HYENA detects non-coding genes activated by distal enhancers in cancer

Ali E. Yesilkanal [1, *], Anqi Yu [1, *], Ashish Thakur [2], Fan Wang [1], Xiaoyang Wu [1,3], Alexander Muir [1,3], Xin He [2], Francois Spitz [2], Lixing Yang [1, 2, 3,**]

1. Ben May Department for Cancer Research, University of Chicago, Chicago IL, USA
2. Department of Human Genetics, University of Chicago, Chicago IL, USA
3. University of Chicago Comprehensive Cancer Center, Chicago, IL, USA

* these authors contributed equally

** correspondence: lixingyang@uchicago.edu

Abstract

Somatic structural variations (SVs) in cancer can shuffle DNA content in the genome, relocate regulatory elements, and alter genome organization. Enhancer hijacking occurs when SVs relocate distal enhancers to activate gene expression. However, most enhancer hijacking studies have only focused on protein-coding genes. Here, we develop a computational algorithm "HYENA" to identify candidate oncogenes (both protein-coding and non-coding) activated by enhancer hijacking based on tumor whole-genome and transcriptome sequencing data. HYENA detects genes whose elevated expression is associated with somatic SVs by using a rank-based regression model. We systematically analyze 1,148 tumors across 25 types of adult tumors and identify a total of 192 candidate oncogenes including many non-coding genes. A long non-coding RNA *TOB1-AS1* is activated by various types of SVs in 10% of pancreatic cancers through altered 3-dimension genome structure. We find that high expression of *TOB1-AS1* can promote cell invasion and metastasis. Our study highlights the contribution of genetic alterations in non-coding regions to tumorigenesis and tumor progression.

Introduction

At mega-base-pair scale, linear DNA is organized into topologically associating domains (TADs) [1], and gene expression is regulated by DNA and protein interactions governed by 3D genome organization. Enhancer-promoter interactions are mostly confined within TADs [2–4]. Structural variants (SVs), including deletions, duplications, inversions and translocations, can dramatically change the TAD organization and gene regulation [5], and subsequently contribute to tumorigenesis. Previously, we discovered that *TERT* is frequently activated in chromophobe renal cell carcinoma by the relocation of distal enhancers [6], a mechanism referred to as enhancer hijacking (**Fig. 1a**). In fact, many oncogenes, such as *BCL2* [7], *MYC* [8], *TAL1* [9], *MECOM/EVI1* [10], *GFI1* [11], *IGF2* [12], *PRDM6* [13], and *CHD4* [14], are activated through this mechanism. These examples demonstrate that genomic architecture plays an important role in pathogenesis. However, the vast majority of the known enhancer hijacking target genes are protein-coding genes, and few non-coding genes have been reported to promote diseases through enhancer hijacking. Non-coding genes are known to play important roles in many biological processes [15] and some are known to drive tumorigenesis [16,17]. In this study, we will focus on identifying oncogenic non-coding genes activated by enhancer hijacking.

Several existing algorithms can detect enhancer hijacking target genes based on patient cohorts, such as CESAM [12] and PANGEA [14]. These two algorithms perform linear regression and elastic net model (also based on linear regression) to associate elevated gene expression with nearby SVs, respectively. PANGEA also considers the effects of somatic single nucleotide variants (SNVs) on gene expression. However, a major drawback of these algorithms is that linear regression is quite sensitive to outliers. Outliers are very common in gene expression data from cancer samples and can seriously impair the performances of these algorithms. In addition, CESAM is optimized for microarray data, while PANGEA depends on annotation of tissue-specific promoter-enhancer pairs which are not readily available for many tumor types. Cis-X [18] and NeoLoopFinder [19] can detect enhancer hijacking target genes based on individual samples. However, these tools have limitations in detectable genes and input data. Cis-X detects *cis*-activated genes based on allele-specific expression, which requires the genes to carry heterozygous SNVs. NeoLoopFinder takes Hi-C, Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET), or similar data measuring chromatin interactions as input. Furthermore, to identify oncogenes, recurrent events in patient cohorts will still be required. Finally, no non-coding genes have been reported as enhancer hijacking targets by the above algorithms. A recent study on SVs altering gene expression in Pan-Cancer Analysis of Whole Genomes (PCAWG) samples [20] only considered protein-coding genes but not non-coding genes.

