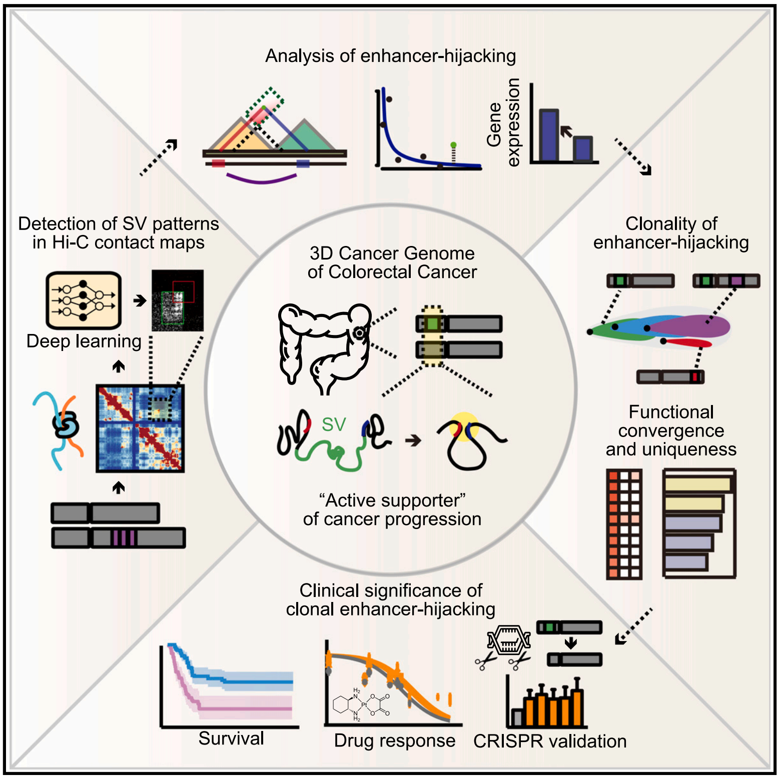
Cell Reports

Article

Spatial and clonality-resolved 3D cancer genome alterations reveal enhancer-hijacking as a potential prognostic marker for colorectal cancer

Graphical abstract


Analysis of enhancer-hijacking
Gene expression
Detection of SV patterns in Hi-C contact maps
Deep learning
3D Cancer Genome of Colorectal Cancer
SV
"Active supporter" of cancer progression
Clonality of enhancer-hijacking
Functional convergence and uniqueness
Clinical significance of clonal enhancer-hijacking
Survival
Drug response
CRISPR validation

Authors

Kyukwang Kim, Mooyoung Kim,
Andrew J. Lee, ..., Young-Joon Kim,
Tae-You Kim, Inkyung Jung

Correspondence

kimty@snu.ac.kr (T.-Y.K.),
ijung@kaist.ac.kr (I.J.)

In brief

Kim et al. delve into the effect of structural variations on the 3D cancer genome and how they activate proto-oncogenes in colorectal cancer. Their findings highlight that clonal enhancer-hijacking genes are functionally convergent and could potentially serve as reliable prognostic markers, leading to more accurate interpretation of patient-specific cancer genomes.

Highlights

- Cancer genome undergoes frequent and genome-wide 3D genome disorganization
- Disorganized 3D genome rewires enhancer-promoter interactions and activates oncogenes
- Clonal enhancer-hijacking genes are recurrently enriched to oncogenic functions
- Clonal enhancer-hijacking genes may serve as prognostic markers for colorectal cancer

Kim et al., 2023, Cell Reports *42*, 112778
July 25, 2023 © 2023 The Author(s).
https://doi.org/10.1016/j.celrep.2023.112778

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

Spatial and clonality-resolved 3D cancer genome alterations reveal enhancer-hijacking as a potential prognostic marker for colorectal cancer

Kyukwang Kim,[1,9] Mooyoung Kim,[1,9] Andrew J. Lee,[1,9] Sang-Hyun Song,[2,3,9] Jun-Kyu Kang,[2,3] Junghyun Eom,[1] Gyeong Hoon Kang,[4] Jeong Mo Bae,[4] Sunwoo Min,[1] Yeonsoo Kim,[2,3] Yoojoo Lim,[2,5] Han Sang Kim,[6] Young-Joon Kim,[7] Tae-You Kim,[2,3,5,8,*] and Inkyung Jung[1,10,*]

[1]Department of Biological Sciences, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Korea
[2]Cancer Genomics Research Laboratory, Cancer Research Institute, Seoul National University, Seoul 03080, Korea
[3]Department of Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, Seoul National University, Seoul 03080, Korea
[4]Department of Pathology, Seoul National University Hospital, Seoul 03080, Korea
[5]Department of Internal Medicine, Seoul National University Hospital, Seoul 03080, Korea
[6]Yonsei Cancer Center, Division of Medical Oncology, Department of Internal Medicine, Graduate School of Medical Science, Brain Korea 21 Project, Severance Biomedical Science Institute, Yonsei University College of Medicine, Seoul 03722, Korea
[7]Department of Biochemistry, College of Life Science and Biotechnology, Yonsei University, Seoul 03722, Korea
[8]IMBdx, Inc., Seoul 08506, Korea
[9]These authors contributed equally
[10]Lead contact
*Correspondence: kimty@snu.ac.kr (T.-Y.K.), ijung@kaist.ac.kr (I.J.)

https://doi.org/10.1016/j.celrep.2023.112778

SUMMARY

The regulatory effect of non-coding large-scale structural variations (SVs) on proto-oncogene activation remains unclear. This study investigated SV-mediated gene dysregulation by profiling 3D cancer genome maps from 40 patients with colorectal cancer (CRC). We developed a machine learning-based method for spatial characterization of the altered 3D cancer genome. This revealed a frequent establishment of "*de novo* chromatin contacts" that can span multiple topologically associating domains (TADs) in addition to the canonical TAD fusion/shuffle model. Using this information, we precisely identified super-enhancer (SE)-hijacking and its clonal characteristics. Clonal SE-hijacking genes, such as *TOP2B*, are recurrently associated with cell-cycle/DNA-processing functions, which can potentially be used as CRC prognostic markers. Oncogene activation and increased drug resistance due to SE-hijacking were validated by reconstructing the patient's SV using CRISPR-Cas9. Collectively, the spatial and clonality-resolved analysis of the 3D cancer genome reveals regulatory principles of large-scale SVs in oncogene activation and their clinical implications.

INTRODUCTION

The genome is organized in a hierarchical 3D chromatin structure, in which distant regulatory elements are often juxtaposed in the nuclear space to control target gene expression. In the classical model of multi-layered genome organization, active and inactive chromatin regions are spatially segregated into compartment A/B and further partitioned into topologically associating domains (TADs) or loop domains, which can span several megabases.[1–3] The structural formation of TADs or loops is associated with gene regulation, as gene expression is often controlled by distal *cis*-regulatory elements located within the same TAD or loop domain.[4–6]

Multiple studies have highlighted the pathogenic disruption of TADs or loop domains, in which germline and somatic structural variations (SVs) rewire spatial contacts between *cis*-regulatory elements and distal target promoters, leading to aberrant gene expression.[7] Large-scale germline deletions and duplications proximal to *cis*-regulatory elements that cause gene dysregulation are associated with congenital disorders.[8,9] In the cancer genome, which frequently carries large-scale somatic SVs, the altered 3D genome structure has been previously suggested as a mechanism for oncogene activation (known as "enhancer-hijacking"). For instance, SV-mediated juxtaposition of *GFI1*/*GFI1B* in medulloblastoma,[10] the micro-deletion of the loop boundary containing *LMO2* in T cell acute lymphoblastic leukemia (T-ALL),[11] and the duplication-mediated *de novo* domain formation comprising *IGF2* and a super enhancer (SE)[12] in the pancancer analysis have been reported. Additionally, recent studies showed that recurrent SV-mediated enhancer-hijacking causes upregulation of *FLT3* in ALL.[13] Also, a systematic investigation of large-scale SVs in various cancer cell lines has further revealed the functional consequences of SVs in cancer.[14] Examination of Hi-C (high-throughput chromosome conformation capture) contact maps of patients with T-ALL identified recurrent TAD fusion containing dysregulated *MYC*.[15]

However, inconsistencies in defining the exact role of 3D genome alterations in oncogene activation were also reported. Specifically, the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium concluded that 3D genome alterations have a marginal effect on gene regulation and that only the fusion between the TADs in repressed and active compartments has a significant impact on gene dysregulation.[16] Furthermore, substantial 3D genome compartmental reorganization alters the transcription program to be tumor suppressive rather than oncogenic.[17]

We hypothesized that such inconsistencies were because of the hindered comprehensive understanding of the 3D cancer genome caused by the following limitations: (1) large patient cohorts from which primary tumor 3D genome data could be obtained are a highly scarce resource, (2) the scope of previous systematic investigations was limited to simple TAD boundary disruptions,[16] and (3) the genetic heterogeneity of cancer cells in the tumor was not taken into consideration.

To overcome such limitations, we comprehensively investigated 3D cancer genome alterations by conducting *in situ* Hi-C experiments on 40 colorectal cancer (CRC) patient tumor tissues and 10 adjacent normal tissues. A new machine learning-based method was developed to accurately identify SV-mediated cancer genome-specific aberrant chromatin contacts (termed "*de novo* chromatin contacts"). Using this resource, we demonstrated genome-wide rewiring of SE and promoter interactions and the underlying principles of SE-hijacking caused by the rearranged 3D genome. Our findings elucidated the functional and clinical implications of SE-hijacking genes and re-highlighted the oncogenic effect of large-scale SVs.

RESULTS

Identifying 3D genome alterations using a new machine learning-based method

We collected tumor tissues (n = 40) and matched adjacent normal tissues (n = 40) from patients with late-stage CRC (labeled as P1–P40) along with detailed clinical information (Figure 1A; Table S1; see STAR Methods). These samples were subjected to whole-genome sequencing (WGS) and total RNA sequencing (RNA-seq; see STAR Methods). We ensured the quality of the collected specimens and their biological relevance to late-stage CRC based on the examination of archived tissue slides (Figure S1; Table S1), somatic variant profiling of known driver genes (Figure S2A), computational tumor purity (Figure S2B), segregation of transcriptome profiles between tumor and adjacent normal tissues (Figure S2C), distribution of consensus molecular subtypes (CMSs) consistent with published data[18] (Figures S2D and S2E), and enrichment of well-known oncogenes in differentially expressed genes (DEGs) (Table S2; see STAR Methods).

Next, we conducted *in situ* Hi-C experiments with 40 CRC tumors and 10 adjacent normal tissues, yielding a total of 10.3 billion long-range chromatin interactions (Table S1; see STAR Methods). The number of long-range chromatin interactions collected in this study was sufficient to profile compartment A/B and TADs. A pan-normal Hi-C contact map to be used as a control was generated based on the observation that Hi-C contact maps are highly reproducible across the adjacent normal tissues (Figures S2F and S2G; see STAR Methods). Examination of tumor Hi-C contact maps revealed aberrant long-range chromatin interactions with a distance exceeding several megabases (*de novo* chromatin contacts), which were highly coupled with the breakpoints of WGS-identified large-scale SVs (Figures 1B and 1C; Table S3; see STAR Methods).

To systematically identify such *de novo* chromatin contacts, we developed a new machine learning method named "chromatin anomaly pattern recognition and size estimation" (CAPReSE), comprising a convolutional neural network (CNN)-based feature extractor combined with an XGBoost classifier (Figure 1D; see STAR Methods). CAPReSE utilizes a unique chromatin contact signature of SVs that shows enriched contact frequencies at the break ends of SVs and a gradual decrease in contact frequencies along the rearranged genomic regions[14,19] (Figure S3A). The input tumor Hi-C contact map was normalized against a pan-normal Hi-C contact map (Figure S3B), which leaves abnormally strong long-range or inter-chromosomal contact signals originating from the large-scale genomic rearrangements of each sample (Figure S3C). Then, a series of image processing algorithms were applied to identify the SVs' unique chromatin contact signatures (see STAR Methods). The SVs supported by both WGS and Hi-C data were used as a ground-truth set for the final classifier (Figure S3D; see STAR Methods). As a result, CAPReSE achieved around 90% test accuracy (F1 score) in 2-fold cross validations (Figure S3E). The performance of CAPReSE achieved a low false positive rate in a benchmark test, outperforming conventional software[14] (Figures S3F and S3G; see STAR Methods). Also, robustness in performance (∼90% recall and ∼99% precision) regardless of tumor purity (Figure S3H; see STAR Methods) was confirmed. By applying CAPReSE to 40 CRC patient Hi-C contact maps, we identified a total of 562 *cis*- and 235 *trans-de novo* chromatin contacts (Figure 1E; Table S3; see STAR Methods). Patients classified as CMS2/4 (Figure 1E) or with accumulated *TP53* mutations (Figure S3I) tended to show more *de novo* chromatin contacts. We also observed massive and complex forms of disorganized 3D genomes in 35% of the patients, most likely associated with chromothripsis or chromoplexy (Table S3; see STAR Methods).

Fine mapping of enhancer-hijacking based on 3D genome alterations

Notably, genes associated with *de novo* chromatin contacts tended to be upregulated compared with the matched adjacent normal tissues, which cannot be simply explained by the copy-number gains (Figure 2A). The majority of the upregulated genes (∼77%) in the *de novo* chromatin contacts did not show copy-number gains (Figure S4A; Table S4; see STAR Methods). In consideration of previous reports and models about enhancer-hijacking (SV-driven juxtaposition of enhancers to originally distal promoter),[11,20] we analyzed the regulatory effect of 3D genome alterations to further explain the activation of these putative oncogenes. To this end, we simplified our analysis by focusing on 1,888 SEs defined by H3K27ac chromatin

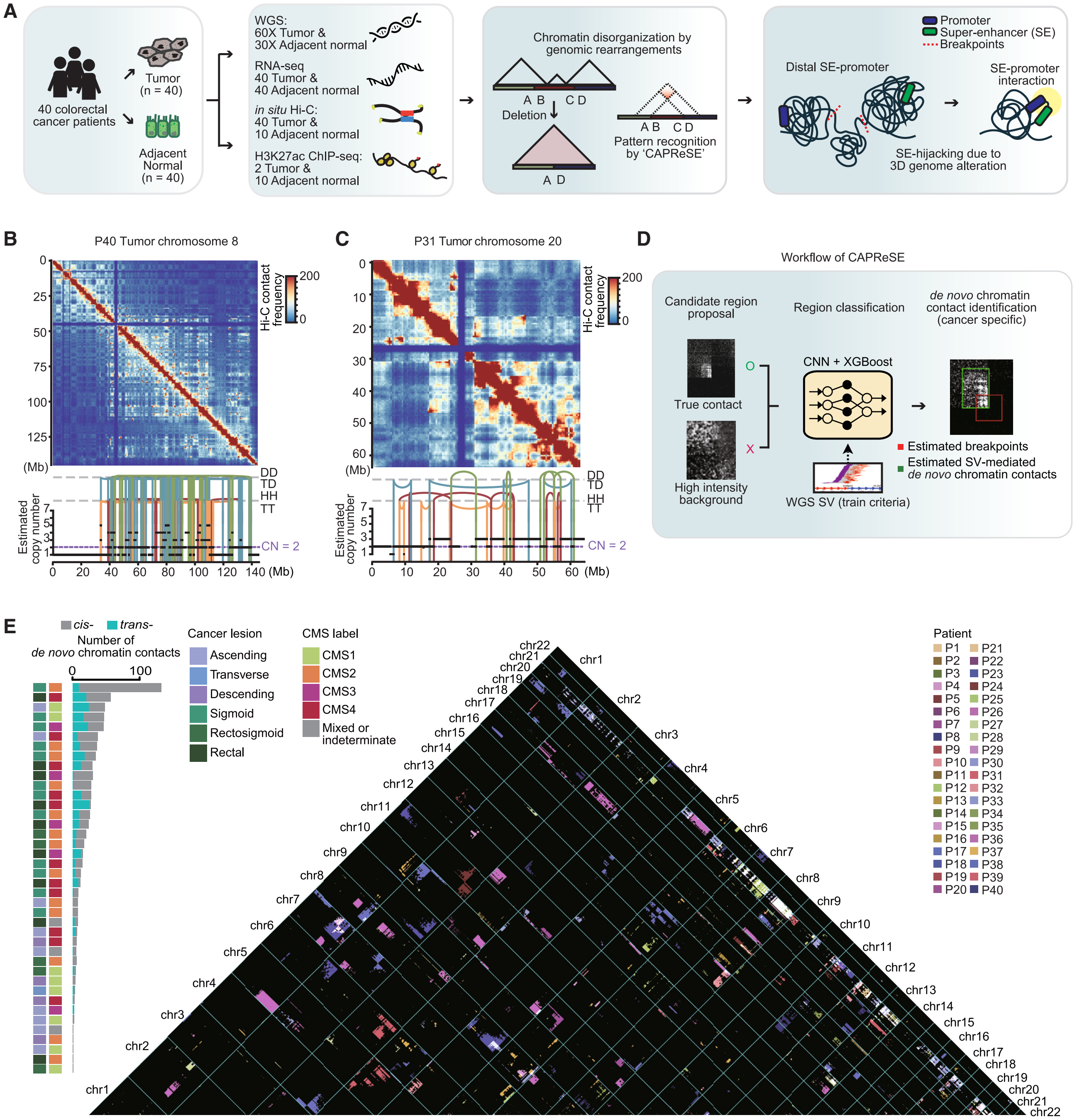


Figure 1. Identification of widespread *de novo* chromatin contacts in colorectal cancer

(A) A schematic of the study design illustrating the generation of sequencing-based omics data derived from 40 patients with CRC and downstream computational analysis to characterize the effects of large-scale genomic rearrangement.

