 can be used to identify loci that exhibit significantly recurrent variants likely due to positive selection.
239 While this problem has received considerable attention for SNVs, studies for tissue-specific neutral

models of the genomic distribution of SVs are lacking. We find that a GAM is able to describe the breakpoint proximity distribution of SVs in cancer genomes with the explained deviance ranging from 10% in lung to 57% in lymph-nodes cancer. The use of a GAM allows us to interpret the results graphically and provides estimates for both the magnitude and statistical significance of the contribution of each feature. We find that the 3D chromosomal conformation plays an important role in the genome-wide distribution of SVs for most cancer types with TAD-recurrence, TAD-B recurrence and their interaction terms with other covariates contributing significantly to the model.

Our method is able to identify the known cancer drivers and nominate novel candidates as those that exhibit higher breakpoint proximity than expected by random chance, such as *DMD* and *LRRN3* in esophageal cancer, *NF1* in ovarian cancer, *CDK12* in uterine cancer and *WWOX* in colorectal cancer. We note that the pan-cancer analysis by PCAWG allowed the identification of regions that are not likely to gain significance in single cancer analyses due to limited cohort sizes but failed to identify many novel regions that gain significance in our tissue-specific analysis. While our method is able to identify 24 enhancers as putative drivers, they are usually in the vicinity of other coding exons with similarly high element rearrangement scores and further functional validation will be needed to decipher their role in tumorigenesis. However, one prominent example supported by RNA-seq data is *CTDSP2* enhancer that is a candidate driver in glioblastoma.

Finally, while we could identify the features that contribute significantly to the breakpoint proximity curve, we find that the relationships are usually tissue-specific, complex and non-linear, often forbidding straightforward interpretations. Furthermore, while we focused our analysis on the peaks in *BPpc*, the valleys could potentially provide insights about the loci showing depletion of breakpoints due to negative selection in future studies.

262 **Methods**

263 **Cancer somatic structural variations data**

264 All somatic SVs were obtained from the PCAWG project Working Group 6 consensus calls [1, 18]. We
 265 selected the cohorts based on the median number of SVs per sample aiming to get data representative of
 266 the cancer types more impacted by genomic rearrangements. It covers 15 different organs systems
 267 including 32 distinct histological cancer subtypes. They are Skin: MELA (Melanoma), SKCM (Skin
 268 Cutaneous melanoma); Pancreas: PACA (Pancreatic Cancer), PAEN (Pancreatic Endocrine Neoplasms);
 269 Lymphatic system: DLBC (Lymphoid Neoplasm Diffuse Large B-cell Lymphoma), MALY (Malignant
 270 Lymphoma); Esophagus: ESAD (Esophageal Adenocarcinoma); Lung: LUAD (Lung Adenocarcinoma),
 271 LUSC (Lung Squamous cell carcinoma); Uterus: UCEC (Uterine Corpus Endometrial Carcinoma);
 272 Ovary: OV (Ovarian Cancer); Breast: BRCA (Breast Cancer); Brain (central nervous system): PBCA
 273 (Pediatric Brain Cancer), LGG (Brain Lower Grade Glioma), GBM (Brain Glioblastoma Multiforme);
 274 Bone: SARC (Sarcoma), BOCA (Bone Cancer); Colorectal: COAD (Colon Adenocarcinoma), READ
 275 (Rectum Adenocarcinoma); Kidney: KICH (Kidney Chromophobe), KIRC (Kidney Renal Clear Cell
 276 Carcinoma), KIRP (Kidney Renal Papillary Cell Carcinoma), RECA (Renal clear cell carcinoma);
 277 Liver: LICA (Liver Cancer), LIRI (Liver Cancer - RIKEN), LIHC (Liver Hepatocellular carcinoma),
 278 LINC (Liver Cancer - NCC); Stomach: GACA (Gastric Cancer), STAD (Gastric Adenocarcinoma);
 279 Prostate: EOPC (Early Onset Prostate Cancer), PRAD (Prostate Adenocarcinoma). Particularly for
 280 prostate cancer, we additionally collected SVs from a cohort of 124 metastatic samples (37,9605 SVs)
 281 obtained from calls reported in [19] and [20]. These datasets provide information from 2,382 high-
 282 quality WGS cancer donors allowing us to analyze 324,838 high confidence unique somatic SVs. We

analyzed SVs in autosomes and chromosome X only. The genome reference used is hg19 build. The details are shown in Supplementary Table 1.