Here, we developed Hijacking of Enhancer Activity (HYENA) using normal-score regression and permutation test to detect candidate enhancer hijacking genes (both protein-coding and non-coding genes) based on tumor whole-genome and transcriptome sequencing data from patient cohorts. HYENA identified a non-coding gene, *TOB1-AS1*, as a regulator of cancer cell invasion in vitro and metastasis in vivo.

Results

HYENA workflow

Conceptually, the SVs leading to elevated gene expression are expression quantitative trait loci (eQTLs). The variants are SVs instead of commonly used germline single nucleotide polymorphisms (SNPs) in eQTL analysis. With somatic SVs and gene expression measured from the same tumors through whole-genome sequencing (WGS) and RNA sequencing (RNAseq), we can identify enhancer hijacking target genes by eQTL analysis. However, the complexities of cancer and SVs pose many challenges. For instance, there is tremendous inter-tumor heterogeneity—no two tumors are identical at the molecular level. In addition, there is substantial intra-tumor heterogeneity as tumor tissues are always mixtures of tumor, stromal and immune cells. Moreover, genome instability is a hallmark of cancer, and gene dosages are frequently altered [21]. Furthermore, gene expression networks are widely rewired [22], and outliers of gene expression are common.

Here, we developed an algorithm HYENA to overcome the challenges described above (see more details in Methods Section). We used a gene-centric approach to search for elevated expression of genes correlated with the presence of nearby SVs (**Fig. 1b**). Although duplicated enhancers can upregulate genes [23,24], we do not consider these as enhancer hijacking events since no neo-promoter-enhancer interactions are established. However, small deletions can remove TAD boundaries or repressive elements and lead to neo-promoter-enhancer interactions (**Fig. 1a**). Therefore, small tandem duplications were discarded, and small deletions were retained. For each gene, we annotated SV status (presence or absence of nearby SVs) for all samples. Samples in which the testing genes were highly amplified were discarded since many of these genes are amplified by circular extrachromosomal DNA (ecDNA) [25] and ecDNA can promote accessible chromatin [26] with enhancer rewiring [27]. Only genes with nearby SVs in at least 5% of tumors were further considered. In contrast to CESAM and PANGEA, we did not use linear regression to model the relationships between SV status and gene expression because linear regression is sensitive to outliers and many false significant associations would be detected [28]. Instead, we used a rank-based normal-score regression approach. After quantile normalization of gene expression for both protein-coding and non-coding genes, we added small Gaussian noises to gene expression for tie breaking, ranked the genes based on quantile-normalized noise-added expression, and transformed the ranks to the quantiles of standard normal distribution. We used the z scores (normal scores) of the quantiles as dependent variables in regression. In the normal-score regression model, tumor purity, copy number of the testing gene, patient age, and sex were included as covariates since these factors confound gene expression. We also included gene expression principal components (PCs) that were not correlated with SV status to model unexplained variations in gene expression. To deduce a better null distribution, we permuted the gene expression 100 times and ran the same regression models. All *P* values from the permutations were pooled together and used as the null distribution to calculate empirical *P* values. Then, multiple test corrections were performed on one-sided *P* values since we are only interested in elevated gene expression under the influence of nearby SVs. Finally, genes were

discarded if their elevated expression could be explained by germline eQTLs. The remaining genes were candidate enhancer hijacking target genes.

Benchmarking performances

There is no gold standard available to comprehensively evaluate the performance of HYENA. Here, we used eight known enhancer hijacking target genes including *MYC* [8], *BCL2* [7], *CCNE1* [29], *TERT* [6], *IGF2* [12,29], *IGF2BP3* [30] and *IRS4* [12] from six types of adult tumors profiled by PCAWG [20] to assess the sensitivity of HYENA (**Supplementary Table S1**). The six tumor types were malignant lymphoma (MALY), stomach/gastric cancer (STAD), chromophobe renal cell carcinoma (KICH), colorectal cancer (COAD/READ), thyroid cancer (THCA) and lung squamous cell carcinoma (LUSC). We also expect immunoglobulin genes to be detected as enhancer hijacking candidates in malignant lymphoma due to V(D)J recombination since the lymphomas in PCAWG are B-cell derived Burkitt lymphomas [31]. In B cells, V(D)J recombination occurs to join different variable (V), joining (J) and constant (C) segments to produce antibodies with a wide range of antigen recognition ability. Therefore, certain segments have elevated expression, and the recombination events can be detected as somatic SVs. In addition to HYENA, we also tested two other algorithms—CESAM and PANGEA. All three algorithms were run on the same somatic SVs and gene expression data from the PCAWG cohort. Note that PANGEA depends on promoter-enhancer interactions predicted from cell lines which were not available for thyroid tissue. Therefore, thyroid cancer data were not analyzed by PANGEA.