(B and C) Tumor Hi-C contact maps (500 kb resolution) and WGS-based SVs of (B) P40 chromosome 8 and (C) P31 chromosome 20, respectively. Types of somatic SVs include deletion (DD), tandem duplication (TD), head-to-head inversion (HH), and tail-to-tail inversion (TT) with copy numbers (black dots) are shown in the graphs below the contact maps. These vertical lines are not expected in normal tissue. Purple dashed lines indicate a copy number of 2 (CN = 2, expected in normal tissue).

(D) An overview of CAPReSE, the machine learning-based method for systematically identifying large-scale genomic rearrangements on Hi-C contact maps.

(E) Concatenated Hi-C contact maps of chromosomes 1–22 from 40 patients with CRC. The colored areas indicate regions where each patient's contact signal increases more than 3-fold compared with the pan-normal. The stacked bar plot on the left shows the number of CAPReSE-detected *de novo* chromatin contacts. Intra- (*cis*-, gray) and inter- (*trans*-, cyan) chromosomal patterns are represented by the bar color, along with the location of the cancer lesion and the CMS for each patient.

See also Figure S3 and Table S3.

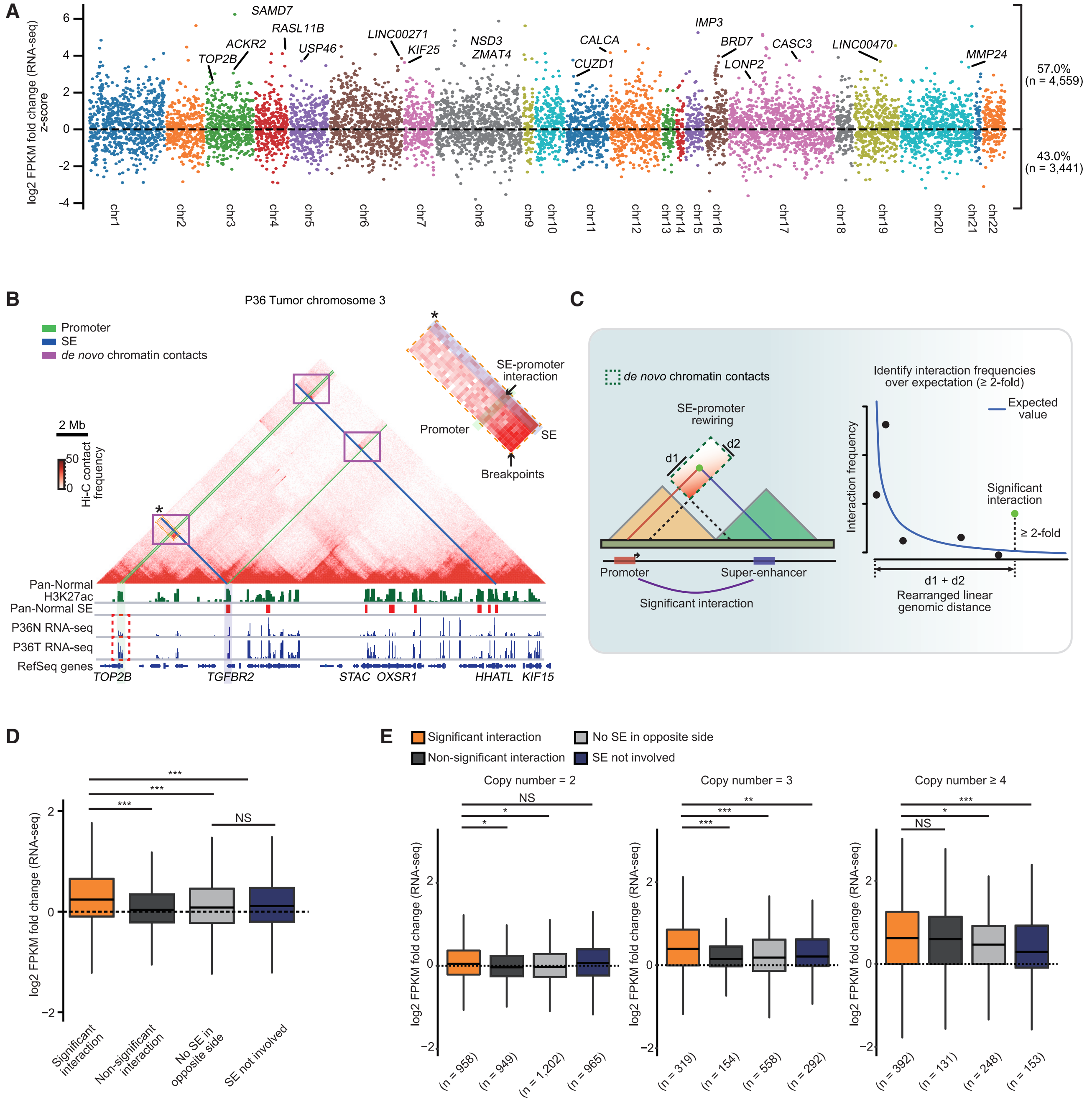


Figure 2. Regulatory effects of 3D genome alterations on gene expression

(A) A scatterplot showing the expression levels (*Z* score of log2 fold changes compared with the matched adjacent normal tissue) of genes located in the identified *de novo* chromatin contacts. Upregulated genes associated with CRC development and progression are annotated.

(B) A tumor Hi-C contact map showing *de novo* chromatin contacts in chromosome 3 of P36, along with genome browser tracks of pan-normal H3K27ac ChIP-seq, super enhancers (SEs), and P36 RNA-seq results. The representative genomic rearrangement (orange dashed line box, marked with an asterisk) shows increased chromatin interactions of a rewired pair of the *TOP2B* promoter (highlighted in solid/translucent green bar) and the SE (highlighted in solid/translucent blue bar).

(C) A schematic illustration of the strategy used to measure the interaction significance of rewired SE-promoter pairs. The light green dot represents a significant SE-promoter interaction.

(D) A boxplot showing the effects of newly established SE-promoter interactions on gene expression (dark orange, n = 1,695, log2 fold changes compared with the matched adjacent normal tissue). The dark gray box indicates the SE-promoter pairs without significant interactions (n = 1,311). The light gray box represents promoters involved in the *de novo* chromatin contacts but not associated with relocated SEs (n = 2,095). The dark blue box indicates promoters in the *de novo*

(legend continued on next page)

immunoprecipitation (ChIP)-seq results (Table S5; see STAR Methods), since SEs are well conserved across adjacent normal and tumor tissues, present a far higher regulatory activity than other typical enhancers, and are enriched in *de novo* chromatin contacts (Figures S4B–S4D). To fine map SE-hijacking events, Hi-C contact maps of individual patients' primary tumor tissues were used, unlike a previous study,[16] with the following criteria: (1) SE-promoter pair was originally located at distal loci, (2) the pair is rewired by the *de novo* chromatin contacts, and (3) the interaction strength of the pair is stronger than the expectation (≥2-fold) based on the rearranged genomic distance (Figures 2B and 2C; see STAR Methods).

As a result, a total of 1,695 genes were targeted by relocated SEs with significant interactions (SE-hijacking genes). These genes were significantly upregulated in tumors compared with the control gene sets (SE-promoter pairs with non-significant interactions, promoters without newly introduced SEs, or promoters residing in the *de novo* chromatin contacts without SE involvement) (Figures 2D and S4E). Additionally, we excluded the effect of copy-number gains by confirming the upregulated expression of SE-hijacking genes that correspond to each copy-number variation (2, 3, and ≥4) (Figure 2E). Thus, we can conclude that the spatial interactions between SE-promoter pairs make meaningful changes in gene expression. Altogether, our results demonstrated the regulatory effect of SE-hijacking in CRC and emphasized the importance of precise interpretation of each patient's 3D cancer genome to predict oncogenic SVs.

Prevalent SE-hijacking mediated by multi-TAD spanning *de novo* chromatin contacts

In the canonical model, enhancer-hijacking is mediated by SV-linked fusion or shuffle between two TADs.[20] However, intriguingly, we observed that a large proportion of *de novo* chromatin contacts span multiple TADs: the contacts are not strictly insulated by the boundaries of "adjacent TADs" (two fused or shuffled TADs, where breakpoints are located within TADs) (n = 681, 45.49%) (Figures 3A and S5A–S5C). These multi-TAD-spanning *de novo* chromatin contacts often produce many SE-hijacking candidates, as exemplified in the overexpression case of proto-oncogene *CCDC117*, which is associated with DNA repair and cell-cycle progression (Figure 3B). Newly established chromatin contacts were observed between a SE and the promoter of *CCDC117* located outside of the adjacent TADs (Figure 3B). We found that both canonical and multi-TAD-crossing (extended model) SE-hijacking models support the upregulation of SE-hijacking genes well (Figure 3C).

To further understand the principle of large-scale SVs generating multi-TAD-spanning *de novo* chromatin contacts, we investigated multiple characteristics of the genomic regions participating in chromatin reorganization. First, we found that multi-TAD-spanning *de novo* chromatin contacts are significantly associated with weak TAD boundary insulation scores (Figure 3D; see STAR Methods) and short boundary lengths (Figure S5D). Additionally, SVs located in active compartments (compartment A) often induced the TAD-boundary-crossing 3D genome reconfiguration (Figures 3E and 3F). These results propose that weakly stratified spatial chromatin domains within active regions are more vulnerable to the effect of SVs on the 3D genome. Our results demonstrated that the regulatory effects of large-scale SVs can expand much larger than anticipated by the previous simple TAD fusion model.[11,16]

Resolving the clonality of *de novo* SE-promoter interactions

Clonal evolution is a major hallmark in cancer progression, where cancer subclones with genetic diversity emerge at various time points and experience different selection pressures. Although previous studies have examined how the enhancer-promoter pairs are juxtaposed by SVs, the "clonality" of such events in primary tumors has never been investigated due to the lack of a proper algorithm. We addressed this challenge by leveraging somatic variants found in Hi-C reads. Since *de novo* chromatin contacts mediated by SVs are mostly cancer specific, the clonality of somatic variants within Hi-C reads indicates the clonality of the corresponding *de novo* chromatin contacts. To this end, we first measured the mutation timing (clonality) of each WGS somatic variant[21] (see STAR Methods). The clonality estimation was verified by confirming the clonal enrichment of the major CRC driver mutations (*TP53*, *APC*, *KRAS*, *PIK3CA*, and *SMAD4*) (Figure S5E). After that, by assessing the existence of clonality-defined variants within the Hi-C reads that constitute the *de novo* chromatin contacts, we categorized 891 upregulated SE-hijacking events (fold change > 0, compared with the matched adjacent normal tissue) into clonal, subclonal, undetermined, and non-testable (Figure 4A; see STAR Methods). Approximately 46.4% of the SE-hijacking events were determined to be clonal, which is likely to occur at a relatively early stage shared by a major cancer cell population in a tumor tissue (Figure 4B; Table S6). As expected, the genes associated with the clonal SE-hijacking also showed a higher level of gene expression changes compared with those associated with subclonal SE-hijacking (Figure 4C).

Functional recurrence of clonal SE-hijacking genes

The regulatory role of large-scale SVs in cancer progression is not well understood due to their sporadic location and various types of rearrangements. Although several SVs and SE-hijacking incidents were recurrently observed in this study, there was little commonality at the gene level among patients. In this regard, we

chromatin contacts without SEs (n = 1,469). Statistical differences between the two groups were calculated by performing two-sided Kolmogorov-Smirnov (KS) tests (***p < 0.001, and NS, not significant).

(E) Boxplots showing the effects of newly established SE-promoter interactions on gene expression (log2 fold changes compared with the matched adjacent normal tissue) based on CNs. Box colors indicate SE-promoter combination cases. Genes with fragmented CNs were not included. Statistical differences between the two groups were calculated by performing two-sided KS tests (*p < 0.05, ***p < 0.001, and NS, not significant). For boxplots, the box represents the interquartile range (IQR), and the whiskers correspond to the highest and lowest points within 1.5× IQR.

See also Figure S4 and Tables S4 and S5.