Genomic feature annotations and tissue-specific epigenomic covariates

We use several genome features as the *BP* covariates in the modeling of the expected background of *BPpc*. They include tissue-specific chromatin state marks from Roadmap Epigenomics Mapping Consortium [23], which annotates genomic regions based on histone modifications and chromatin DNA accessibility [64]. We used the average RT signal per 1Mb genomic bins across eight different cell types (liver, breast, brain, lymphoma system, skin, blood, lung and prostate) as described in the CNC-Driver method [11]. The datasets were collected from ENCODE and constitute wavelet-smoothed Repli-seq data [25]. The GC content was computed in windows of 101 bp centered in each BP using the function ‘GCcontent’ from the R Package ‘biovizBase’ [65]. The gene density was computed per 1 Mb window using the function *kpPlotDensity* from the R Package *kpPlotDensity* [66]. The chromosomal Fragile Sites (FS) annotation was obtained from the database HumCFS [26].

We included two higher-order structural 3D chromosomal conformations, TADs and LADs, in the set of covariates. For TADs, we use the recurrence of the regions as well as their boundaries (TAD-B) annotated across 37 datasets obtained from the 3D Genome Browser [24]. We also classify each sub-segment of TAD that shows different recurrence by computing the coverage of all 15 chromatin marks within every TAD segment, similar to the approach used in [16] (Supplementary Fig. 7). We found that three principal components explain most of the variance of this coverage (Supplementary Fig. 8). Thus, we grouped the TAD segments in three clusters according to the coverage of chromatin marks. The three clusters are quiescent/heterochromatin, low-active and active. We confirm that the clustering is robust

304 across the 16 cohorts analyzed in this study (Supplementary Fig. 9). More details of the sources of
305 genomic feature annotations are shown in Supplementary Table 2.

306 The putative functional effect for each significantly rearranged locus is predicted by annotating the
307 potential drivers on the basis of the impact of the SVs breakpoints on the coding and noncoding
308 elements within these regions. For the coding drivers, we use the subset of protein-coding genes
309 extracted from the comprehensive gene list obtained from the Genecode Release 29 (GRCh37). For the
310 non-coding elements, we use the active tissue-specific cCRE marks gathered from ENCODE 3 [25].

311 **CSVDriver workflow.**

312 CSVDriver aims to identify cancer-driving rearrangement events by analyzing the focal trend of
313 breakpoint clustering.

314 Input and preprocessing of data

315 The method takes as input the data of cancer somatic SVs calls and analyzes the combined impact of all
316 rearrangement types including insertions (Ins), deletions (Del), inversions (Inv) and translocations (Tra).
317 The expected file format is SV-table containing 12 columns named as: (cohort_code, donor_id,
318 variant_type, sv_id, chr_from, chr_from_bkpt, chr_from_bkpt, chr_from_strand, chr_to, chr_to_bkpt,
319 chr_to_bkpt, chr_to_strand). Each row defines a single SV by the cohort (cancer type), the donor ID, the
320 SV type (Ins, Del, Inv, Tra), the SVs unique ID and the genome coordinate [chr, start, end, strand] for
321 both rearrangement breakpoints (from, to). CSVDriver checks SVs from all donors and decomposes this
322 SV-table into a BP-table sorted by genomic position per chromosome.

323 Establishing the observed breakpoint proximity-curve (BPpc)

The method bases the SV analysis on the description of the genome-wide distribution of breakpoint proximity taking all unique breakpoint coordinates for each distinct sample in the given cohort. Thus, if a breakpoint coordinate is present in two samples, it is represented twice. Then we sort all breakpoints based on their coordinates and obtained an ordered list of breakpoints encompassing all samples, each one represented by BP_i , where i is the ordered index. Next, we annotate each BP_i by computing their neighbor reachability ($BPnr_i$) defined as the average genome distance to reach the two adjacent breakpoints:

$$BPnr_i = (\Delta(BP_i, BP_{i-1}) + \Delta(BP_i, BP_{i+1}))/2.$$

Thereupon we define the breakpoint proximity (BPp_i) as the normalized reverse scale, applying logarithmic transformation: $BPp_i = -\log_{10}(BPnr_i + 1)$. Then, we compute the breakpoint proximity curve ($BPpc$), which is a smooth curve resulting from the nonparametric local polynomial regression (LOESS) fitted to the $BPnr_i$ values (Fig. 1A). This curve shows the trend of focal clustering of the breakpoints because the fitting result is weighted toward the nearest surrounding values. The span argument ($\alpha = 0.2$) controls the size of the near surrounding. It reflects the interval as a proportion of the total breakpoints and regulates the grade of smoothness in the resulting $BPpc$. This curve depicts the breakpoint genomic distribution and captures the regions of high and low proximity between breakpoints. Chromosomes with a low load of breakpoints that show no clear trend of clustering will not reflect a reliable signal. Therefore, we do not consider them. That is the case for small chromosomes (e.g. chr21, chr22) in some cancer cohorts.