Out of the eight known genes, HYENA detected five, CESAM detected three and PANGEA did not detect any (**Fig. 2a**). In the five tumor types analyzed by all three algorithms, HYENA identified a total of 25 candidate genes, CESAM identified 19, whereas PANGEA identified 255 genes (**Fig. 2b, Supplementary Tables S2, S3 and S4**). Six genes were detected by both HYENA and CESAM. However, PANGEA had little overlap with the other algorithms (**Fig. 2b**). Both *IGF2* and *IRS4* were initially discovered by CESAM as enhancer hijacking target genes using copy number variation (CNV) breakpoints profiled by microarray with much larger sample sizes (378 colorectal cancers and 497 lung squamous cell carcinomas) [12]. In PCAWG, the sample sizes with both WGS and RNAseq were smaller (51 colorectal cancers and 47 lung squamous cell carcinomas). HYENA detected *IGF2* in colorectal cancer but not *IRS4*, whereas CESAM and PANGEA detected neither. In stomach/gastric cancer, *IGF2* and *CCNE1* were identified as enhancer hijacking genes in a cohort of 208 samples [29]. Neither of these genes were detected by any of the algorithms because there were only 29 stomach tumors in PCAWG. Therefore, known genes missed by HYENA were likely due to small sample size. Of the 16 genes detected by HYENA in malignant lymphoma, there were two immunoglobulin light chain genes from lambda cluster (*IGLC7* and *IGLJ7*) (**Supplementary Table S2**). CESAM detected 11 genes with one being immunoglobulin gene (*IGLC7*) (**Supplementary Table S3**). In contrast, PANGEA detected 30 candidate genes, but none were immunoglobulin genes (**Supplementary Table S4**). In summary, HYENA had the best sensitivity of the three algorithms.

To evaluate specificity of the algorithms, we generated 20 random datasets by shuffling gene expression data in both malignant lymphoma and breast cancer and ran all three algorithms

again. Since these gene expression data were random, there should be no associations between SVs and gene expression, and all genes detected should be false positives. In malignant lymphoma with observed gene expression, HYENA, CESAM and PANGEA detected 16, 11 and 30 candidate genes respectively (**Supplementary Tables S2, S3 and S4**). In the 20 random gene expression datasets, HYENA detected an average of 0.5 genes per dataset (**Fig. 2c**), and CESAM detected 0.5 genes, whereas PANGEA detected 40 genes (**Supplementary Fig. S1**). In breast cancer with observed gene expression, HYENA, CESAM and PANGEA detected 61, 9 and 2,309 candidate genes, respectively (**Supplementary Tables S2, S3 and S4**). In 20 random gene expression datasets, HYENA, CESAM and PANGEA detected 0.35, 0.9 and 2,296 genes on average (**Fig. 2c** and **Supplementary Fig. S1**). In both tumor types, the numbers of false positives called by PANGEA in random datasets were comparable to the numbers of genes detected with observed gene expression (**Supplementary Fig. S1**). In summary, HYENA predicted the least number of false positives among the three algorithms.

Overall, HYENA has superior sensitivity and specificity in the detection of enhancer hijacking candidate genes.