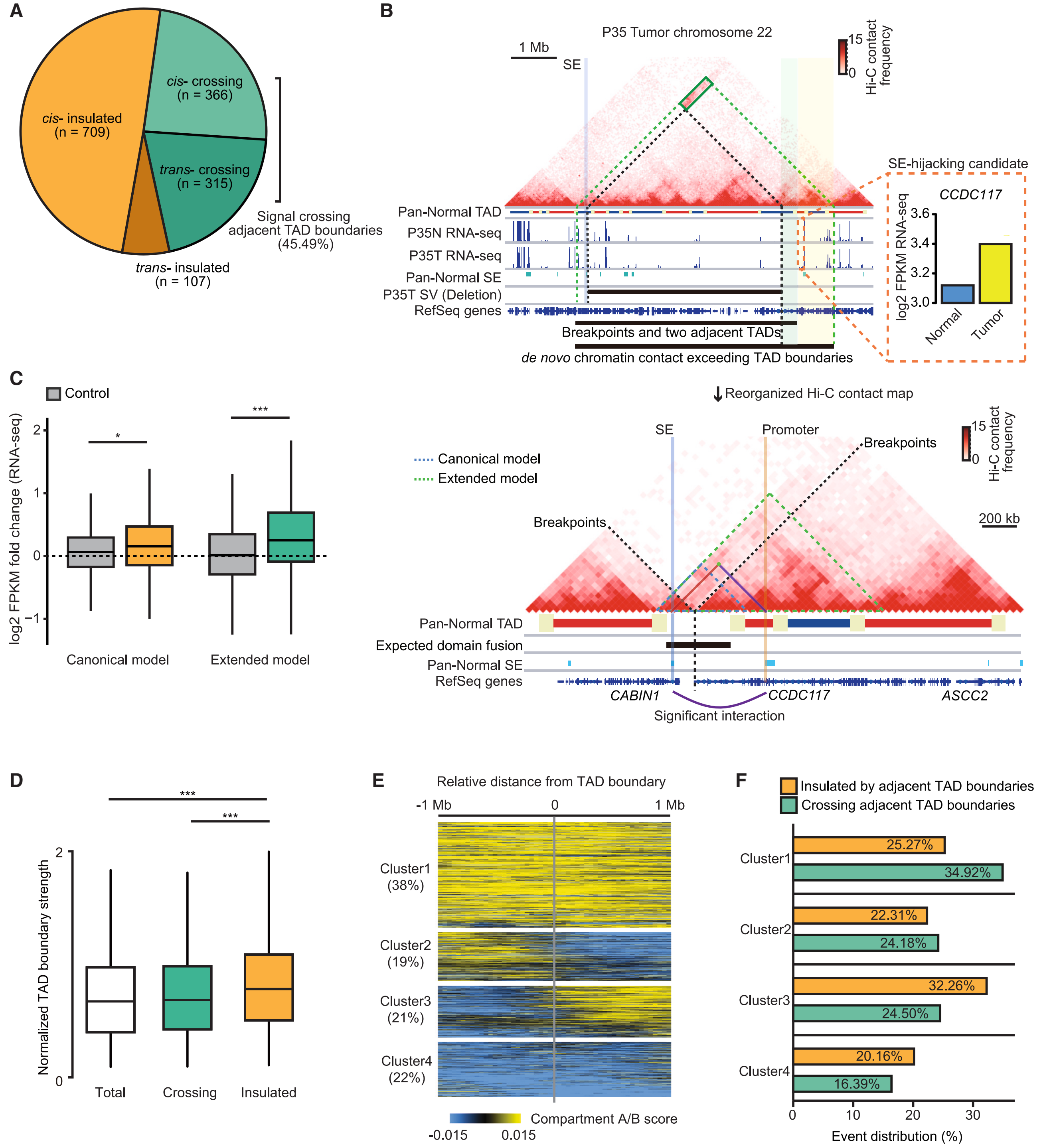


Figure 3. SE-hijacking candidates through multi-TAD-spanning *de novo* chromatin contacts

(A) A pie chart showing the proportion of *de novo* chromatin contacts either insulated by "adjacent TAD" boundaries or crossing adjacent TAD boundaries.

(B) Top panel: a normalized Hi-C contact map of P35's tumor showing a representative example of adjacent TAD-boundary-exceeding SE-hijacking at chromosome 22 (chr22):22,419–30,268 kb (40 kb resolution) as well as genome browser tracks of pan-normal TAD, RNA-seq, pan-normal SEs, and WGS-based SVs. Bottom panel: a reorganized Hi-C contact map showing multi-TAD-spanning *de novo* chromatin contacts. Fused domains are marked with dashed line triangles (canonical model: blue, extended model: green). Translucent blue and orange marks show SE and *CCDC117* promoter locations, respectively. The SE-promoter rewired interaction is marked with red and blue solid lines and light green dot.

(legend continued on next page)

hypothesized that the functions of SE-hijacking genes can be shared among patients, considering that the SE-hijacking events are clonal and favorable to cancer progression.

To examine the hypothesis about the functional convergence of SE-hijacking genes, we performed a gene set enrichment analysis with the 891 upregulated SE-hijacking genes using Metascape.[22] Notably, these SE-hijacking genes showed functional enrichment in cell cycle and DNA processing terms, seemingly associated with biological functions that contribute to cancer progression (Figure 4D). Multiple representative markers for tumorigenesis and metastasis (e.g., SP1, E2F1, and MYC) were also identified as upstream regulators of SE-hijacking genes, which further support the biological significance of SE-hijacking on tumorigenesis (Figure 4D). We also performed the same analysis with high-copy-number-gain genes ($\geq$3 copies and recurrently found in $\geq$16 patients, n = 1,011) and frequently mutated genes (mutated in $\geq$3 patients, n = 466) of 40 patients with CRC. Interestingly, all three gene sets showed a quite unique gene composition (Figure S5F) and functional enrichment characteristics (Figure 4D), indicating that genes activated by SE-hijacking could exert a unique contribution to tumorigenesis.

We further questioned whether the functional convergence of SE-hijacking genes depends on the clonality and applied the gene set enrichment analysis according to the clonality types (clonal, subclonal, or undetermined). The results showed that most of the cancer-related functions of SE-hijacking genes were reproduced with the clonal SE-hijacking genes, while subclonal or undetermined SE-hijacking genes exhibited weak or no enrichment (Figure 4E). Additionally, the enrichment analysis conducted with the PaGenBase (Pattern Gene Database) showed that only clonal SE-hijacking genes were enriched with a ‘‘colorectal adenocarcinoma’’ term (Figure 4F). The functional association of clonal SE-hijacking genes in tumorigenesis supports our hypothesis that early generated SE-hijacking events are highly associated with CRC progression and may have a positive effect on survival during clonal selection.

The prognostic potential of clonal SE-hijacking genes

Given the functional significance of SE-hijacking events, we further examined whether SE-hijacking genes can act as a CRC prognostic marker. As it is difficult to estimate the effect of multiple gene expressions on the CRC prognosis, a RISK score metric was used[23] (see STAR Methods). Using all valid protein-coding genes as candidates, 2-fold cross validations were conducted 100 times with the Center for Integrative Omics and Precision Medicine (CoPM) consortium dataset to define the gene set that is recurrently related to the CRC prognosis (see STAR Methods).

In a total of 100 repeated tests, the patients with a high RISK score (RISK score $\geq$ average RISK score) showed a significant decrease in survival time (progression-free survival [PFS] of CoPM dataset) compared with the patients with a low RISK score (Figure 5A). We found that 23% of SE-hijacking genes overlap with the ‘‘frequently included genes’’ (significant in $\geq$50 times of cross validations; see STAR Methods) (Figure 5B), which is a higher proportion compared with the control gene sets (Figure 5C). Among the SE-hijacking genes found in the frequently included genes, 82 genes were classified as clonal (defined as ‘‘clonal RISK genes’’) and 11 genes were subclonal (3 overlapping genes) (pie chart in Figure 5C).

Next, a hazard ratio (HR) analysis was used to compare the effects of clonal RISK genes on the CRC prognosis with other factors. For comparison, 4 mutation features (*APC*, *KRAS*, *TP53*, and microsatellite instability status), 5 types of gene set markers,[24] and age (control) data were used (see STAR Methods). In the HR analysis, the RISK score of clonal RISK genes showed a HR of 3.6, which was the highest value among the 9 (two non-significant factors filtered; see STAR Methods) factors (Figure 5D). This is a higher value than a well-known marker such as *KRAS* mutation or previously published marker gene sets. The prognostic ability of clonal RISK genes was further verified by the HR analysis using The Cancer Genome Atlas (TCGA) dataset (see STAR Methods). Again, the clonal RISK genes showed the highest HR of 1.8 in the TCGA dataset, indicating that clonal RISK genes can work as a robust marker (Figure S5G). Overall, our results highlighted the potential clinical practicability of SE-hijacking genes as CRC prognostic markers.

Validation of *TOP2B* clonal SE-hijacking case and its functional consequence

Gene upregulation due to 3D genome alterations and their potential as CRC prognostic markers were discussed in the previous sections. For validation, the CRISPR-Cas9 system was used on the HCT116 CRC cell line to mimic a SE-hijacking event observed in a patient. We selected *TOP2B*, which controls DNA integrity and cell cycle, as a validation target because of its relevant function on cancer progression,[25] the clonal nature of this SE-hijacking event, and its upregulated expression in the primary tumor.

(C) A boxplot illustrating the gene expression (log2 fold changes compared with the matched adjacent normal tissue) characteristics of SE-hijacking candidates based on their adjacent TAD-boundary-exceeding property compared with the controls (gray, expression of the other 39 samples). Yellow: both SE and promoters are located within the adjacent TADs (canonical enhancer-hijacking model, n = 413). Cyan: either SEs or promoters located outside of the adjacent TAD boundaries (extended enhancer-hijacking model, n = 1,242). Statistical differences between the two groups were calculated by performing two-sided KS tests (*p < 0.05 and ***p < 0.001).

(D) Boxplots showing the normalized TAD boundary strength compared with background (average of each chromosome) for all TAD boundaries of 40 patients with CRC (white, n = 103,720), crossing TAD boundaries (cyan, n = 1,257), and insulated by TAD boundaries (orange, n = 744). Statistical differences between the two groups were calculated by performing two-sided KS tests (***p < 0.001). For boxplots, the box represents the IQR, and the whiskers correspond to the highest and lowest points within 1.5× IQR.

(E) K-means clustering of compartment A/B values around (2 Mb window) TAD boundaries (n = 2,001) reside in *de novo* chromatin contacts regions.

(F) Bar plots showing the proportion of *de novo* chromatin contacts that are insulated by (orange, n = 744) or crossing the adjacent TAD boundaries (cyan, n = 1,257) in each compartment A/B value cluster type.

See also Figure S5.

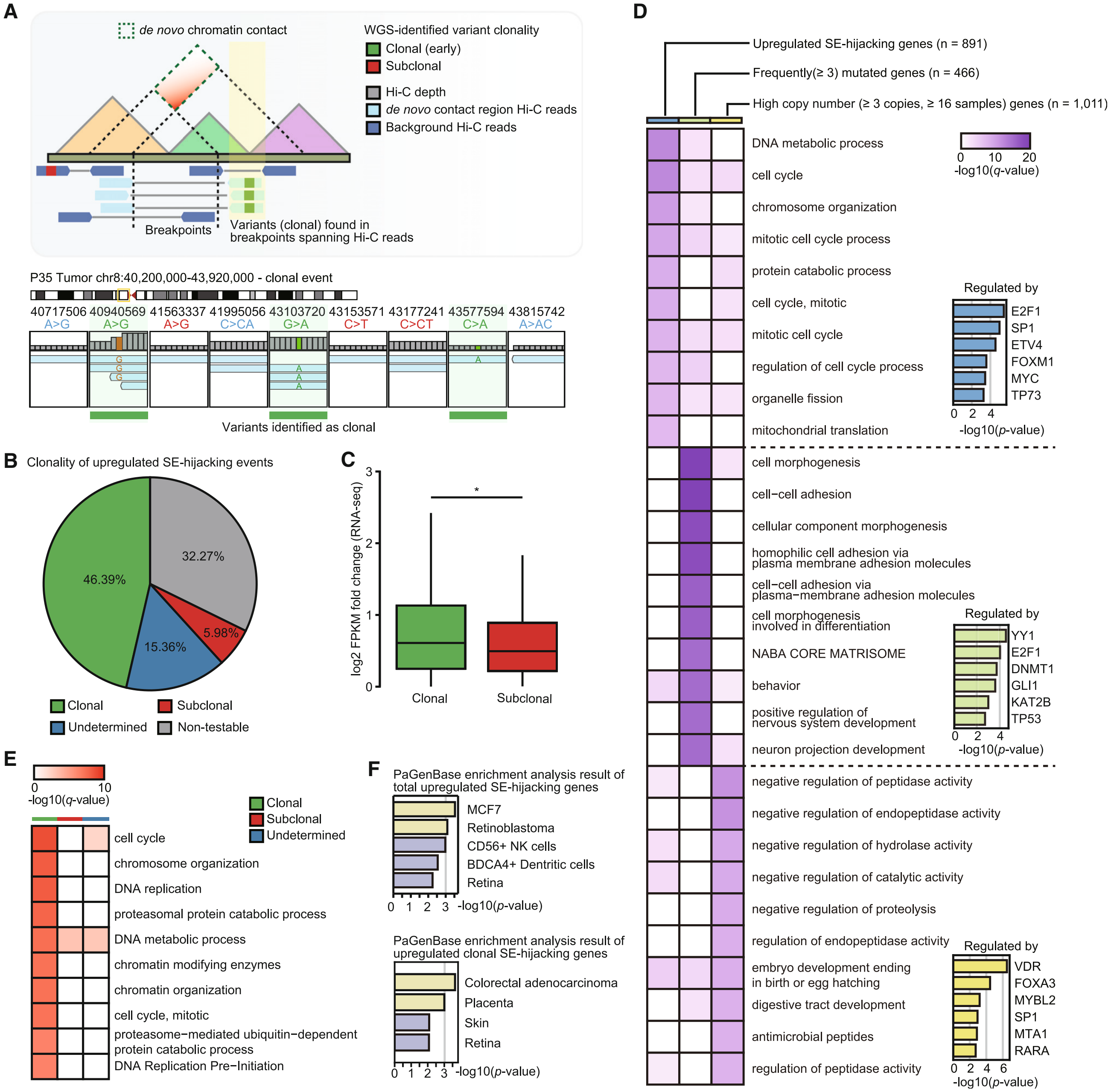


Figure 4. Oncogenic functionality and relative clonality of SE-hijacking genes

(A) A schematic showing how clonality types of the *de novo* chromatin contacts are determined. Hi-C reads harboring a variant with known clonality type (green square: clonal and red square: subclonal) spanning the *de novo* chromatin contacts (light blue reads) are highlighted in translucent yellow (top panel). A genome track example shows *de novo* chromatin contacts found in patient P35 tumor chromosome 8. Translucent green and green bars show mapped variants supporting the clonality type of the Hi-C interactions consisting of the *de novo* chromatin contacts (bottom panel).

(B) A pie chart showing the clonality types of upregulated SE-hijacking events.

(C) A boxplot showing the gene expressions (log2 fold changes compared with the matched adjacent normal tissue) for clonal and subclonal SE-hijacking genes, respectively (one-tailed KS test, *p < 0.05). For boxplots, the box represents the IQR, and the whiskers correspond to the highest and lowest points within 1.5× IQR.

(D) A heatmap showing the enriched biological pathways involved in the upregulated SE-hijacking genes (left column), frequently mutated genes (middle column), and high-CN genes (right column). Bar plots in each gene set sector show the upstream transcription factors regulating the gene set.

(legend continued on next page)

Introduction of large-scale genomic deletion mimicking P36 patient's *TOP2B* gene SE-hijacking event in the HCT116 cell line (D134 clone) reproduced the altered 3D cancer genome observed in the patient well (Figures 6A, 6B, S6A, and S6B; see STAR Methods). After the 3D genome alteration was induced, all mutant clones showed significant overexpression of *TOP2B* compared with the control clones. However, the expression levels of nearby (*THRB*) and distal (*SATB1*) genes (no significant interactions with the relocated SE) did not show consistent changes (Figures 6C and S6C; see STAR Methods). Our genome editing results provide evidence that SV-mediated 3D genome alterations can have a significant impact on proto-oncogene activation. In light of the association between topoisomerases and drug resistance, we also questioned the clinical consequences of *TOP2B* overexpression mediated by SE-hijacking. We examined the viability of HCT116 mutant clones after treatment with oxaliplatin, a platinum-based, DNA-damaging antineoplastic drug used in standard CRC chemotherapy (see STAR Methods).[26] The results of MTT assays to measure the viability of the HCT116 mutant clones with *TOP2B* overexpression showed increased viability compared with the control HCT116 cell line at various oxaliplatin concentrations (Figure 6D). The mutant clones also showed increased viability against irinotecan/SN38 and etoposide, an inhibitor of topoisomerase I and II α (Figures 6D and S6D). In contrast, *TOP2B* knockout mutants (K13 and K14 clones; Figure S6E; see STAR Methods) presented the opposite effect (Figure 6E). These results show that certain SE-hijacking genes, such as *TOP2B*, provide a survival advantage when DNA is topologically stressed by the sustained cell-cycle progression of cancer. Furthermore, our findings demonstrate the clinical significance of SE-hijacking genes and their potential as CRC prognostic markers.