Modeling the expected BPpc by using generalized additive model (GAM) with tissue-specific genomic covariates

345 The model is conceived to expand the linear regression analysis of genomic covariates by introducing
 346 the capacity to investigate the potential non-linear relationships between genomic covariates and the
 347 distribution of breakpoints. Additionally, it accounts for the contribution of the covariates' interaction to
 348 the model. Thus, CSVDriver models the expected background $BPpc$ using GAM, a parametric
 349 regression method, which models the $BPpc$ (dependent variable) with respect to tissue-specific data of
 350 genomic covariates (predictors or independent covariates).

351 GAM is a flexible extension of generalized linear models (GLM). Using a GAM, we can fit a linear
 352 model, which allows us to consider either linear or non-linear contributions of the genomic covariates to
 353 the model of $BPpc$. To fit the GAM, CSVDriver uses the R package 'mgcv' version 1.8-28 [67, 68]. The
 354 values of the observed $BPpc$ for each cohort fit gamma distribution (Supplementary Fig. 10). Therefore,
 355 the model assumes each expected $BPpc$ to be generated from a gamma distribution for the identity link
 356 function of the response. The model follows the equation:

$$\begin{aligned} 357 \quad \widehat{BPpc}_i = & \beta_0 + FS_i + LAD_i + repClass_i + ChromMark_i + s(RT_i)LAD_i + s(GC_i)LAD_i + \\ 358 \quad & s(TAD.recurr_i)LAD_i + s(TADB.recurr_i)LAD_i + s(GeneDensity)LAD_i + \\ 359 \quad & te(GeneDensity_i, TAD.recurr_i)TADsegm_{class_i} + te(GeneDensity_i, RT_i)TADsegm_{class_i} + \\ 360 \quad & te(RT_i, TAD.recurr_i)TADsegm_{class_i} \end{aligned}$$

361 where $i=1, \dots, N$ (total number of breakpoints), $\widehat{BPpc}_i$ is the expected breakpoints proximity, β_0 is the
 362 intercept and the remaining terms are the genomic covariates used as predictors. The covariates FS,
 363 LAD, repClass, and ChromMark are modeled as linear factorial terms. The thin plate regression spline
 364 smooth function (s) can describe the non-linearity in the contribution of the covariates (RT, GC,
 365 GeneDensity, TAD, and TAD-B recurrence) and for each one the model gets the interaction with the
 366 status of the factor LAD. The model investigates the main effects of the predictors as well as the effects

of the tensor product interaction (te) for the interplay between GeneDensity and TAD recurrence; GeneDensity and RT; and TAD and RT. This accounts for the status of TAD-segment class. Nonetheless, a high number of interaction terms increases the chance of over-fitting. Furthermore, GAM can be computationally expensive to reach convergence. Consequently, we try to balance the complexity of the model and its ability to explain the deviance by covariate interactions reasonably by including the interaction only of covariates that contribute at least 10% in at least one cohort in single predictor models.

374 Computing the adjusted $BPpc$

The goal of CSVDriver is to capture the loci where breakpoint clusters potentially arise due to selective pressure unlike the clusters associated with neutral non-selective forces. Therefore, the adjusted curve $\dot{BPpc} = BPpc - \widetilde{BPpc}$ represents a corrected BPpc where regions with values close to the $\widetilde{BPpc}$ will be considered expected biases. The signal in the y-axis depicts how close the breakpoints are in a given region (Fig. 1c). The positive values represent regions where the breakpoints are closer than expected while regions with negative values are loci where the breakpoints are sparser than expected from the GAM. The values around zero in the adjusted curve ($\dot{BPpc}$) correspond to regions of observed BPpc close to the expected one.

383 Detecting the significantly rearranged peaks in the $BPpc$

The method takes the regions corresponding to the top 25% of the summit of each peak, which represent the regions of local maximum breakpoint clustering. Then we identify the loci that potentially correspond to the positively selected regions (Fig. 1d). Each peak is described by their peak recurrence score (PRs) defined as:

$$PRs = \frac{N_{smp}}{N_{sv}} \times \frac{Peak_A}{Peak_{GR}}$$

where N_{smp} is the number of unique samples in the peak, N_{sv} is the number of unique SVs, $Peak_A$ is the area under the peak and $Peak_{GR}$ is the genome range of the peak. Thus, PRs is the highest for peaks where many samples (high N_{smp}) contribute to create tight clusters of breakpoints (high $Peak_A$) over narrow regions (low $Peak_{GR}$). The score is square root transformed to reduce the dispersion in the values while keeping the trend of interest. We take into account the inherent limitation of the background model and the differential explained deviance per tissue type (Fig. 2a). Consequently, the approach does not use any absolute pre-established threshold, instead, it considers the cohort-specific distributions of all PRs (Fig. 4b density distribution) as the expectation of the combined effect of several processes. Next, for each cohort, we identify peaks with PRs significantly higher (Fig. 4b, QQ-plots) than the fitted gamma distribution theoretical density using the ‘fitdistrplus’ R Package [69]. After controlling the false discovery rate (FDR) [70] for multiple hypothesis testing, the significant loci were defined as those with $FDR < 0.2$.