Enhancer hijacking candidate genes in PCAWG

We used HYENA to analyze a total of 1,146 tumors across 25 tumor types in PCAWG with both WGS and RNAseq data. When each tumor type was analyzed individually, we identified 192 candidate enhancer hijacking genes in total (**Supplementary Tables S1** and **S2**), five of which were known enhancer hijacking targets (**Fig. 3a**). *TERT* was the only gene identified in two tumor types/cohorts (chromophobe renal cell carcinoma [KICH] from the US and renal cell carcinoma [RECA] from Europe). Almost all genes were only detected in one tumor type, highlighting high tumor type specificity of the findings. The number of genes detected in each tumor type also differed dramatically (**Fig. 3b**). No genes were detected in bladder cancer (BLCA), cervical cancers (CESC), glioblastoma multiforme (GBM), and low-grade gliomas (LGG) probably due to their small sample sizes. Breast cancers (BRCA) had the greatest number of candidate genes likely due to the large sample size as well as the abundance of SVs resulting from homologous recombination deficiency (HRD) [32]. Although ovarian cancers (OV) also suffer from HRD and their sample size was comparable with breast cancers, there were much fewer enhancer hijacking genes detected. Thyroid cancer (THCA) genomes were among the most stable genomes in PCAWG [33]. However, the 15 enhancer hijacking genes identified in thyroid cancers exceeded the number of candidate genes in ovarian cancer as well as many other tumor types. Among these 15 genes, *IGF2BP3* was a known oncogene activated by enhancer hijacking [30,34]. Interestingly, there were two liver cancer cohorts with comparable sample sizes—LIHC from the US and LIRI from Japan. A total of 18 genes were identified in the US cohort whereas no genes were found in the Japanese cohort. One possible reason for such a drastic difference could be that hepatitis B virus (HBV) infection is more common in liver cancers in Japan [35], and virus integration in the tumor genome can result in oncogene activation [36]. In Chronic Lymphocytic Leukemia (CLLE), a total of nine genes were detected, and seven were immunoglobulin genes from both lambda and kappa clusters. Given that sample size and genome instability can only explain a small fraction of variations of enhancer hijacking genes detected in

different tumor types, the landscape of enhancer hijacking in cancer seems to be mainly driven by the underlying disease biology. Intriguingly, out of the 192 candidate genes, 73 (38%) were non-coding genes including long non-coding RNAs (lncRNAs) and microRNAs (**Fig. 3b**).

Neo-TADs formed through somatic SVs

Next, we focus on the most frequently altered enhancer hijacking candidate non-coding gene in pancreatic cancer *TOB1-AS1* (**Fig. 4a**), a long non-coding RNA (lncRNA). Seven (9.6%) out of 74 tumors had a variety of somatic SVs near *TOB1-AS1* including translocations, deletions, inversions and tandem duplications (**Fig. 4b** and **Supplementary Table S5**). For example, tumor 9ebac79d-8b38-4469-837e-b834725fe6d5 had a translocation between chromosomes 17 and 19 (**Fig. 4c**). The breakpoints were upstream of *TOB1-AS1* and upstream of *UQCRFS1* (**Fig. 4d**). In tumor 748d3ff3-8699-4519-8e0f-26b6a0581bff, there was a 19.3 Mb deletion which brought *TOB1-AS1* next to a region downstream of *KCNJ2* (**Fig. 4c** and **4e**).

We used Hi-C maps to investigate the genome organization of the loci in their normal context and Akita [37], a convolutional neural network that predicts 3D genome organization, to estimate the 3D architecture of the loci impacted by SVs. While 3D structures are dynamic and may change with cell-type and gene activity, TAD boundaries are often more stable and remain similar between different cell-types [1]. TAD boundaries are defined locally by the presence of binding sites for CCCTC-binding factor (CTCF), a ubiquitously expressed DNA-binding protein [1,38], and their formation arises from the stalling of the cohesin-extruded chromatin loop by CTCF at these positions [39]. For this reason, one can reliably expect that upon chromosomal rearrangement, new TADs will be formed as directed by the position of relocated TAD boundaries. This assumption has been validated by direct experimental evidences examining the "neo-TADs" associated with SVs at different loci [40–42]. The wildtype *TOB1-AS1* locus had a TAD between a CTCF binding site in *RSAD1* and another one upstream of *SPAG9* (**Fig. 4d** and **Supplementary Fig. S2**). There were TADs spanning *UQCRFS1* and downstream of *KCNJ2* in the two partner regions (**Fig. 4d**, **4e** and **Supplementary Fig. S2**). In tumor 9ebac79d-8b38-4469-837e-b834725fe6d5, the translocation was predicted to lead to a neo-TAD resulting from merging the TADs of *TOB1-AS1* and *UQCRFS1* into one (**Fig. 4d**). In tumor 748d3ff3-8699-4519-8e0f-26b6a0581bff, another neo-TAD was predicted to form as a result of the deletion that merged the TADs of *TOB1-AS1* and the downstream portion of *KCNJ2* into one (**Fig. 4e**). Within these predicted neo-TADs, Akita predicted new strong interactions involving several CTCF binding sites across two SV partner regions (**Fig. 4d** and **4e** pink dots in the right panels). Similar neo-TADs could be observed in two additional tumors (**Supplementary Fig. S3**). In the two tumors with 317 kb and 226 kb tandem duplications, the *TOB1-AS1* TAD was expanded (**Supplementary Fig. S4a**). However, in the last tumor, *TOB1-AS1* was predicted to remain in its original TAD after translocation (**Supplementary Fig. S4b**). In summary, at least four out of the seven tumors harboring somatic SVs near *TOB1-AS1* were predicted to result in neo-TADs including *TOB1-AS1*.