DISCUSSION

Frequent SV is a hallmark of the cancer genome.[27,28] Despite the success in discovering translocations that directly exert an effect on the coding region by forming oncogenic fusion genes,[29,30] the impact of these genomic aberrations on nearby gene regulation is not yet fully understood. In this study, we focused on the SV-mediated 3D genome alterations in primary CRC tumor tissues and investigated how they affect gene dysregulation in cancer by thoroughly examining the newly established chromatin contacts between the promoter and SE.

A few studies have previously questioned the impact of the disorganized 3D genome on gene expression. For example, the results of the pan-cancer genome analysis examining TAD boundary deletion indicate that only 16% of genes are affected,[16] which rebuts the impact of the disorganized 3D genome on gene expression. However, due to the limitations of the data and detection methodology, the analysis was conducted on cell line Hi-C data rather than the patient-originated specimens. Furthermore, the scope of the investigations was limited by handling simple deletion cases, which are insufficient to draw a generalized conclusion and to estimate the overall impact of enhancer-hijacking. Also, even though the presence of significant interactions with relocated regulatory elements is a critical factor, such information was not incorporated. In contrast, herein, we focused on the newly established chromatin interactions mediated by SVs with a large patient cohort. The machine learning-based method was developed for the detection and analysis of the disorganized 3D genome (Figure 1D), which allowed us to find enhancer-hijacking candidates more precisely regardless of SV types or TAD boundary conditions. By combining a large CRC cohort's high-depth Hi-C data with the precise analysis of *de novo* chromatin contacts, we showed that the oncogenic function of SVs can be better understood in the context of the 3D genome of the individual patient than relying on WGS-based analysis only (Figures 2D and 2E).

We propose two new principles underlying enhancer-hijacking: (1) multi-TAD-spanning SV-mediated chromatin contacts (Figure 3) and (2) clonality of enhancer-hijacking (Figure 4). The enhancer-hijacking was conventionally explained by TAD fusion or shuffling that only considers two TADs where breakpoints of the SV reside in these domains.[20] However, our analysis showed that the formation of *de novo* chromatin contacts can further extend beyond adjacent TADs rather than being strictly limited to the local 3D genome organization. Thus, we can expand the canonical TAD fusion/shuffle model to the multi-TAD fusion/shuffle model. Also, we devised a new approach to determine the clonality of enhancer-hijacking through combined analysis of mutation timing of WGS variants and the variant containing Hi-C reads. Our approach highlighted that the clonality-resolved cancer-specific 3D genome provides new insight about enhancer-hijacking during tumorigenesis.

Sustained proliferative signaling of cancer leads to clones with mutations favorable for proliferation becoming predominant. However, oncogene mutations usually activate signaling circuits for unlimited cell growth and simultaneously activate failsafe mechanisms such as apoptosis and senescence. Therefore, to overcome these barriers, an inactivating mutation in the surveillance system such as in *TP53* is inevitable, which enables massive somatic mutations and genome instability. SVs possibly generated by defects in the DNA repair system could cause SE-hijacking and lead to clonal selection for cell survival. Remarkably, we found that the majority of SE-hijackings are clonal events that occur relatively earlier in cancer progression than subclonal events. It seems that the clonal selection of cancer cells with upregulated DNA repair genes by SE-hijacking have the advantages of overcoming failsafe mechanisms such as apoptosis and senescence, which leads to tumor progression and chemoresistance. In this study, of the SE-hijacking genes

(E) A heatmap showing the enriched biological pathways associated with clonal (left column, green marker), subclonal (middle column, red marker), and undetermined (right column, blue marker) gene sets.

(F) Bar plots showing the results of PaGenBase enrichment analysis for total upregulated SE-hijacking genes (top panel) and upregulated clonal SE-hijacking genes (bottom panel).

For (B), (C), (E), and (F), the same gene sets (clonal = 422, undetermined = 145, subclonal = 58, and non-testable = 304) were used.

See also Figure S5 and Table S6.

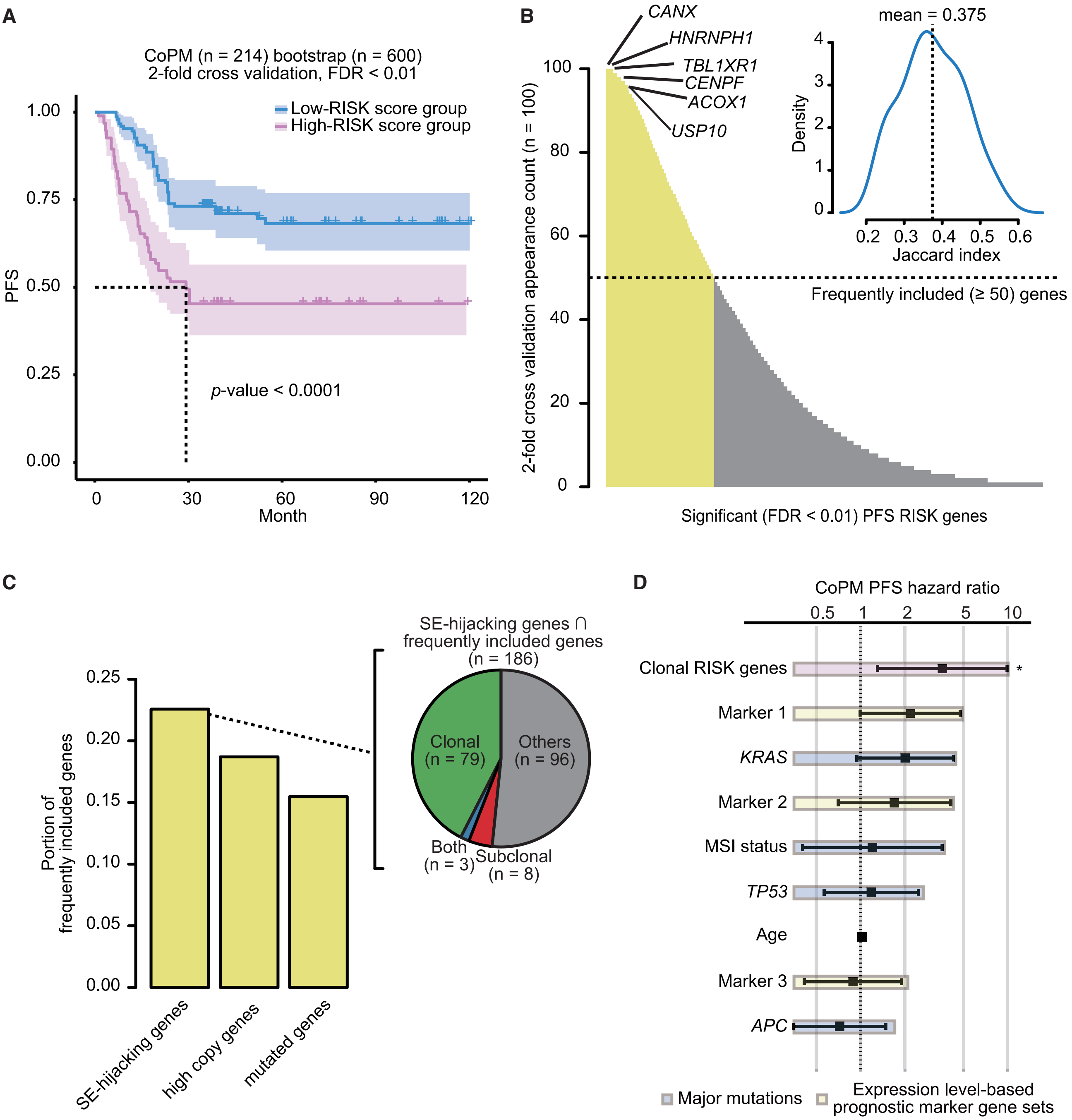


Figure 5. The clinical implications of SE-hijacking genes

(A) An example Kalan-Meier curve showing the survival probability (progression-free survival [PFS]) of high- (purple) and low-RISK-score (blue) patient groups during 2-fold cross validations using the Center for Integrative Omics and Precision Medicine (CoPM) consortium bootstrap dataset.

(B) A bar plot showing the appearance counts of genes during 100 repeats of 2-fold cross validations. Yellow bars indicate that genes appeared more than 50 times in the significant gene set list (false discovery rate [FDR] < 0.01, n = 3,623). Kaplan-Meier (log rank) p value <0.05 was obtained in all tests, and a p value <0.001 was observed in 95 tests. A density plot shows the distribution of the Jaccard index between the frequently included genes and the significant gene set list of each cross validation.

(C) A bar plot showing the proportion of frequently included genes in each tested gene set (testable gene numbers: SE-hijacking genes = 824, high-copy-gain genes = 941, and frequently mutated genes = 446). A pie chart indicates the number of clonal and subclonal SE-hijacking genes in frequently included genes.

(legend continued on next page)

involved in DNA integrity and the cell-cycle checkpoint, *TOP2B* is potentially involved in genome stability in CRC. The outcomes of a recent study suggest that *TOP2B* generates double-strand breaks at loop anchors to resolve topological strains associated with chromosome organization.[25] Thus, an elevated level of *TOP2B* could overcome genome instability by resolving topological strains for unlimited cell growth. In the present study, we validated this hypothesis that an increased level of *TOP2B* via SE-hijacking resulted in higher cell survival after drug treatment, while *TOP2B* knockout increased drug sensitivity. Similarly, although *CHEK1* (another upregulated clonal SE-hijacking gene found in the present study; Table S6) functions in the DNA damage response and activates various downstream effectors to trigger cellular response upon DNA damage and to protect the genome, it is frequently overexpressed in cancer. Moreover, an elevated level of *CHEK1* increases cell transformation and is related to therapy resistance. These examples show that overexpression of DNA repair genes by enhancer-hijacking efficiently contributes to relieving DNA damage stress in tumor environments and facilitates cancer cell survival. As it is well known that genetic mutation or fusion genes are necessary for the development of cancer, it is difficult to call enhancer-hijacking a driver. Instead, we propose that enhancer-hijacking may help cancer progression from a relatively early point by acting as an "active supporter."

Predicting the drug response for each patient is key to effective medical treatment. The diversity of large-scale SVs found across patients with CRC in our data shows that integrating Hi-C into multi-omics data analysis provides a new dimension to explain the complex dysfunctions of cancer. As seen by the dysregulation of *TOP2B*, SE-hijacking caused by the disorganized 3D genome may lead to drug resistance against multiple anticancer drugs. When considering cancer heterogeneity and clonality, the influence of the disorganized 3D genome can be large and diverse. In this aspect, our results emphasize the importance of the disorganized 3D genome in cancer and bridge the gap between genomic rearrangement and gene expression.

Limitations of the study

We systemically identified 3D genome alterations in tumors from 40 patients with CRC using a new SV detection method and characterized the regulatory role of the large-scale SVs. However, one of the challenges in utilizing clinical tumor specimens lies in tissue heterogeneity. In the present study, we resolved this issue, to a large extent, by designing our algorithm to subtract pan-normal signals in the individual 3D cancer genome to specifically isolate cancer-relevant 3D genome alterations. However, the construction of 3D genome maps in single-cell resolution is eventually required to resolve the exact underlying mechanism of oncogene activation and clonal evolution. In this regard, recent developments in single-cell Hi-C technology have shown promising advances,[31–33] and the application of these methods in individual tumor tissues will better portray cancer-specific gene dysregulations caused by large-scale SVs. Also, the CAPReSE pipeline should be further modified for application to broader cases with diverse input formats and sample types such as cancer cell line data.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - ○ Colorectal cancer tissue samples and clinical data
  - ○ Cell culture
- METHOD DETAILS
  - ○ Tumor tissue purity evaluation
  - ○ Immunohistochemistry
  - ○ Whole-genome sequencing
  - ○ Whole-genome sequencing data processing
  - ○ Somatic SNV and INDEL calling
  - ○ Copy number variation calling
  - ○ Large-scale somatic structural variation calling
  - ○ Computational tumor purity measurement
  - ○ RNA sequencing
  - ○ RNA-sequencing data processing
  - ○ Identification of CMS for tumor samples
  - ○ *In situ* Hi-C library preparation
  - ○ Hi-C data processing
  - ○ Topologically associating domain calling and pan-normal Hi-C contact map generation
  - ○ Measurement of insulation score and TAD boundary strength
  - ○ Development of a machine learning-based method to detect SVs from Hi-C contact maps
  - ○ Refinement of the identified breakpoints
  - ○ Performance evaluation of the developed method
  - ○ Tumor purity robustness test of the developed algorithm
  - ○ Identification of complex genomic rearrangements
  - ○ ChIP-seq library preparation
  - ○ ChIP-seq data processing
  - ○ Identification of pan-normal super-enhancers
  - ○ Identification of SE-hijacking candidates
  - ○ Clonality determination of *de novo* chromatin contacts
  - ○ RISK score metric calculation
  - ○ Survival analysis based on RISK score

(D) A forest plot showing PFS hazard ratios of the 82 clonal RISK genes (translucent red), major mutations (translucent blue), and expression level-based CRC prognostic markers (translucent yellow) using the CoPM patient dataset (n = 96). Markers were excluded if all genes showed non-significant p value (>0.05) during Cox regression (markers 4 and 5). Each box represents the median of the hazard ratio, while the bars at both ends represent the minimum and maximum hazard ratio values. The significance of each factor was calculated based on log rank tests (*p < 0.05).
See also Figure S5.

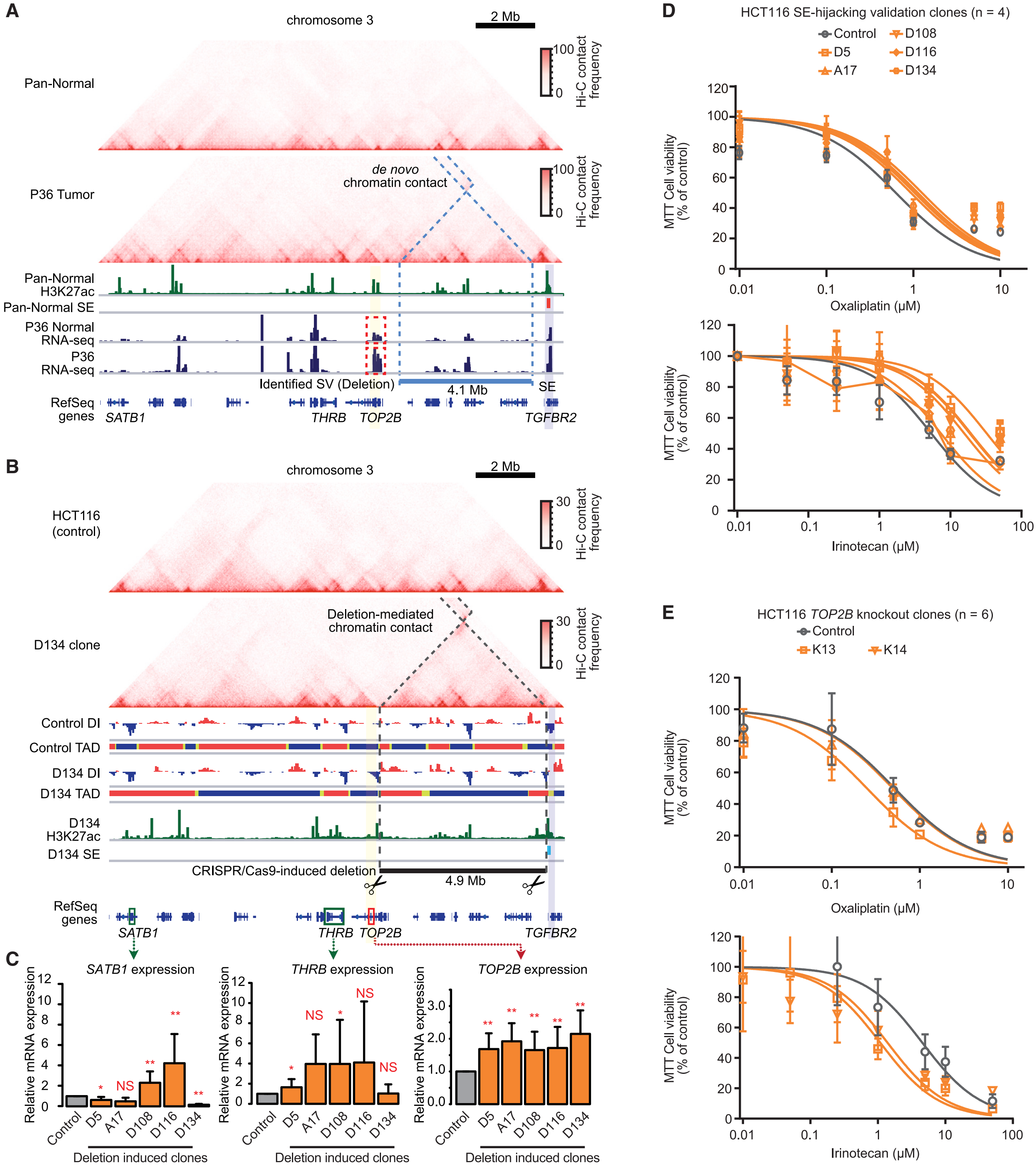


Figure 6. Validation and functional inference of the SE-hijacking genes

(A) Normalized Hi-C contact maps for pan-normal and P36 tumor sample at chr3:24.6–45.1 Mb (20 kb resolution) along with genome browser tracks for H3K27ac ChIP-seq and RNA-seq for the P36 tumor and the matched adjacent normal tissue. Positions of pan-normal SEs (red box) and identified sample SVs (blue stick) are also annotated. The *de novo* chromatin contacts are highlighted by a dashed blue line. *TOP2B* gene expression is highlighted by a dashed red box.