Detecting the driver candidates within the significantly rearranged peaks

The potential functional effect of the significantly rearranged regions for each cancer is directly associated with the effects on coding and non-coding elements by their deletion, disruption or relocation. CSV-Driver determines the functional elements that are the most likely targets of positive selection within each significantly rearranged peak.

For each functional element (protein-coding exons, lncRNA, enhancer, CTCF-insulator) located at a significantly rearranged peak, the method computes the element rearrangement-score (RS_E)

$$RS_E = \frac{N_{smp_E}}{N_{smp}} \times \frac{N_{sv_E}}{L_E}$$

where N_{smp_E} is the number of unique samples that have SVs overlapping the element, N_{smp} is the number of unique samples in the peak, N_{sv_E} is the number of SVs overlapping the element and L_E is the length of the element. High N_{sv_E} values could be due to multiple alleles affected or multiple clonal/sub-clonal cell populations with SVs at the given locus. Normalization by L_E takes into account that longer elements are more likely to have more SVs randomly. RS_E is a metric of the relative selective pressure acting upon elements within the significantly rearranged region. It takes into account the relative recurrence in samples and the number of SVs that impact the element normalized by the length of the element. Due to the larger genomic length of genes and different mechanisms of SVs that can impact their function, i.e. (a) amplification, deletion, relocation or (b) gene fusion, CSVDriver computes two rearrangement scores RS_E for coding genes under significant peaks. One score accounts for the number of SVs that engulf the entire gene and the other accounts for the number of breakpoints that fall within the gene's body. The genes with the highest RS_E are the most likely targets of positive selection.

The method reports the element rearrangement scores for the highest scoring elements of all types (protein-coding exons, enhancer or lncRNA) at a given peak (Supplementary Fig. 6 and Supplementary Table 5). While at a specific locus altered by SVs, it is fair to assume that affected coding exons of genes are more likely to have greater impact than the non-coding elements, CSV-Driver allows the analysis of the potential importance of all possible driver elements. One peak may contain multiple driver elements which can represent alternative paths of disrupted regulation, or even some subgroup of samples with slightly different, yet related genomic drivers. In the current analysis, if a peak contains multiple elements of different type, all elements with highest rearrangement scores are reported in Supplementary Fig. 6 and Supplementary Table 5. However, if one of those elements is or is associated

430 with a known cancer gene, we consider it as the most likely candidate in a given peak shown in Fig. 5.

431 Results report

432 For each input cohort, CSVDriver reports the graphics of the *BPpc* annotated with the driver candidates
433 (coding genes and non-coding elements) at each significantly rearranged peak (Supplementary Fig. 6). It
434 also provides the summary tables with the catalog of drivers (Supplementary Table 5). The tool also
435 provides the set of plots resulting from GAM covariates modeling to analyze the non-linearity of
436 relationships between the genomic features and the genomic distribution of breakpoints (Supplementary
437 Fig. 3)

438 **Verification of the driver candidates**

439 We verified our results of driver candidates using CancerMine [45] and COSMIC [46]. In addition, for
440 the significantly rearranged peaks that have enough samples with RNA-seq, we check for the significant
441 differential expression between samples with SVs at the given locus relative to the samples without SVs.

442 **Acknowledgement**

443 This work is supported by NIH grant R01CA218668 to E.K.

444 **Author Contributions**

445 E.K. and A.MF. conceived and designed the project. A.MF. implemented the method and performed the
446 data analysis. A.D. performed the RNA-seq differential expression analysis. A.MF. and E.K. wrote the
447 manuscript. E.K. supervised the project.