These results demonstrated that HYENA can detect genes activated by reorganization of 3D genome architecture.

Oncogenic functions of *TOB1-AS1*

TOB1-AS1 has been reported as a tumor suppressor in several tumor types [44,45]. However, it was predicted to be an oncogene by HYENA in pancreatic cancers. To test if *TOB1-AS1* has oncogenic functions in pancreatic cancer, we performed both in vitro and in vivo experiments. We surveyed pancreatic cancer cell line RNAseq data from Cancer Cell Line Encyclopedia (CCLE) and identified that the commonly transcribed isoform of *TOB1-AS1* was ENST00000416263.3 in pancreatic cancers (**Supplementary Fig. S5**). The synthesized *TOB1-AS1* cDNA was cloned and overexpressed in two pancreatic cancer cell lines, PANC-1 and PATU-8988T, both of which had low expression of *TOB1-AS1* (**Supplementary Fig. S6**). In both cell lines, overexpression of *TOB1-AS1* promoted in vitro cell invasion (**Fig. 4f**). In addition, three weeks after tail vein injection, PANC-1 cells with *TOB1-AS1* overexpression caused higher metastatic burden in immunodeficient mice than the control cells (**Fig. 4g**). Our results showed that the lncRNA *TOB1-AS1* is oncogenic and has pro-metastatic function in pancreatic cancer, and HYENA is able to detect novel proto-oncogenes activated by distal enhancers.

Discussion

Here, we report a computational algorithm HYENA to detect candidate oncogenes activated by distal enhancers via somatic SVs. These SV breakpoints fell in the non-coding regions of the genome and caused shuffling of regulatory elements and altered gene expression. The candidate genes we detected were not limited to protein-coding genes but also included non-coding genes. Our in vitro and in vivo experiments showed that a lncRNA identified by HYENA, *TOB1-AS1*, was a potent oncogene in pancreatic cancers.

HYENA detects candidate genes based on patient cohorts rather than individual samples. Genes need to be recurrently rearranged in the cohort to be detectable. Therefore, sample size is a major limiting factor. Of the eight ground truth cases, HYENA only detected five (**Fig. 2a**). The undetected ones were likely due to small sample size. HYENA aims to identify oncogenes recurrently activated by somatic SVs since these events are under positive selection. However, genes detected in individual tumors by tools such as cis-X and NeoLoopFinder may not be oncogenes. To identify candidate oncogenes among them, recurrence would still be required.

The candidate genes identified by HYENA have statistically significant associations between nearby somatic SVs and elevated expression. However, the relationship may not be causal. It is possible that the presence of SVs and gene expression are unrelated, but both are associated with another factor. We modeled other factors to the best of our ability including gene dosage, tumor purity, patient gender, patient age and principal components of gene expression. In addition, it is also possible that the high gene expression caused somatic SVs. Open chromatin and unwounded double helix regions during transcription are prone to double-strand DNA breaks which may produce somatic SVs. Therefore, it is possible that some of the candidate genes are not oncogenes. Functional studies are required to determine the disease relevance of the candidate genes.

Note that the predicted 3D genome organization is not cell-type-specific. Akita was trained on five high quality Hi-C and Micro-C datasets (HFF, H1hESC, GM12878, IMR90 and HCT116) [37] and predicts limited cell-type-specific differences. Therefore, the predicted TADs reflect conserved 3D genome structure in the five cell types (foreskin fibroblast, embryonic stem cell, B-lymphocyte, lung fibroblast and colon cancer). There were minor differences between HFF and H1hESC (**Supplementary Fig. S2**) in genome organization. For example, the left boundary of the TAD at the *UQCRFS1* locus was different between HFF and H1hESC (**Supplementary Fig. S2a**). Nonetheless, the translocation between chromosomes 17 and 19 removed the left boundary and the right side of the *UQCRFS1* TAD merged with the *TOB1-AS1* TAD (**Fig. 4d**). Therefore, the cell-type difference likely does not have major impact on our results.