(B) Normalized Hi-C contact maps for the control and the D134 clone along with genome browser tracks for the H3K27ac ChIP-seq signal, SE positions, directionality index (DI) scores, and TADs. In the D134 clone, CRISPR-Cas9-mediated deletion of 4.9 Mb (scissors) was induced to simulate the genomic rearrangement in the P36 tumor.

(legend continued on next page)

- ○ Hazard ratio analysis
- ○ Experimental validation of the *TOP2B* gene overexpression by genomic deletion
- ○ Cell proliferation analysis
- ○ RNA-seq of CRISPR-Cas9 treated cells for cross validation
- ○ Western blot analysis
- ○ Growth inhibition assay
- ● QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2023.112778.

ACKNOWLEDGMENTS

We thank members of the Jung laboratory for their support and critical suggestions throughout the course of this work. We are thankful to Dr. Seongyeol Park, Dr. Jinho Jang, and Dr. Semin Lee for advising SV analysis; Dr. Eun-ji Choi for sharing raw WGS and RNA-seq data; and Dr. Jung-Kyoon Choi, Dr. Woong-Yang Park, and Dr. Hae-Ock Lee for their support. This work was supported by the Suh Kyungbae Foundation (to I.J.) and the Korea Ministry of Science and ICT (NRF-2023R1A2C3002773 and NRF-2021M3H9A2096767 to I.J. and NRF-2017M3A9A7050610 to T.-Y.K.).

AUTHOR CONTRIBUTIONS

K.K., T.-Y.K., and I.J. conceived the study. J.-K.K., S.-H.S., J.E., A.J.L., and Y.K. performed experiments. K.K., M.K., A.J.L., and J.-K.K. performed data analysis. Y.-J.K. contributed to WGS and RNA-seq data generation. G.H.K. and J.M.B. contributed to quantifying sample purity. Y.L. contributed to the provision of clinical and colon cancer tissue samples. H.S.K. contributed to data interpretation. K.K. and I.J. prepared the manuscript with assistance from M.K., A.J.L., S.-H.S., J.-K.K., S.M., and T.-Y.K. All authors read and commented on the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 26, 2022
Revised: May 4, 2023
Accepted: June 23, 2023
Published: July 13, 2023

REFERENCES

1. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380. https://doi.org/10.1038/nature11082.
2. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293. https://doi.org/10.1126/science.1181369.
3. Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., and Aiden, E.L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021.
4. Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331–336. https://doi.org/10.1038/nature14222.
5. Jung, I., Schmitt, A., Diao, Y., Lee, A.J., Liu, T., Yang, D., Tan, C., Eom, J., Chan, M., Chee, S., et al. (2019). A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat. Genet. *51*, 1442–1449. https://doi.org/10.1038/s41588-019-0494-8.
6. Schoenfelder, S., and Fraser, P. (2019). Long-range enhancer-promoter contacts in gene expression control. Nat. Rev. Genet. *20*, 437–455. https://doi.org/10.1038/s41576-019-0128-0.
7. Spielmann, M., Lupiáñez, D.G., and Mundlos, S. (2018). Structural variation in the 3D genome. Nat. Rev. Genet. *19*, 453–467. https://doi.org/10.1038/s41576-018-0007-0.
8. Franke, M., Ibrahim, D.M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W.L., et al. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature *538*, 265–269. https://doi.org/10.1038/nature19800.
9. Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012–1025. https://doi.org/10.1016/j.cell.2015.04.004.
10. Northcott, P.A., Lee, C., Zichner, T., Stütz, A.M., Erkek, S., Kawauchi, D., Shih, D.J.H., Hovestadt, V., Zapatka, M., Sturm, D., et al. (2014). Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature *511*, 428–434. https://doi.org/10.1038/nature13379.
11. Hnisz, D., Weintraub, A.S., Day, D.S., Valton, A.L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Fan, Z.P., Sigova, A.A., et al. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science *351*, 1454–1458. https://doi.org/10.1126/science.aad9024.
12. Weischenfeldt, J., Dubash, T., Drainas, A.P., Mardin, B.R., Chen, Y., Stütz, A.M., Waszak, S.M., Bosco, G., Halvorsen, A.R., Raeder, B., et al. (2017). Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. Nat. Genet. *49*, 65–74. https://doi.org/10.1038/ng.3722.
13. Yang, M., Safavi, S., Woodward, E.L., Duployez, N., Olsson-Arvidsson, L., Ungerbäck, J., Sigvardsson, M., Zaliova, M., Zuna, J., Fioretos, T., et al. (2020). 13q12.2 deletions in acute lymphoblastic leukemia lead to upregulation of FLT3 through enhancer hijacking. Blood *136*, 946–956. https://doi.org/10.1182/blood.2019004684.
14. Dixon, J.R., Xu, J., Dileep, V., Zhan, Y., Song, F., Le, V.T., Yardımcı, G.G., Chakraborty, A., Bann, D.V., Wang, Y., et al. (2018). Integrative detection and analysis of structural variation in cancer genomes. Nat. Genet. *50*, 1388–1398. https://doi.org/10.1038/s41588-018-0195-8.
15. Kloetgen, A., Thandapani, P., Ntziachristos, P., Ghebrechristos, Y., Nomikou, S., Lazaris, C., Chen, X., Hu, H., Bakogianni, S., Wang, J., et al. (2020).

(C) Bar plots showing the expression of *TOP2B* (marked with a red box) and adjacent genes (*THRB* and *SATB1*, marked with green boxes) measured by using qRT-PCR in multiple deletion-induced HCT116 clones (dark orange: D5, A17, D108, D116, and D134) compared with the control clones (gray). The error bars indicate the standard deviation of the results for the mutant clones after multiple replications, and the y axis indicates the mean values of the results. All of the mutant clones had significantly increased *TOP2B* expression after deletion based on two-sided KS tests (**p < 0.01, *p < 0.05, and NS, not significant).

(D and E) Dose-response plots showing the cell viability of the HCT116 clones (D, SE-hijacking induced; E, *TOP2B* knockout) against oxaliplatin and irinotecan via MTT assays. The color indicates the treatment (orange) and control (gray) clones. For the dose-response plots, the "n" indicates the number of replicates. The mean (marker) and standard deviation (error bar) of the replicated results for each clone are shown together.

See also Figure S6.

Three-dimensional chromatin landscapes in T cell acute lymphoblastic leukemia. Nat. Genet. *52*, 388–400. https://doi.org/10.1038/s41588-020-0602-9.

16. Akdemir, K.C., Le, V.T., Chandran, S., Li, Y., Verhaak, R.G., Beroukhim, R., Campbell, P.J., Chin, L., Dixon, J.R., Futreal, P.A., et al. (2020). Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. Nat. Genet. *52*, 294–305. https://doi.org/10.1038/s41588-019-0564-y.

17. Johnstone, S.E., Reyes, A., Qi, Y., Adriaens, C., Hegazi, E., Pelka, K., Chen, J.H., Zou, L.S., Drier, Y., Hecht, V., et al. (2020). Large-scale topological changes restrain malignant progression in colorectal cancer. Cell *182*, 1474–1489.e23. https://doi.org/10.1016/j.cell.2020.07.030.

18. Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., et al. (2015). The consensus molecular subtypes of colorectal cancer. Nat. Med. *21*, 1350–1356. https://doi.org/10.1038/nm.3967.

19. Kim, K., Eom, J., and Jung, I. (2019). Characterization of structural variations in the context of 3D chromatin structure. Mol. Cells *42*, 512–522. https://doi.org/10.14348/molcells.2019.0137.

20. Valton, A.L., and Dekker, J. (2016). TAD disruption as oncogenic driver. Curr. Opin. Genet. Dev. *36*, 34–40. https://doi.org/10.1016/j.gde.2016.03.008.

21. Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S.C., Gonzalez, S., Rosebrock, D., Mitchell, T.J., Rubanova, Y., Anur, P., Yu, K., et al. (2020). The evolutionary history of 2,658 cancers. Nature *578*, 122–128. https://doi.org/10.1038/s41586-019-1907-7.

22. Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C., and Chanda, S.K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat. Commun. *10*, 1523. https://doi.org/10.1038/s41467-019-09234-6.

23. Liu, Q., Diao, R., Feng, G., Mu, X., and Li, A. (2017). Risk score based on three mRNA expression predicts the survival of bladder cancer. Oncotarget *8*, 61583–61591. https://doi.org/10.18632/oncotarget.18642.

24. Koncina, E., Haan, S., Rauh, S., and Letellier, E. (2020). Prognostic and predictive molecular biomarkers for colorectal cancer: updates and challenges. Cancers *12*, 319. https://doi.org/10.3390/cancers12020319.

25. Canela, A., Maman, Y., Jung, S., Wong, N., Callen, E., Day, A., Kieffer-Kwon, K.R., Pekowska, A., Zhang, H., Rao, S.S.P., et al. (2017). Genome organization drives chromosome fragility. Cell *170*, 507–521.e18. https://doi.org/10.1016/j.cell.2017.06.034.

26. Woynarowski, J.M., Faivre, S., Herzig, M.C., Arnett, B., Chapman, W.G., Trevino, A.V., Raymond, E., Chaney, S.G., Vaisman, A., Varchenko, M., and Juniewicz, P.E. (2000). Oxaliplatin-induced damage of cellular DNA. Mol. Pharmacol. *58*, 920–927. https://doi.org/10.1124/mol.58.5.920.

27. Ho, S.S., Urban, A.E., and Mills, R.E. (2020). Structural variation in the sequencing era. Nat. Rev. Genet. *21*, 171–189. https://doi.org/10.1038/s41576-019-0180-9.

28. Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J.O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. Nat. Rev. Genet. *14*, 125–138. https://doi.org/10.1038/nrg3373.

29. Kohno, T., Ichikawa, H., Totoki, Y., Yasuda, K., Hiramoto, M., Nammo, T., Sakamoto, H., Tsuta, K., Furuta, K., Shimada, Y., et al. (2012). KIF5B-RET fusions in lung adenocarcinoma. Nat. Med. *18*, 375–377. https://doi.org/10.1038/nm.2644.

30. Shtivelman, E., Lifshitz, B., Gale, R.P., and Canaani, E. (1985). Fused transcript of abl and bcr genes in chronic myelogenous leukaemia. Nature *315*, 550–554. https://doi.org/10.1038/315550a0.

31. Lee, D.S., Luo, C., Zhou, J., Chandran, S., Rivkin, A., Bartlett, A., Nery, J.R., Fitzpatrick, C., O'Connor, C., Dixon, J.R., and Ecker, J.R. (2019). Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. Nat. Methods *16*, 999–1006. https://doi.org/10.1038/s41592-019-0547-z.

32. Li, G., Liu, Y., Zhang, Y., Kubo, N., Yu, M., Fang, R., Kellis, M., and Ren, B. (2019). Joint profiling of DNA methylation and chromatin architecture in single cells. Nat. Methods *16*, 991–993. https://doi.org/10.1038/s41592-019-0502-z.

33. Nagano, T., Lubling, Y., Várnai, C., Dudley, C., Leung, W., Baran, Y., Mendelson Cohen, N., Wingett, S., Fraser, P., and Tanay, A. (2017). Cell-cycle dynamics of chromosomal organization at single-cell resolution. Nature *547*, 61–67. https://doi.org/10.1038/nature23001.

34. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.

35. Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. *31*, 213–219. https://doi.org/10.1038/nbt.2514.

36. Saunders, C.T., Wong, W.S.W., Swamy, S., Becq, J., Murray, L.J., and Cheetham, R.K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics *28*, 1811–1817. https://doi.org/10.1093/bioinformatics/bts271.

37. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

38. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

39. Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., Janoueix-Lerosey, I., Delattre, O., and Barillot, E. (2012). Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics *28*, 423–425. https://doi.org/10.1093/bioinformatics/btr670.

40. Talevich, E., Shain, A.H., Botton, T., and Bastian, B.C. (2016). CNVkit: Genome-wide copy number detection and visualization from targeted DNA sequencing. PLoS Comput. Biol. *12*, e1004873. https://doi.org/10.1371/journal.pcbi.1004873.

41. Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, i333–i339. https://doi.org/10.1093/bioinformatics/bts378.

42. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

43. Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360. https://doi.org/10.1038/nmeth.3317.

44. Pertea, M., Kim, D., Pertea, G.M., Leek, J.T., and Salzberg, S.L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat. Protoc. *11*, 1650–1667. https://doi.org/10.1038/nprot.2016.095.

45. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

46. Favero, F., Joshi, T., Marquard, A.M., Birkbak, N.J., Krzystanek, M., Li, Q., Szallasi, Z., and Eklund, A.C. (2015). Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Ann. Oncol. *26*, 64–70. https://doi.org/10.1093/annonc/mdu479.

47. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137. https://doi.org/10.1186/gb-2008-9-9-r137.

48. Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319. https://doi.org/10.1016/j.cell.2013.03.035.

49. Bankhead, P., Loughrey, M.B., Fernández, J.A., Dombrowski, Y., McArt, D.G., Dunne, P.D., McQuaid, S., Gray, R.T., Murray, L.J., Coleman, H.G., et al. (2017). QuPath: Open source software for digital pathology image analysis. Sci. Rep. *7*, 16878. https://doi.org/10.1038/s41598-017-17204-5.

50. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222. https://doi.org/10.1093/bioinformatics/btv710.

51. Lee, J.J.K., Park, S., Park, H., Kim, S., Lee, J., Lee, J., Youk, J., Yi, K., An, Y., Park, I.K., et al. (2019). Tracing oncogene rearrangements in the mutational history of lung adenocarcinoma. Cell *177*, 1842–1857.e21. https://doi.org/10.1016/j.cell.2019.05.013.

52. Kim, K., and Jung, I. (2021). covNorm: an R package for coverage based normalization of Hi-C and capture Hi-C data. Comput. Struct. Biotechnol. J. *19*, 3149–3159. https://doi.org/10.1016/j.csbj.2021.05.041.

53. Crane, E., Bian, Q., McCord, R.P., Lajoie, B.R., Wheeler, B.S., Ralston, E.J., Uzawa, S., Dekker, J., and Meyer, B.J. (2015). Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature *523*, 240–244. https://doi.org/10.1038/nature14450.