448 **Competing Interests statement**

449 The authors declare no competing interests.

References

1. Li, Y., et al., *Patterns of somatic structural variation in human cancer genomes*. Nature, 2020. **578**(7793): p. 112-121.
2. Yang, L., et al., *Diverse mechanisms of somatic structural variations in human cancer genomes*. Cell, 2013. **153**(4): p. 919-29.
3. Maciejowski, J. and M. Imielinski, *Modeling cancer rearrangement landscapes*. Curr Opin Syst Biol, 2017. **1**: p. 54-61.
4. Wang, W.J., L.Y. Li, and J.W. Cui, *Chromosome structural variation in tumorigenesis: mechanisms of formation and carcinogenesis*. Epigenetics Chromatin, 2020. **13**(1): p. 49.
5. Rheinbay, E., et al., *Analyses of non-coding somatic drivers in 2,658 cancer whole genomes*. Nature, 2020. **578**(7793): p. 102-111.
6. Zhang, Y., et al., *Global impact of somatic structural variation on the DNA methylome of human cancers*. Genome Biol, 2019. **20**(1): p. 209.
7. Khurana, E., et al., *Role of non-coding sequence variants in cancer*. Nat Rev Genet, 2016. **17**(2): p. 93-108.
8. Zhang, Y., et al., *A Pan-Cancer Compendium of Genes Deregulated by Somatic Genomic Rearrangement across More Than 1,400 Cases*. Cell Rep, 2018. **24**(2): p. 515-527.
9. Hadi, K., et al., *Distinct Classes of Complex Structural Variation Uncovered across Thousands of Cancer Genome Graphs*. Cell, 2020. **183**(1): p. 197-210 e32.
10. Chen, H., et al., *Comprehensive assessment of computational algorithms in predicting cancer driver mutations*. Genome Biol, 2020. **21**(1): p. 43.
11. Liu, E.M., et al., *Identification of Cancer Drivers at CTCF Insulators in 1,962 Whole Genomes*. Cell Syst, 2019. **8**(5): p. 446-455 e8.
12. Liu, E.M., et al., *CNCDatabase: a database of non-coding cancer drivers*. Nucleic Acids Res, 2021. **49**(D1): p. D1094-D1101.
13. Polak, P., et al., *Cell-of-origin chromatin organization shapes the mutational landscape of cancer*. Nature, 2015. **518**(7539): p. 360-364.
14. Yi, K. and Y.S. Ju, *Patterns and mechanisms of structural variations in human cancer*. Exp Mol Med, 2018. **50**(8): p. 1-11.
15. Dixon, J.R., et al., *Integrative detection and analysis of structural variation in cancer genomes*. Nat Genet, 2018. **50**(10): p. 1388-1398.
16. Akdemir, K.C., et al., *Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer*. Nat Genet, 2020. **52**(3): p. 294-305.
17. Wang, X., et al., *Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes*. Nat Methods, 2021. **18**(6): p. 661-668.
18. Consortium, I.T.P.-C.A.o.W.G., *Pan-cancer analysis of whole genomes*. Nature, 2020. **578**(7793): p. 82-93.
19. Quigley, D.A., et al., *Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer*. Cell, 2018. **175**(3): p. 889.
20. Viswanathan, S.R., et al., *Structural Alterations Driving Castration-Resistant Prostate Cancer Revealed by Linked-Read Genome Sequencing*. Cell, 2018. **174**(2): p. 433-447 e19.
21. Cheloshkina, K. and M. Poptsova, *Comprehensive analysis of cancer breakpoints reveals signatures of genetic and epigenetic contribution to cancer genome rearrangements*. PLoS Comput Biol, 2021. **17**(3): p. e1008749.
22. Ernst, J. and M. Kellis, *Chromatin-state discovery and genome annotation with ChromHMM*. Nat Protoc, 2017. **12**(12): p. 2478-2492.
23. Roadmap Epigenomics, C., et al., *Integrative analysis of 111 reference human epigenomes*. Nature, 2015. **518**(7539): p. 317-30.