54. Yang, D., Jang, I., Choi, J., Kim, M.S., Lee, A.J., Kim, H., Eom, J., Kim, D., Jung, I., and Lee, B. (2018). 3DIV: a 3D-genome interaction viewer and database. Nucleic Acids Res. *46*, D52–D57. https://doi.org/10.1093/nar/gkx1017.

55. Canny, J. (1986). A computational approach to edge-detection. IEEE T Pattern Anal. *8*, 679–698. https://doi.org/10.1109/Tpami.1986.4767851.

56. Liang, Y.D., and Barsky, B.A. (1984). A new concept and method for line clipping. ACM Trans Graph. *3*, 1–22.

57. Korbel, J.O., and Campbell, P.J. (2013). Criteria for inference of chromothripsis in cancer genomes. Cell *152*, 1226–1236. https://doi.org/10.1016/j.cell.2013.02.023.

58. Song, S.H., Jeon, M.S., Nam, J.W., Kang, J.K., Lee, Y.J., Kang, J.Y., Kim, H.P., Han, S.W., Kang, G.H., and Kim, T.Y. (2018). Aberrant GATA2 epigenetic dysregulation induces a GATA2/GATA6 switch in human gastric cancer. Oncogene *37*, 993–1004. https://doi.org/10.1038/onc.2017.397.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Histone H3K27ac antibody (pAb)	Active Motif	Cat# 39133; RRID: AB_2561016
Cytokeratin (AE1/AE3)	Dako	Cat# M3515; RRID: AB_2132885
Biological samples		
Tumor and normal colorectal primary tissues	Seoul National University Hospital	N/A
Chemicals, peptides, and recombinant proteins		
MboI restriction enzyme	NEB	Cat# R0147
Protease Inhibitor Cocktail Tablets	Roche	Cat# 04-693-159-001
T4 DNA Ligase	NEB	Cat# M0202
Proteinase K	NEB	Cat# P8102
Dynabeads^{TM} Protein A	Thermo Fisher	Cat# 10001D
RNase A	QIAGEN	Cat# 19101
Critical commercial assays		
AllPrep DNA/RNA Mini Kit	QIAGEN	Cat# 80204
TruSeq DNA Nano Low Throughput Library Prep Kit	Illumina	Cat# 20015964
TruSeq Stranded Total RNA Library Prep Kit with Ribo-Zero Gold	Illumina	Cat# RS-122-2301
NEBNext® Ultra^{TM} II DNA Library Prep Kit	NEB	Cat# E7645
NucleoSpin RNA XS kit	Macherey-Nagel	Cat# MN740902.50
ViraPower^{TM} Lentiviral Packaging Mix	Invitrogen	Cat# K497500
Deposited data		
In situ Hi-C raw sequencing data	This paper	GEO: GSE137188
Raw WGS and RNA-seq data	Center for integrative Omics and Precision Medicine (CoPM) consortium	N/A
CRC RNA-seq data	TCGA	N/A
Experimental models: Cell lines		
HCT116 cells	Korean Cell Line Bank (KCLB)	Cat# 10247
Oligonucleotides		
ERCC (External RNA controls consortium) RNA Spike-In Mix	Thermo Fisher	4456740
CRISPR gRNA	This paper	Method details
gDNA PCR primers	This paper	Method details
RT-qPCR primers	This paper	Method details
Recombinant DNA		
lentiCRISPRv2 vector	Addgene	Cat# 52961
Software and algorithms		
Code for CAPReSE	This paper	https://github.com/kaistcbfg/CAPReSEv1
Genome analysis toolkit (GATK) ver.4.1.4 and Mutect2	McKenna et al. and Cibulskis et al.[34,35]	https://github.com/broadinstitute/gatk
Strelka2 ver. 2.9.7	Saunders et al.[36]	https://github.com/Illumina/strelka
SAMtools	Danecek et al.[37]	https://github.com/samtools/samtools
Burrows-Wheeler aligner (BWA) ver. 0.7.17	Li and Durbin[38]	https://bio-bwa.sourceforge.net/
FREEC ver. 11.5	Boeva et al.[39]	http://boevalab.inf.ethz.ch/FREEC/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
CNVkit ver. 0.9.6	Talevich et al.[40]	https://github.com/etal/cnvkit
DELLY ver. 0.7.6	Rausch et al.[41]	https://github.com/dellytools/delly
Trimmomatic ver. 0.38	Bolger et al.[42]	http://www.usadellab.org/cms/?page=trimmomatic
HISAT2 ver. 2.1.0	Kim et al.[43]	http://daehwankimlab.github.io/hisat2/
Stringtie ver. 1.3.5	Pertea et al.[44]	https://ccb.jhu.edu/software/stringtie/
DESeq2 ver. 1.22.2	Love et al.[45]	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
Sequenza ver. 2.1.2	Favero et al.[46]	https://cran.r-project.org/web/packages/sequenza/vignettes/sequenza.html
CMS classifier	Guinney et al.[18]	https://github.com/Sage-Bionetworks/CMSclassifier
MACS2 ver. 2.1.1	Zhang et al.[47]	https://github.com/macs3-project/MACS
Rank ordering of super-enhancers (ROSE)	Whyte et al.[48]	http://younglab.wi.mit.edu/super_enhancer_code.html
Domaincaller	Dixon et al.[1]	https://github.com/XiaoTaoWang/domaincaller
hic_breakfinder	Dixon et al.[14]	https://github.com/dixonlab/hic_breakfinder
MutationTimeR	Gerstung et al.[21]	https://github.com/gerstung-lab/MutationTimeR
QuPath ver. 0.4.1	Bankhead et al.[49]	https://qupath.github.io/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Inkyung Jung (ijung@kaist.ac.kr).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Hi-C and ChIP-seq data used in this study have been deposited in the GEO repository (GEO: GSE137188) and are publicly available as of the date of publication. Visualization of processed Hi-C data is available at 3DIV database (http://3div.kr/cancer_hic). Scanned tissue sample slide images with examination results, details, and figures are available at custom web repository (http://junglab.kaist.ac.kr/Dataset/CellReports_tissueImage.html).
- Custom code supporting this work has been deposited in the GitHub repository (https://github.com/kaistcbfg/CAPReSEv1).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Colorectal cancer tissue samples and clinical data

Tumor and adjacent normal tissues of 40 colorectal cancer patients (38 patients were diagnosed as stage III and two patients were diagnosed as stage IV) were used in this study. The biospecimens for this study were provided by the Seoul National University Hospital (SNUH) Cancer Tissue Bank. All samples derived from the Cancer Tissue Bank of SNUH were obtained with informed consent under institutional review board approved protocols. Institutional Review Board (IRB) approval was obtained from Seoul National University College of Medicine and IRB waiver was approved by KAIST for the use of these tissues.

The tissues of the patients were initially obtained through surgery at SNUH between 2009 and 2016. After the surgery, the tissue samples were placed in the cryotube vial with isopentane solution and stored in the liquid nitrogen (-196°C). Two pieces of 0.2–0.3 cm^3 tissue samples were placed per vial. During the sampling of the tumor tissues, tumor-normal tissue border regions were excluded to avoid inclusion of normal tissues. For the normal tissue sampling, near-mucosa regions which are far from the tumor site were sampled. Pathological information (adenocarcinoma, mucinous carcinoma, or signet-ring cell carcinoma) were also determined by examination of SNUH pathologists right after the surgery. Electronic medical records of the patients including age, gender, tumor site, and progression free survival (PFS) with the pathological information were also provided by SNUH.

Cell culture

HCT116 cells were obtained from the Korean Cell Line Bank (KCLB). The identity of HCT116 cells has been authenticated by KCLB using short tandem repeat (STR) profiling. Cells were cultured in RPMI 1640 (Cytiva, SH30027.01) supplemented with 10% fetal bovine serum (WELGENE, S001-01) and gentamicin (Corning, 30-005-CR) (10 μg/mL) at 37 °C in a 5% CO_2-humidified atmosphere as suggested by the supplier. Cells were regularly tested for mycoplasma contamination using MycoAlert Mycoplasma Detection Kit (Lonza, LT07-318) according to the manufacturer's instructions.

METHOD DETAILS

Tumor tissue purity evaluation

The purity of the banked tumor tissues was re-validated in terms of cellularity after being retrieved for this study. Tumor tissue samples with available hematoxylin and eosin (H&E)-stained formalin-fixed, paraffin-embedded (FFPE) slides (n = 37) were scanned with Aperio AT2 scanner (Leica Biosystems) at 40× magnification and imported into QuPath software (ver. 0.4.1), an open-source software that provides multiple "tools" for the slide image analysis.[49] The tissue border of each sample was manually outlined by using the "Polygon tool". The "Cell detection tool" was used to identify individual cells. Several groups of cells of the same cell type were manually outlined with the Polygon tool, and annotated as "tumor" or "stroma". Then, the entire detected cells (average of 495,046 cells per sample) were classified into tumor cells or stromal cells using "random trees" of the "Object classifier". For the nucleus, cell, and cytoplasm, 17 features related to morphology (area, perimeter, circularity, max/min caliper, and eccentricity) and color (mean, sum, standard deviation, max/min, and range) were measured. By adding the nucleus/cell ratio to the feature list, a total of 52 features were prepared and used for the classification. Tumor purity was quantified as the number of tumor cells divided by the number of total detected cells. Manual processes required for QuPath software were conducted by a trained pathologist (JMB).

To corroborate the purity obtained by QuPath software, a manual assessment of tumor purity was also performed. Another pathologist (GHK) selected random 7 rectangular areas (0.265×0.211 mm^2) on the virtual slides of 10 random tumor samples. Then, all of the tumor cells within the areas were first counted followed by the non-tumor cell nuclei counting. From the median values of the tumor cell count and non-tumor cell count in rectangular areas, the proportion of tumor cells among total cells was determined for each sample. A strong correlation (Pearson's correlation coefficient, PCC = 0.727) between the manual and QuPath-measured tumor purity shows the consistency of our evaluation criteria and validates QuPath-measured purity values.

Immunohistochemistry

Immunohistochemistry (IHC) was also performed for the further validation of QuPath-measured H&E staining-based purity values. Samples with additional available tissue blocks were used (n = 30). Monoclonal mouse anti-human cytokeratin antibodies cocktail (clone AE1/AE3) were used as the markers (Dako, M3515). The marker antibodies were validated by confirming the positive staining in epithelial cells (30 normal tissues were used for the test). Whole IHC procedures were conducted using Ventana BenchMark XT (Roche) automated immunostainer system, according to the manufacturer's protocol.

After the staining, the trained pathologist (JMB) examined the slides to count total tumor and non-tumor cells to measure IHC-based purity. QuPath software was also used for this procedure. Correlation (PCC) was measured with 27 samples which have both H&E staining and IHC-based purity data. A strong correlation of 0.738 was observed which confirms the quality of our retrieved tumor tissues and robustness of purity measurement methods.

Whole-genome sequencing

The whole-genome sequencing (WGS) fastq files for patients' tumor and adjacent normal samples were obtained from the Center for integrative Omics and Precision Medicine (CoPM) consortium. In brief, DNA was extracted from the tissue using AllPrep DNA/RNA Mini Kit (QIAGEN, 80204). The whole-genome library was prepared with TruSeq DNA Nano Low Throughput Library Prep Kit (Illumina, 20015964) and sequenced with Illumina HiSeq X platform with 60X coverage for 40 tumor samples and 30X coverage for adjacent normal tissue samples.

Whole-genome sequencing data processing

Whole-genome sequencing data were processed according to the genome analysis toolkit (GATK) best practice[34] (ver. 4.1.4). Raw WGS reads were converted into unmapped binary alignment map (BAM) files using Picard FastqToSam (ver. 2.18.12) with default parameters. BAM files were aligned to human genome reference (GRCh38) using Burrows-Wheeler aligner (BWA)-mem (ver. 0.7.17). Adapter sequences and polymerase chain reaction (PCR)/optical duplicates were marked. Base quality scores were also recalibrated.

Somatic SNV and INDEL calling

In order to identify somatic single nucleotide variants (SNVs) and Insertion/Deletions (INDELs), we run Mutect2 (GATK ver. 4.1.4) and Strelka2 (ver. 2.9.7)[35,36] with default parameters using 40 CRC tumor-matched adjacent normal tissue aligned BAM files as an input. Of note, Strelka2 also utilizes the INDEL candidates identified by Manta[50] (ver. 1.4.0) to rescue additional INDEL candidates. We further filtered out germline variants and sequencing artifacts based on panel of normal (pan-normal) that were generated by Mutect2 using 391 normal Korean blood samples provided by the CoPM consortium. The union of identified somatic SNVs and INDELs from Mutect2 and Strelka2 was used for the downstream analysis.

Copy number variation calling

Copy Number Variation (CNV) profiles of tumor samples were estimated using control-FREEC[39] (ver. 11.5) and CNVkit[40] (ver. 0.9.6.dev0). 40 CRC tumor-matched adjacent normal tissue WGS BAM files were given as input for both methods. We converted output CNV profiles from each method into 40 kb resolution by selecting the mode values of each bin. When the copy number values were not consistent between the two methods, a value similar to the inverse of the B-Allele Frequency (BAF) was selected. BAF indicates a normalized measure of the allelic intensity ratio between two alleles, which was calculated by processing the WGS BAM files with Samtools (option 'mpileup –AB –Q 20 –q') and normalized by control-FREEC (tumor-normal paired input mode). If neither of the method output matches the BAF, copy number was not assigned to the corresponding region. The copy number values of the centromere and the telomere region (+1 Mb) were excluded due to low mapping quality (MQ).

Large-scale somatic structural variation calling

A large-scale somatic structural variations (SVs) were identified with DELLY[41] (ver. 0.7.6) and filtered by a previously published guideline.[51] First, SV candidates were obtained by processing WGS BAM files with DELLY (tumor-normal paired input mode).

We further refined somatic SV candidates by applying automated filtration processes to identify precise coordinates of breakpoints and rescue missed breakpoints by examining SA (other canonical alignments in a chimeric alignment) tags of aligned reads proximal (700 bp forward, 100 bp backward direction) to the breakpoints of SV candidates. The coordinates of breakpoints were determined if the coordinates of SA tag and the corresponding DELLY breakpoints are consistent. If they were inconsistent, loci supported by most SA tags were selected as the coordinates of breakpoints.

The refined somatic SV candidates have filtered out again according to the following criteria: 1) overlap with germline SVs identified at least one matched normal samples, 2) minimum MQ < 20 or median MQ < 40 for both breakpoints, 3) all SV supporting reads are mapped within 2 bp distance from the breakpoints, 4) no supporting split reads for duplication event with breakpoints distance <10 kb or deletion event with breakpoints distance <1 kb, and 5) if less than 5 supporting split reads exists for duplication, deletion, and inversion event smaller than 1 kb, 1 kb, and 5 kb distance between breakpoints, respectively.

Finally, the filtered SVs were manually curated to remove false positives which are difficult to be removed by the automated filtration processes. To this end, we use the MQ and three tag information (AS: alignment score generated by aligner, UQ: Phred likelihood of the segment, conditional on the mapping being correct, and XS: suboptimal alignment score) of aligned reads nearby breakpoints to identify true SVs with high confidence. We kept only SV candidates satisfying the following criteria.

(1) for SV breakpoints with supporting split reads, at least 3 paired-end reads support the SV event and at least one read supporting each breakpoint satisfies MQ = 60, AS = 151, XS $\leq$ 40, and UQ $\leq$ 20.
(2) for SVs breakpoints without supporting split reads, at least 4 paired-end reads support the SV event and at least one read supporting each breakpoint satisfies MQ = 60, AS = 151, XS $\leq$ 30, and UQ = 0.

Of note, in the case of candidates in low MQ regions such as repeat sequence, less stringent criteria were applied.

Computational tumor purity measurement

Sequenza pipeline[46] (ver. 2.1.2) was used to determine tumor purity. After receiving WGS bam files as input, preprocessing procedures such as seqz file format conversion and copy number smoothing were performed before the purity was calculated. Sequenza was applied with a default option to obtain the purity of each sample. The purity values provided by the two software (Control-FREEC and CNVkit) were also examined. Sequenza showed the best-fit purity value (mean = 0.484) with the purity calculated with the H&E stained tumor tissue slide using QuPath (mean = 0.522); however, the values of Control-FREEC (mean = 0.77) and CNVkit (mean = 0.75) showed a significant difference in distribution (KS-test, p value = 5.55e-09 for both methods when compared with Sequenza). Therefore, the value of Sequenza was used as a representative computational purity value.

RNA sequencing

RNA-seq fastq files for patients' tumor and matched normal samples were obtained from the CoPM consortium. In brief, RNA was extracted from the patients' tissue samples using AllPrep DNA/RNA Mini Kit (QIAGEN, 80204). Total RNA sequencing libraries were constructed using TruSeq Stranded Total RNA Library Prep Kit with Ribo-Zero Gold (Illumina, RS-122-2301). RNA-seq library was sequenced in 100 bp paired-end mode on Illumina HiSeq 4000 system at 120 million reads per sample. ERCC (External RNA controls consortium) RNA Spike-In Mix (Thermo Fisher, 4456740) were included for quality assurance.

RNA-sequencing data processing

The adapter sequences were trimmed with the trimmomatic[42] (ver. 0.38) software using TrueSeq3-PE-2 adapter file. The trimmed sequences were then aligned to the reference genome (hg38 with ERCC) using HISAT2 software[43] (ver. 2.1.0) and transcripts were quantified by StringTie[44] (ver. 1.3.5) software. GENCODE v27 with ERCC was used as a gene transfer format (GTF) reference input file; a list of annotated protein-coding genes and long non-coding RNAs (lncRNAs) (level 1 and 2 for protein-coding genes and level 1 for lncRNAs, n = 27,313) were used for quantification. The calculated fragments per kilobase of transcripts per million mapped reads (FPKM) values obtained from StringTie were log2 transformed with a pseudocount value of 1 and quantile normalized. The ERCC correlation of each sample against the ERCC input and other samples was examined to check sequencing quality and reproducibility. To calculate differentially expressed genes (DEGs) between normal and tumor samples, the StringTie assembled data was converted into gene count format using prepDE.py script provided by StringTie. The gene count file was provided as an input for DESeq2 R package[45] (ver. 1.22.2) to calculate log2 fold change and adjusted p value for each gene. The genes over 2-fold change with adjusted p value <0.01 were selected as the final DEGs (n = 4,518). The principle component analysis (PCA) was applied to check separation of 40 CRC patients' adjacent normal tissue and tumor samples in terms of gene expression. Due to functional clarity, we focused on protein-coding genes only in downstream analysis.

Identification of CMS for tumor samples

We identified the consensus molecular subtype (CMS) of the CRC tumor samples based on expression levels of 5,973 genes as an input for the CMS classifier.[18] The predicted CMS subtypes were validated by using known mutation signatures for each subtype and comparing the ratio of each subtype with the published data.

In situ Hi-C library preparation

In situ Hi-C experiment was performed on 40 CRC tumors, 10 adjacent normal samples, and HCT116 cell line clones. Approximately 30–50 mg of tissues was placed in sterilized aluminum foil and repeatedly hammered while frozen in liquid nitrogen to produce tissue powder. The pulverized tissue was cross-linked with 1% formaldehyde. *In situ* Hi-C experiment was then conducted on the samples using the protocol modified from a previously study.[3] Cross-linked cells were lysed (10 nM Tris-HCl, 10 mM NaCl, and 0.2% IGEPAL CA-630 (Sigma-Aldrich, 18896), pH 8.0) and digested with 100U *Mbo*I (NEB, R0147). Digested fragments were labeled with Biotin-14-dCTP (Invitrogen, 19518018) and proximally ligated with T4 DNA Ligase (NEB, M0202), followed by reverse cross-linking (Proteinase K (NEB, P8102) and 10% SDS). The fragments were then sonicated (Covaris, S220) and purified with AMPure XP Reagent (Beckman Coulter, A63880). The final ligated DNA was pulled down with Dynabeads MyOne Streptavidin T1 (Invitrogen, 65602), followed by manual library preparation. Hi-C library was sequenced with Illumina HiSeq 4000 or NovaSeq 6000 platforms.

Hi-C data processing

Reads from Hi-C sequenced data were mapped to the reference genome (hg38) using BWA-mem ('–M' option). We removed low-quality reads (MQ < 10), reads that span ligation sites, and self-ligation reads where two fragments are located within 15 kb. The read-pairs were merged together as paired-end aligned BAM files, and PCR duplicates were removed with Picard (ver. 2.18.12). To remove multiple sources of intrinsic biases of raw Hi-C data, we used covNorm[52] (ver. 1.0.0) and created normalized Hi-C contact maps in 40 kb resolution.

Topologically associating domain calling and pan-normal Hi-C contact map generation

We systemically identified topologically associating domains (TADs) of normalized 40 kb resolution Hi-C contact maps by processing the directionality index (DI) score with Domain Caller.[1] Then, the genome was partitioned into TAD, boundaries, and unstructured regions. Unstructured regions were defined when the boundary was larger than 400 kb. The pan-normal TADs were generated by processing merged Hi-C data of 10 adjacent normal samples. Also, we combined (pixel-wise median) Hi-C contact maps of 10 adjacent normal samples and defined a pan-normal Hi-C contact map.

Measurement of insulation score and TAD boundary strength

The insulation score of each 40 kb bin was obtained by calculating a mean of 400 kb × 400 kb ([bin_index-10, bin_index], [bin_index, bin_index+10]) square window along the diagonal of the Hi-C contact map.[53] The raw insulation score was normalized by the log2 ratio of the mean of all bins' insulation scores in each chromosome. The insulation score delta (Δ) of a given bin was defined as the insulation score difference between the upstream 3 bins and the downstream 3 bins from the given bin. Boundary strength is defined as a subtraction between the nearest Δ local maximum and the local minimum.

Development of a machine learning-based method to detect SVs from Hi-C contact maps

To identify SVs using Hi-C contact maps, we developed a new method using image processing and machine learning algorithms. The developed method was implemented by Python including python-OpenCV (ver. 2.4.9.1), PyTorch (ver. 0.4.1), and python-XGBoost (ver. 0.82) packages. 20 kb resolution for *cis*- and 500 kb resolution *trans*- Hi-C contact maps were prepared as an input of our method. Quantile normalization across the *trans*- Hi-C contact maps was applied to allow a robust comparative analysis. For 20 kb resolution *cis*- Hi-C contact maps, quantile normalization was not applicable due to the requirements of large computing power.

Instead, the mean and standard deviation of each *cis*- Hi-C contact map within the same chromosome were measured and the average mean and standard deviation were computed by adopting procedures in the previous publication.[54] All values were scaled such that mean and standard deviation of each Hi-C contact map were equal to (or similar) the average mean and standard deviation.

To identify each sample specific long-range chromatin interactions, normalized 20 kb resolution *cis*- and 500 kb resolution *trans*- Hi-C contact maps were divided by the pan-normal Hi-C contact map (pseudocount of 1 was added). In the case of normalizing adjacent normal samples, a sample specific pan-normal Hi-C contact map was generated by excluding the sample of interest to avoid the removal of sample specific signals. By dividing Hi-C contact maps with the pan-normal Hi-C contact map, potential artifacts were crossed out. The signals at the centromere regions were removed due to low mapping quality. The normalized Hi-C contact maps were converted into image files after ceiling the maximum value to 5-fold for *cis*- and 15-fold for *trans*-.

Candidate regions for the SV-mediated tumor specific chromatin contacts were obtained by applying a series of image filters to the given sample's normalized Hi-C contact map. The median blur was applied to remove 'salt and pepper' noise generated by division. A kernel size of 3 was used as single pixel-sized signals cannot have any gradient. The bilateral filter was applied to reduce texture-like patterns and enhance edge-like patterns. The SVs are known to generate 'gradient' shaped patterns on the Hi-C contact map.[19] These '*de novo* chromatin contacts' were defined as where the signal propagated from the breakpoints is no longer distinguished from the background. To obtain this information, Canny edge detection algorithm[55] was applied with a minimum threshold parameter corresponding to a 1-fold change (no change compared to the pan-normal) of the normalized Hi-C contact map, which was used to obtain the distinguishable edges from the background. The contour search algorithm was applied to find separated contours formed by the edges. Liang-Barsky algorithm[56] was used to remove contours overlapping the diagonal axis which are potential false positive signals generated by the strong chromatin contact frequencies at a closer distance. Identified contours were converted to equal sized crop (32 × 32) centered at the potential breakpoints; after dividing the crop into four square regions, the vertex coordinate of the crop located in the square region with the strongest pixel intensity was selected.

Applying a series of image processing techniques, a large number of abnormally enriched, distant signals from the Hi-C contact maps could be identified. Previously published software also tried to find SV-mediated Hi-C signals with their own approaches such as statistical significance testing. However, there are multiple areas on the Hi-C contact map that show a sharp signal difference, such as the boundary of low coverage regions or compartment patterns, even though they are not SV. These patterns are difficult to properly classify into a simple metric and thereby generate many false positives. By adding a machine learning-based classifier at the final stage, false positives were reduced as much as possible. The crops prepared at the previous steps were used as the input of the machine learning architecture and classified whether it is true *de novo* chromatin contacts or not.

For the training of the machine learning architecture, the breakpoints centered *de novo* chromatin contacts were collected. For the unbiased selection of the positive dataset, *de novo* chromatin contacts consistent with the WGS-identified SV breakpoints were selected. Crops from the randomly selected regions were used as a negative dataset. 127 positives and 150 negatives were selected from 12 different samples and image augmentation (three times of 90-degree rotation and subsampling) was applied to increase the train set into 2,540 positives and 2,400 negatives.

The transferred feature learning technique was used to classify breakpoints centered crops as the number of the available train set was not ideal for deep neural networks (DNN) training from the scratch. Pre-trained DNN was used as a feature extractor and values from the final fully connected layer before the softmax layer were used for an input of the secondary classifier. The Modified National Institute of Standards and Technology (MNIST) dataset was used for the DNN pre-training as the features of the MNIST dataset was similar to the gradient patterns (black and white, lines and corners only). Also, the MNIST dataset can be well-trained to light weighted networks, which decreases the computational burdens of the method. The official MNIST classification example offered by Pytorch GitHub was used. The convolutional neural network (CNN) with two convolution layers followed by two fully connected layers which generates a 10-length feature vector was used. The batch size of 64, 10 epochs, 0.01 learning rate, and SGD optimizer with 0.5 momentum were used as the hyperparameters which were provided by the official Pytorch example. The XGBoost as the binary decision mode was used as a secondary classifier which makes the final classification decision. Total 10 runs of 2-fold cross validation were done to check the train phase accuracy and robustness of the trained model.

For the best result, the WGS-guided mode was also developed. The known WGS SV breakpoints were used as a region proposal if the WGS SV breakpoints were covered by none of the image-based detection results. For all candidate contour regions initially classified as 'false', the second classification trial was applied if the contour region matches the WGS breakpoints (contain the breakpoint in region or breakpoint located within single bin distance). In this case, WGS breakpoints were used as a new crop center. Very small SVs or misclassified patterns were rescued and included in the final *de novo* chromatin contacts list. Serious errors were finally filtered out by manual curation.

Refinement of the identified breakpoints

For the exact estimation of the breakpoints and SV (*de novo* chromatin contacts) detection results, fine mapping of the identified breakpoints was conducted. The developed method contains multiple image denoising processes. These processes help faster and easier identification of the breakpoints but pixel-level exact information can be missed as contour boundaries are manipulated. The raw Hi-C contact map crops corresponding to the called regions were extracted. The best fit point with the highest raw interaction frequencies was selected as breakpoints. To avoid miss selection by noise, nearby pixel information was also considered when selecting the best fit point. The 20 kb and 40 kb resolution were used for fine mapping of *cis*- and *trans*-results, respectively. In the case of the translocation, the coordinates of *de novo* chromatin contacts were also readjusted into the 40 kb resolution by re-searching contours containing the corresponding breakpoints in the raw 40 kb Hi-C contact map.

Performance evaluation of the developed method

For the measurement of the performance, the proposed method was compared with the published code. The 'hic_breakfinder' software[14] was used as a benchmark. The benchmark program receives filtered paired-read bam file and author-provided inter-/intra-chromosomal expectation file as input and computes the sub matrices of the original contact map which may contain structural rearrangements and breakpoints. The Hi-C BAM files of 17 CRC tumor samples with manually curated WGS-identified SV information and 10 adjacent normal samples were processed by hic_breakfinder and the performance of called results were compared with the proposed method. The WGS-identified SVs were used as a true positive set for the measuring sensitivity and the called SVs from 10 adjacent normal samples were used for measuring specificity. Germline SVs were removed in specificity measurement by excluding DELLY called germline SV candidates (MQ > 20 with more than 3 supporting reads). For detailed comparison, both results from the no-WGS mode and WGS guided mode were compared with the benchmark data.

Tumor purity robustness test of the developed algorithm

Two tests were applied to check whether the SV (*de novo* chromatin contacts) detection is affected by purity or whether our developed method functions robustly under various tumor purity conditions. Firstly, 40 CRC tumor samples were sorted based on the number of Hi-C found SVs, and the Sequenza purities of the top-10 and bottom-10 samples were compared. The 20 samples from both extremes showed no or marginal purity differences with a non-significant p value (0.164, KS test).

The second test was conducted by using 'diluted' tumor Hi-C contact maps. The P40 sample has a ~60% tumor purity, and this sample's chromosome 8 shows >100 SVs which makes it as a good testbed to check a wide variety of purity conditions. After adding the scale factor-multiplied (0.2, 0.4, 0.6, 0.8, and 1.0) normal Hi-C contact maps to the tumor Hi-C contact map, the scale was adjusted using the mean and standard deviation normalization method mentioned in the previous section. Scaled diluted tumor Hi-C contact maps were then divided by pan-normal and tested. The detection result of the original (not diluted P40 chromosome 8) Hi-C contact map was used as a true positive for measuring robustness.

Identification of complex genomic rearrangements

The complex genomic rearrangements were identified according to previously reported criteria.[51,57] For the chromoplexy, we systematically scanned multiple translocation events linking at least three chromosomes. Linked SVs were defined as SVs sharing breakpoints within 5 Mb. In the case of complex *cis*-rearrangements, chromosomes containing more than 10 *de novo* chromatin contacts were checked. The chromosomes with the oscillatory copy number in 40 kb resolution were identified as a potential chromothripsis which occurred by a single catastrophic event while sequentially accumulated complex genomic rearrangements tend to show a stepwise increase of the copy numbers.