- 495 24. Wang, Y., et al., *The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-*
496 *range chromatin interactions.* Genome Biol, 2018. **19**(1): p. 151.
- 497 25. Consortium, E.P., et al., *Expanded encyclopaedias of DNA elements in the human and mouse genomes.* Nature,
498 2020. **583**(7818): p. 699-710.
- 499 26. Kumar, R., et al., *HumCFS: a database of fragile sites in human chromosomes.* BMC Genomics, 2019. **19**(Suppl
500 9): p. 985.
- 501 27. Navarro Gonzalez, J., et al., *The UCSC Genome Browser database: 2021 update.* Nucleic Acids Res, 2021.
502 **49**(D1): p. D1046-D1057.
- 503 28. Briand, N. and P. Collas, *Lamina-associated domains: peripheral matters and internal affairs.* Genome Biol,
504 2020. **21**(1): p. 85.
- 505 29. Quigley, D.A., et al., *Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer.* Cell, 2018.
506 **174**(3): p. 758-769 e9.
- 507 30. Fraser, M., et al., *Genomic hallmarks of localized, non-indolent prostate cancer.* Nature, 2017. **541**(7637): p. 359-
508 364.
- 509 31. Tomlins, S.A., et al., *Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer.* Science,
510 2005. **310**(5748): p. 644-8.
- 511 32. Tomlins, S.A., et al., *Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate*
512 *cancer.* Nature, 2007. **448**(7153): p. 595-9.
- 513 33. Liu, W., et al., *Genetic markers associated with early cancer-specific mortality following prostatectomy.* Cancer, 2013.
514 **119**(13): p. 2405-12.
- 515 34. Hieronymus, H., et al., *Copy number alteration burden predicts prostate cancer relapse.* Proc Natl Acad Sci U S
516 A, 2014. **111**(30): p. 11139-44.
- 517 35. Takeda, D.Y., et al., *A Somatic Acquired Enhancer of the Androgen Receptor Is a Noncoding Driver in*
518 *Advanced Prostate Cancer.* Cell, 2018. **174**(2): p. 422-432 e13.
- 519 36. Verhaak, R.G., et al., *Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized*
520 *by abnormalities in PDGFRA, IDH1, EGFR, and NF1.* Cancer Cell, 2010. **17**(1): p. 98-110.
- 521 37. Lopez-Gines, C., et al., *Association of chromosome 7, chromosome 10 and EGFR gene amplification in glioblastoma*
522 *multiforme.* Clin Neuropathol, 2005. **24**(5): p. 209-18.
- 523 38. Zhang, X.M., et al., *[Detection of bcl-2/IgH fusion gene in lymphoma by real-time polymerase chain reaction and its*
524 *clinical significance].* Zhongguo Shi Yan Xue Ye Xue Za Zhi, 2009. **17**(2): p. 368-72.
- 525 39. Miao, Y., et al., *CCND1-IGH Fusion-Amplification and MYC Copy Number Gain in a Case of Pleomorphic*
526 *Variant Mantle Cell Lymphoma.* Am J Clin Pathol, 2016. **146**(6): p. 747-752.
- 527 40. Einerson, R.R., et al., *Novel FISH probes designed to detect IGH-MYC and IGL-MYC rearrangements in B-cell*
528 *lineage malignancy identify a new breakpoint cluster region designated BVR2.* Leukemia, 2006. **20**(10): p. 1790-
529 9.
- 530 41. Yang, F., et al., *FHIT and C-MYC expression in cervical histology and cytology as biomarkers for detecting high-*
531 *grade intraepithelial neoplasia in human papillomavirus-positive women.* Cancer Biomark, 2020. **27**(3): p. 303-
532 312.
- 533 42. Tanna, M. and R.I. Aqeilan, *Modeling WWOX Loss of Function in vivo: What Have We Learned?* Front
534 Oncol, 2018. **8**: p. 420.
- 535 43. Fan, Y., et al., *Desmoplastic Melanoma: A Clinicopathological Analysis of Three Cases in the Chinese Population.*
536 Onco Targets Ther, 2021. **14**: p. 2651-2660.
- 537 44. Iwasaki, M., et al., *CD93 Marks a Non-Quiescent Human Leukemia Stem Cell Population and Is Required for*
538 *Development of MLL-Rearranged Acute Myeloid Leukemia.* Cell Stem Cell, 2015. **17**(4): p. 412-21.
- 539 45. Lever, J., et al., *CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer.* Nat
540 Methods, 2019. **16**(6): p. 505-507.
- 541 46. Sondka, Z., et al., *The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers.* Nat
542 Rev Cancer, 2018. **18**(11): p. 696-705.

- 543 47. Li, K., et al., *Genetic progression in gastrointestinal stromal tumors: mechanisms and molecular interventions*.
544 Oncotarget, 2017. **8**(36): p. 60589-60604.
- 545 48. Dumont, N.A., et al., *Dystrophin expression in muscle stem cells regulates their polarity and asymmetric division*.
546 Nat Med, 2015. **21**(12): p. 1455-63.
- 547 49. Chang, N.C., F.P. Chevalier, and M.A. Rudnicki, *Satellite Cells in Muscular Dystrophy - Lost in Polarity*.
548 Trends Mol Med, 2016. **22**(6): p. 479-496.
- 549 50. Liu, B., et al., *Leucine-rich repeat neuronal protein-1 suppresses apoptosis of gastric cancer cells through regulation of*
550 *Fas/FasL*. Cancer Sci, 2019. **110**(7): p. 2145-2155.
- 551 51. Fukamachi, K., et al., *Rat neuronal leucine-rich repeat protein-3: cloning and regulation of the gene expression*.
552 Biochem Biophys Res Commun, 2001. **287**(1): p. 257-63.
- 553 52. Redig, A.J., et al., *Clinical and Molecular Characteristics of NF1-Mutant Lung Cancer*. Clin Cancer Res, 2016.
554 **22**(13): p. 3148-56.
- 555 53. Uusitalo, E., et al., *Breast cancer in neurofibromatosis type 1: overrepresentation of unfavourable prognostic factors*.
556 Br J Cancer, 2017. **116**(2): p. 211-217.
- 557 54. Larribere, L., et al., *NF1-RAC1 axis regulates migration of the melanocytic lineage*. Transl Oncol, 2020. **13**(12):
558 p. 100858.
- 559 55. Liang, S., et al., *CDK12: A Potent Target and Biomarker for Human Cancer Therapy*. Cells, 2020. **9**(6).
- 560 56. Liu, H., et al., *CDK12 and PAK2 as novel therapeutic targets for human gastric cancer*. Theranostics, 2020.
561 **10**(14): p. 6201-6215.
- 562 57. Pospiech, K., E. Pluciennik, and A.K. Bednarek, *WWOX Tumor Suppressor Gene in Breast Cancer, a*
563 *Historical Perspective and Future Directions*. Front Oncol, 2018. **8**: p. 345.
- 564 58. Qin, H.R., et al., *A role for the WWOX gene in prostate cancer*. Cancer Res, 2006. **66**(13): p. 6477-81.
- 565 59. Dutta, M., et al., *Whole genome sequencing analysis identifies recurrent structural alterations in esophageal squamous*
566 *cell carcinoma*. PeerJ, 2020. **8**: p. e9294.
- 567 60. Srivastava, S., et al., *Downregulated Expression of WWOX in Cervical Carcinoma: A Case-Control Study*. Int J
568 Mol Cell Med, 2020. **9**(4): p. 273-288.
- 569 61. Kolat, D., Z. Kaluzinska, and E. Pluciennik, *Fragile Gene WWOX Guides TFAP2A/TFAP2C-Dependent*
570 *Actions Against Tumor Progression in Grade II Bladder Cancer*. Front Oncol, 2021. **11**: p. 621060.
- 571 62. Krasnov, G.S., et al., *Tumor suppressor properties of the small C-terminal domain phosphatases in non-small cell*
572 *lung cancer*. Biosci Rep, 2019. **39**(12).
- 573 63. Winans, S., et al., *Integration of ALV into CTDSPL and CTDSPL2 genes in B-cell lymphomas promotes cell*
574 *immortalization, migration and survival*. Oncotarget, 2017. **8**(34): p. 57302-57315.
- 575 64. Ernst, J. and M. Kellis, *ChromHMM: automating chromatin-state discovery and characterization*. Nat Methods,
576 2012. **9**(3): p. 215-6.
- 577 65. Yin, T., M. Lawrence, and D. Cook, *biovizBase: Basic graphic utilities for visualization of genomic data*. 2018:
578 R package version 1.30.1.
- 579 66. Gel, B. and E. Serra, *karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary*
580 *data*. Bioinformatics, 2017. **33**(19): p. 3088-3090.
- 581 67. Wood, S.N., *Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric*
582 *generalized linear models*. Journal of the Royal Statistical Society Series B-Statistical Methodology, 2011.
583 **73**(1): p. 3-36.
- 584 68. Wood, S.N., *Generalized Additive Models: An Introduction with R*. 2 ed. 2017: Chapman and Hall/CRC.
- 585 69. Delignette-Muller, M.L. and C. Dutang, *fitdistrplus: An R Package for Fitting Distributions*. Journal of
586 Statistical Software, 2015. **64**(4): p. 1--34.
- 587 70. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate - a Practical and Powerful Approach to*
588 *Multiple Testing*. Journal of the Royal Statistical Society Series B-Statistical Methodology, 1995. **57**(1):
589 p. 289-300.

Figure 1

Input: Structural variation from tumor whole-genome sequencing

Figure1: CSVDriver workflow. The input data is the standard cancer somatic SV calls that include a pair of breakpoint coordinates (BP1, BP2) for each SV_id per sample. **(a)** Step 1 computes the breakpoint neighbor reachability (BPnri grey dots) for each breakpoint (BP_i), where ' i ' is the sorted index. Then, for each chromosome, the $BPpc$ is generated as the smooth curve (grey line) resulting from the nonparametric local polynomial regression fitted to the dataset of $BPnr_i$, after reverse scale normalization $-\log_{10}(BPnr_i+1)$. **(b)** In step 2, based on the observed $BPpc$ distribution (solid line), the method assesses the expected ($\widehat{BPpc}$) background distribution (dashed line) by using a generalized additive model (GAM) that includes multiple tissue-specific breakpoint covariates. **(c)** In step 3, it computes adjusted $BPpc$ (observed – expected). **(d)** In step 4, the method calls peaks across the adjusted $BPpc$ and identifies those that potentially correspond to positively selected loci. It computes the peak recurrence score (PRs) and based on the empirical density of PRs (dashed line distribution) it identifies the peaks with PRs significantly higher (QQ-plot FDR < 0.2) than the fitted theoretical density (red line distribution). **(e)** In the last step 5, the driver candidates are identified as the genomic elements (CDS (coding sequence), enhancer, CTCF-Insulator, and lncRNA) with the highest rearrangement scores within the peak region.

Figure 2

Figure 2: Performance of the GAM for the expected background *BPpc* for each cancer type. **(a)** Explained deviance of the GAM for the multivariate model with and without interplay and single covariate models. **(b)** Correlation plot of the performance of the model for each cohort using different tissue-specific covariates. **(c)** Observed *BPpc* distributions and heatmap of the tail probabilities. Black curve is the predicted *BPpc* distribution from the GAM model. **(d)** Heatmap of the significance of the partial contribution of each covariate to the explained deviance of the model with respect to the mean of the distribution (null intercept).