ChIP-seq library preparation

We conducted ChIP-seq to profile genome-wide lysine 27 acetylation (H3K27ac) landscape of 10 adjacent normal samples (P6, P7, P10, P14, P15, P24, P30, P32, P33, and P40) and two CRC tumor samples (P14 and P32). Approximately 40–50 mg tissue was placed in tissueTUBE Extra Thick TT05M XT (Covaris, 520140), and repeatedly hammered while frozen in liquid nitrogen to produce tissue powder. The tissue sample was cross-linked in cross-linking buffer (100 mM NaCl, 0.1 mM EDTA, 5 mM HEPES, 1% formaldehyde, pH 8.0) for 10 min at 25°C. The cross-linking was quenched with 125 mM glycine in 25°C for 5 min with rotation, and washed twice with ice-cold phosphate-buffered saline (PBS). The samples were passed through 30 μm strainer (Sysmex, 04-0042-2316) to remove excessive debris, and suspended in SDS lysis buffer (1% SDS, 50 mM Tris-HCl, 10 mM EDTA, pH 8.0) with cOmplete, Mini, EDTA-free Protease Inhibitor Cocktail Tablets (Roche, 04-693-159-001). Mono- and di-nucleosome size chromatin was obtained through sonication (Covaris, S220). The sonicated chromatin was incubated with anti-acetyl-histone H3 (Lys27) antibody (Active Motif, 39133) and Dynabeads Protein A for Immunoprecipitation (Thermo Fisher, 10001D) for 4 h in 4°C with rotation, while a fraction of the input chromatin was stored to be used as input control. The chromatin-antibody-bead complex was subjected to serial washing. The immunoprecipitated complex and input chromatin were treated with RNase A (QIAGEN, 19101), and reverse cross-linked overnight at 68°C. The DNA was recovered using AMPure XP (Beckman Coulter, A63881), and ChIP DNA and input DNA libraries were prepared using NEBNext Ultra II DNA Library Prep Kit (NEB, E7645) following the manufacturer's instruction. The ChIP-seq libraries were sequenced using Illumina NextSeq 550 and HiSeq 4000 platforms.

ChIP-seq data processing

ChIP-seq reads aligned to the human reference genome (hg38) using BWA-mem (ver. 0.7.17, '–M' option). Reads with MQ < 10 were removed, and PCR duplicates were discarded using Picard (ver. 2.18.12). Peaked regions, relative to the input, were identified using MACS2[47] (ver. 2.1.1.20160309) with default parameters.

Identification of pan-normal super-enhancers

H3K27ac peaks were used as an input in the ROSE algorithm to call super-enhancers (SE).[48] To assess the activity of super-enhancers, reads per million mapped reads (RPM) of H3K27ac ChIP-seq data was subtracted from its corresponding input. Although the number of enhancers and the exact genomic location vary from samples, generally the region annotated as super-enhancer showed consistently strong H3K27ac signals in all adjacent normal samples occupying a large portion of the H3K27ac peaks. Thus, we merged super-enhancer regions from all of the normal samples to create a pan-normal super-enhancer list.

Identification of SE-hijacking candidates

The identified *de novo* chromatin contacts by the proposed method indicate tumor-specific newly formed long-range chromatin interactions which can mediate interactions of distal super-enhancers and promoters. To investigate the regulatory effect of the newly established long-range chromatin interactions, all SE and promoter pairs in which their coordinates intersect in the *de novo* chromatin contacts were collected. To find SE-promoter pairs that were inherently distant but spatially close after rearrangement, SE and promoter coordinates located at the opposite side of the *de novo* chromatin contacts on the perspective of the breakpoints were selected as a candidate pair. SE-promoter pairs with significant interaction frequencies were identified by considering genomic distance-dependent background interaction frequencies as an expectation value. The interaction frequency of the SE-promoter pair was calculated by selecting the maximum value in the rectangular region formed by transcription start site (TSS) coordinates (Additional 1 bin was expanded, -1 if upstream gene and +1 if downstream gene) and super-enhancer coordinates on the raw Hi-C contact map. As the genomic distance of breakpoints is zero after rearrangement, SE-promoter distance was defined by the summation of the promoter-breakpoint distance and SE-breakpoint distance. A profile of expected interaction frequencies over genomic distance was generated from the 20 kb resolution pan-normal Hi-C contact map; the genomic distance-interaction frequency value of all *cis*-pan-normal Hi-C contact map were averaged per each distance (20 kb bin) and normalized by the maximum interaction frequency value. The SE-promoter pairs showing higher interaction frequency (over 2-fold) than the expected value at a given genomic distance after rearrangement (SE-promoter distance) were selected as significant interaction pairs.

Clonality determination of *de novo* chromatin contacts

The clonality (or mutation time) of each *de novo* chromatin contacts was determined based on the presence of clonal or subclonal variants on the Hi-C reads. The clonal or subclonal of WGS-identified variants were determined by using MutationTimeR[21] (ver. 1.00.0). Variant call format (VCF) file (trimmed for MutationTimeR input) and a file with position, strand, copy number (major and minor), and purity information (computational purity) were given as an input with ‘n.boot’ number of 10.

As a number of Hi-C interaction reads inversely proportional to the distance, only reads related to the SV event tend to highly enrich on the *de novo* chromatin contacts of large-scale intra-chromosomal SV (breakpoint distance >1 Mb) or translocations (distance undefined). This property enables us to select Hi-C reads only from the SV harboring clones among all mapped Hi-C reads obtained from the heterogeneous condition.

To determine the clonality of *de novo* chromatin contacts, we first filtered *de novo* chromatin contacts with less than 3 Hi-C read coverage at all clonality-resolved variant loci (‘non-testable’). If more than 60% of the subclonal variants are supported by at least two Hi-C reads, the *de novo* chromatin contacts were classified as ‘subclonal’. If the Hi-C reads only contain clonal variants, the *de novo* chromatin contacts were classified as a ‘clonal’ event. The *de novo* chromatin contacts supported by both clonal and subclonal variants harboring Hi-C reads were classified as ‘undetermined’.

RISK score metric calculation

A RISK score metric[23] was used to evaluate the prognostic effect of the gene sets based on their expression level. The univariate Cox regression between each gene’s expression level in the given gene set and progression-free survival (PFS) was conducted. The statistically significant genes (p value or FDR cutoff used) were selected as a RISK gene set. After this process, the coefficient value (β_i) for each gene was obtained. The RISK score of each patient was defined as the sum of the products of the coefficients (β_i) and expression levels (x_i) of RISK genes (total *n* genes). Patients with scores higher $(\geq)$ than the average RISK score were defined as a high-RISK score group and the remains were defined as a low-RISK score group.

$$RISK\ score\ =\ \sum_{i}^{n} \beta_i * x_i$$

Survival analysis based on RISK score

A total of 214 patients’ gene expression level table, recurrence, and progression free survival time were obtained from the CoPM consortium. To avoid the potential overfitting issue of the Cox regression, the following procedures were applied: 1) 600 times of bootstrapping (sampling without replacement) was applied to the CoPM dataset, 2) 2-fold cross validation was performed in which the coefficient was calculated from the half of the resampled dataset and the final RISK score was calculated by applying the coefficients to the expression level of the other half, and 3) all valid protein coding genes (n = 18,221) were used for the regression. By repeating this process for 100 times, genes robustly included in the RISK gene set (false discovery rate, FDR <0.01 for each round) over 50 times were selected (termed ‘frequently included genes’). The representative coefficient value (β_i) of the frequently included genes was determined by the average of coefficient values obtained during 100 times of 2-fold cross validations. R ‘survival’ (ver. 2.41.3) and ‘survminer’ (ver. 0.4.1) packages were used to compute p values and plot Kaplan-Meier curves.

Hazard ratio analysis

A total of 96 CoPM patients’ microsatellite instability and driver gene mutation (*APC*, *KRAS*, and *TP53*) information were obtained from the whole-genome sequencing data. A total of 5 gene sets of published expression level-based prognostic markers for CRC were also used.[24]

Marker 1: *MCTP1*, *LAMA3*, *CTSC*, *PYROXD1*, *EDEM1*, *IL2RB*, *ZNF697*, *SLC6A11*, *IL2RA*, *CYFIP2*, *PIM3*, *LIF*, *PLIN3*, *HSD3B1*, *ZBED4*, *PPARA*, *THNSL2*, and *CA438802*.

Marker 2: *BGN*, *MKI67*, *MYBL2*, *GADD45B*, *FAP*, *INHBA*, *c-MYC*, *ATP5E*, *GPX1*, *PGK1*, *UBB*, and *VDAC2.*

Marker 3: *BMI1*, *ETV6*, *H3F3B*, *RPS10*, and *VEGFA.*

Marker 4: *PIGR*, *CXCL13*, *MMP3*, *TUBA1B*, *SESN1*, *AZGP1*, *KLK6*, *EPHA7*, *SEMA3A*, *DSC3*, *CXCL10*, *ENPP3*, and *BNIP3.*

Marker 5: *OLFM4*, *CXCL9*, *DMBT1*, *UGT2B17*, *SEMA3A*, *NT5E*, and *WNT11*.

The status of each expression level-based prognostic marker (Marker 1 to 5) and clonal RISK genes (clonal SE-hijacking genes found in the frequently included genes) for each patient was labeled using the RISK score metric (higher than average RISK: 1 and lower than average RISK score: 0). Due to the small number of genes, the threshold of Cox regression significance was adjusted to p value <0.05. After labeling, the hazard ratio of the clonal RISK genes compared to the other known markers was calculated. If all genes had non-significant p values, the gene set was excluded from the HR analysis.

For validation purpose, the TCGA dataset (n = 603) was used to measure HR. Only expression level-based prognostic markers were used with the TCGA dataset. Through HR analysis, it was calculated whether clonal SE-hijacking genes have higher HR compared to other factors in TCGA as well as CoPM. R ‘survival’ and ‘survminer’ packages were used to compute the hazard ratio and plot the result.

Experimental validation of the *TOP2B* gene overexpression by genomic deletion

In order to validate whether a large-scale genomic deletion (4.2 Mb) nearby *TOP2B* gene induces *TOP2B* gene overexpression by SE-hijacking (identified as SE-promoter pairs with significant interaction frequencies), we cloned two guide RNAs (gRNAs) to target the downstream of *TOP2B* (gRNA: CTGTATAGTACCCATGCACA) and the upstream of corresponding super-enhancer loci (gRNA: AGCGAATGACTGACCACCAT) into lentiCRISPRv2 vector (Addgene, 52961). The near-diploid HCT116 cells, obtained from KCLB, were infected with pairs of lentiCRISPR vectors targeting *TOP2B* and super-enhancer loci by virus using ViraPower Lentiviral Packaging Mix (Thermo Fisher, K497500) as described previously.[58] The HCT116 clones infected with green fluorescent protein (GFP)-targeting lentiCRISPR vector served as a control (gRNA: GGGCGAGGAGCTGTTCACCG). Transduced cells were selected in 1 μg/mL puromycin (Sigma-Aldrich, P4512) for 7 days. To make single clones, bulk cultures were plated into 96-well plates. Several single colonies were isolated and independently expanded. After approximately 30 days of clonal expansion, genomic DNA was extracted using QIAamp DNA Micro Kit (QIAGEN, 56304). Large-scale genomic deletions induced by CRISPR-Cas9 system were identified by PCR and direct Sanger sequencing (F1/R1 primer: CCCGGCCCGCAATTTTATAC/GGACTGCTCGGAGGCTTTAA and F2/R2 primer: ACACAGACAGGGCAGGTATC/TCAGATGAGAAGCCGGACTC). For the validation of RNA expression level change, *TOP2B*, *THRB*, and *SATB1* mRNA levels were analyzed by quantitative Real-Time PCR (qRT–PCR) and normalized relative to 18S ribosomal RNA in control GFP-targeting vector treated clone (*TOP2B* mRNA-F/R primer: AAGCACAAGAAAAGGCAGCA/CTCGCCCTTTTG CATCTCTC, *THRB* mRNA-F/R primer: AGTCATGTGCCCATTCCTGA/TTTGCTTGCCCACCATTCTC, and *SATB1* mRNA-F/R primer: TCAGTGGAAGCCTTGGGAAT/TTGTCCTTCAGTTTGCCGTG). More than six independent RNA preparations were performed.

Cell proliferation analysis

Cell proliferation was monitored using Fluorescence-activated cell sorting (FACS) analysis. 1 × 10^5 cells were cultured for 3 days and harvested with trypsin. Cells were fixed with cold 70% ethanol, and stored at −20°C for over 24 h. The cells were then washed in PBS and incubated with 10 μg/mL RNase A (Sigma-Aldrich, R6148) at 37°C for 20 min. Next, the cells were stained with 20 μg/mL propidium iodide (Sigma-Aldrich, P4864). DNA contents were quantified using a FACSCanto II Flow Cytometer (BD Biosciences, ver. 3.0). It was confirmed that CRISPR-induced *TOP2B* SE-hijacking did not affect cell proliferation and apoptosis.

RNA-seq of CRISPR-Cas9 treated cells for cross validation

The expression level of three genes (*TOP2B*, *SATB1*, and *THRB*) measured by qRT-PCR was cross validated by the RNA-seq. The D134 clone was selected and compared with the GFP-targeting vector treated control clone. The RNA was extracted by using NucleoSpin RNA XS kit (Macherey-Nagel, MN740902.50). The quality and quantity of the RNA were measured with 4200 TapeStation (Agilent Technologies). TruSeq Stranded mRNA Library Prep (Illumina, 20020594) was used for cDNA synthesis. High-throughput sequencing was performed by 75 bp paired-end sequencing using NextSeq 550 (Illumina). The quality of obtained sequencing results was confirmed by fastqc software and ERCC input correlation as described. The HISAT2-stringtie pipeline mentioned at the previous section was applied for the processing.

Western blot analysis

The whole cell lysate was obtained by incubating cells on ice for 30 min in lysis buffer (50 mM Tris-HCl, 150 mM NaCl, 1% NP-40, 0.1% Na-deoxycholate, 50 mM NaF, 1 mM sodium pyrophosphate, 1 mM EDTA, and protease/phosphatase inhibitors, pH 7.5), followed by centrifugation at 13,000 rpm at 4°C for 15 min. The supernatant was collected and then protein concentrations were quantified with Pierce BCA Protein Assay Kit (Thermo Fisher, 23225). Equal amounts of proteins were size fractionated by 10% SDS-PAGE and transferred to Amersham Protran NC Membranes (Cytiva, 10600002). The membranes were blocked with 1X TBS-T (20 mM Tris-HCl, 150 mM NaCl, 0.1% Tween 20, pH 8.0) containing 5% skim milk (BD Biosciences, 232100) with agitation. Then, the membranes were immunoblotted overnight with TOP2B (Abcam, ab72334) and β-Actin (Cell Signaling Technology, 4976) specific primary antibodies (diluted in 1X TBS-T containing 5% skim milk) at 4°C. The next day, the membranes were washed with 1X TBS-T for 5 times, incubated with Goat anti-Rabbit IgG (H + L) Secondary Antibody (HRP conjugate, Thermo Fisher, 31460) for 1 h at room temperature, followed with 5 times washing. The proteins were then detected using a Pierce ECL Western Blotting Substrate (Thermo Fisher, 32106) and visualized with X-ray film (AGFA, CP-BU).

Growth inhibition assay

The viability of cells was assessed using MTT assays. A total of 3×10^3 cells were seeded in 96-well plates, incubated for 24 h, and treated for 72 h with drugs (oxaliplatin: Selleckchem, S1224, Irinotecan: Selleckchem, S1198, SN-38: Selleckchem, S4908, and Etoposide: Selleckchem, S1225) at 37°C. Following treatment, MTT solution (Sigma-Aldrich, M2128) was added to each well and incubated for 4 h at 37°C. The medium was then removed, and dimethyl sulfoxide (Sigma-Aldrich, D8418) was added and mixed thoroughly for 30 min at room temperature. Cell viability was determined by measuring absorbance at 540 nm using a Multiskan SkyHigh Microplate Spectrophotometer (Thermo Fisher, 51119600) with SkanIt Microplate Reader Software (ver. 5.0). The concentration of drug required to inhibit cell growth by 50% was determined via interpolation from dose-response curves using Calcusyn software (Biosoft, ver. 5.0). Six replicate wells were utilized for each analysis, and at least three independent experiments were conducted.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical significance was calculated using R software. Two-sided Kolmogorov-Smirnov (KS) test was used to calculate the statistical significance between the two groups. In the case of comparing patient survival, we used the Kaplan-Meier (log rank) test. Statistical significance in gene set enrichment was measured by performing a permutation using randomly selected gene sets as the expectation. Hypergeometric test was used for functional enrichment test, implemented in Metascape. For all figures, asterisks denote statistical significance and the associated p values are shown in the legend.