FS (fragile sites), LAD (lamina associated domain), DNA sequence repeat classes that include LTR (long terminal repeat elements), LINE (long interspersed nuclear elements), SINE (short interspersed nuclear elements) including ALUs, LCR (low complexity repeats), simple repeats i.e. micro-satellites, DNA repeat elements (rDNA), RNA repeats (including tRNA, rRNA, snRNA, scRNA, srpRNA), satellite repeats, other repeats including class RC (Rolling Circle) and Unknown complex sequence.

chromHMM tissue-specific chromatin marks that include TssA (Active TSS); TssAFlnk (Flanking Active TSS); TxFlnk (transcription at gene 5' and 3'); Tx (Strong transcription); TxWk (Weak transcription); EnhG (Genic enhancers); Enh (Enhancers); ZNF/Rpts (ZNF genes & repeats); Het (Heterochromatin); TssBiv (Bivalent/Poised TSS); BivFlnk (Flanking Bivalent TSS/Enh); EnhBiv (Bivalent Enhancer); ReprPC (Repressed PolyComb); ReprPCWk (Weak Repressed PolyComb).

The 's' is a thin plate regression spline smooth function that describes the non-linearity in the contribution of the replication timing (RT), GC content (GC), gene density (gene.density), and the recurrence of the topologically associated domain (TAD.recurr) as well as the recurrence of TAD boundary regions (TAD-B.recurr). The partial contribution of each 's' function accounts for the interaction with the corresponding status of LAD. The 'te' is a tensor product interaction that describes the interplay between gene.density vs. TAD.recurr, gene.density vs. RT, and TAD.recurr vs. RT. The partial contribution of each 'te' interplay accounts for the interaction with the corresponding class of TAD-segment that includes quiescent, low-active, and active regions.

Figure 3

Figure 3: Graphics to analyze the results of the partial contribution of covariates to the GAM shown for representative cancer types. **(a)** the smooth function of partial contribution of TAD recurrence. **(b)** the smooth function of the partial contribution of GC content. These graphics represent the linear or non-linear behavior of the partial contribution of each covariate analyzed. The x-axis is the value of the covariate and the y-axis is the corresponding partial effect of the covariate. The confident interval is represented by the grey area. **(c)** 2D graphics for the partial contribution of the interplay between gene density and TAD recurrence for the interaction with the three functional classes of TAD segments. The scale from light yellow to red represents partial contribution for higher to lower values of the distribution. The full set of graphics for all cohorts is shown in Supplementary Fig. 3

Figure 4

a Examples of cancer-specific BPpc

b Empirical and theoretical density Q-Q plot

c

Figure 4: **(a)** Examples of cancer-specific *BPpc* and the peaks of significantly recurrent rearrangements for representative cancer types. Each peak shows a dot colored in the scale of significance for the corresponding peak recurrence score (PRs). The peaks that come mostly from a unique sample (likely a chromothripsis event) are marked with an asterisk. The known driver candidates detected within the significantly rearranged peaks are marked (green for coding sequence and red for enhancers). **(b)** Empirical and theoretical density of PRs for each cohort and the corresponding QQ-plots which show the p-values follow a uniform distribution. **(c)** Scatter plots of the genome length of the peak region vs. number of SVs for each peak, the size of the dots shows the recurrence in the cohort. Histogram of the genome length distribution of the significant peaks.

Figure 5

Figure 5: Driver candidates predicted for each cancer cohort. In green are the genes affected in the coding sequence and in red are the genes impacted at their enhancers. Genes marked with a black dot are the ones previously reported in the corresponding cancer type. Genes marked with a white dot are the ones previously reported as cancer genes but in a different cancer type than our prediction. The IGH@ locus contains 4 enhancers. The entire list of candidates is shown in Supplementary Table 5.

Figure 6

Figure 6: Zoomed in plots of the genome location of the significantly rearranged peaks that show the breakpoint clustering at regions of novel candidates with previous reports in other cancer types. The different types of SV are color coded as (DEL black, DUP blue, h2hINV purple, t2tINV red, TRA orange). **(a)** *DMD* in a peak on chrX in esophagus cancer. **(b)** *LRRN3* in a peak on chr7 in esophagus cancer. **(c)** *NF1* in a peak on chr17 in ovarian cancer. **(d)** *CDK12* in a peak on chr17 in uterine cancer. **(e)** *WWOX* in a peak on chr16 in colorectal cancer. **(f)** *CTDSP2* in a peak on chr12 in brain cancer. **(g)** *MACROD2* in a peak on chr20 in pancreatic cancer. **(h)** *CDKN2B* in a peak on chr9 in liver cancer. The boxplots show the differential expression between samples with and without SVs in the peak region. GBM, glioblastoma multiforme; PACA, pancreatic cancer; LIRI, liver cancer. In green are the genes affected in the coding sequence. In red are the genes impacted in their enhancer regulatory region.