Methods

Datasets

This study used data generated by the Pan-Cancer Analysis of Whole Genomes (PCAWG). We limited our study to a total of 1,146 tumor samples for which both whole-genome sequencing (WGS) and RNAseq data were available. The data set was composed of cancers from 25 tumor types including 23 bladder urothelial cancers (BLCA), 88 breast cancers (BRCA), 20 cervical squamous cell carcinomas (CESC), 68 chronic lymphocytic leukemias (CLLE), 51 colorectal cancers (COAD/READ), 20 glioblastoma multiforme (GBM), 42 head and neck squamous cell carcinomas (HNSC), 43 chromophobe renal cell carcinomas (KICH), 37 renal clear cell carcinomas from United States (KIRC), 31 renal papillary cell carcinomas (KIRP), 18 low-grade gliomas (LGG), 51 liver cancers from United States (LIHC), 67 liver cancers from Japan (LIRI), 37 lung adenocarcinomas (LUAD), 47 lung squamous cell carcinomas (LUSC), 95 malignant lymphomas (MALY), 80 ovarian cancers (OV), 74 pancreatic cancers (PACA), 19 prostate adenocarcinomas (PRAD), 49 renal clear cell carcinomas from European Union/France (RECA), 34 sarcomas (SARC), 34 skin cutaneous melanomas (SKCM), 29 stomach adenocarcinomas (STAD), 47 thyroid cancers (THCA), and 42 uterine corpus endometrial carcinomas (UCEC). More detailed information on the sample distribution and annotation can be found in **Supplementary Table S1**.

WGS and RNA-seq data analysis of tumor and normal samples were performed by the PCAWG consortium as previously described [20]. Somatic and germline SNVs, somatic CNVs, SVs and tumor purity were detected by multiple algorithms and consensus calls were made. Genome coordinates were based on hg19 reference genome and GENCODE v19 was used for gene annotation. Gene expression was quantified by HT-Seq (version 0.6.1p1) as fragments per kilobase of million mapped (FPKM). Clinical data such as donor age and sex were downloaded from the PCAWG data portal (https://dcc.icgc.org/pcawg).

Significant eQTL-gene pairs (v8) were downloaded from the Genotype-Tissue Expression (GTEx) data portal (https://gtexportal.org/home/datasets). Only those eQTLs that had a hg19 liftover variant ID were included in the analysis and hg38 variants with no corresponding hg19 annotation were discarded.

HYENA algorithm

First, small tandem duplications (<10kb) were discarded since they are unlikely to produce new promoter-enhancer interactions. The remaining SVs were mapped to the flanking regions (500kb upstream and downstream of transcription start sites [TSSs]) of annotated genes. SVs that fall entirely within a gene body were also discarded. The SV status of each gene was defined by the presence or absence of SV breakpoints within the gene or its flanking regions for each tumor. The binary variable SV status was used in the normal-score regression model below. Only genes with at least 5% samples carrying SVs were tested. For each gene, samples with that gene highly amplified (>10 copies) were removed from the regression model.

Gene expression normal scores

Gene expression quantifications (fragments per kilobase per million [FPKM]) were quantile normalized (FPKM-QN) using *quantile.normalize()* function from the *preprocessCore* R package to enhance cross-sample comparison. To break the ties for genes with identical FPKM-QN values in multiple samples (especially those caused by FPKM of zero) during ranking, very small Gaussian noises were added to all the FPKM-QN values in all samples by *add.Gaussian.noise(mat, mean = 0.000000001, stddev = 0.000000001, symm = F)* from the *RMThreshold* R package. Since the mean and standard deviation of the noises added were small, the rankings of the non-identical values were not affected. For each gene, samples were ranked based on their noised-added expression values, the ranks were mapped to a standard normal distribution and the corresponding z scores were gene expression normal scores. Normal score conversion forced the expression data into a Gaussian distribution, allowing for parametric comparisons between samples.

Normal-score regression

Generalized linear model was used to test associations between gene expression normal scores and SV status and control for confounding variables such as gene copy number, tumor sample purity, donor age and sex. To capture unobserved variations in gene expression, the first n principal components (PCs) of the expression data were also included in the regression model, where n was determined as 10% of sample size of the cohort and up to 20 if the sample size was more than 200. The regression model was shown as below:

$$\text{Expression_normal_score} \sim \text{sv_status} + \text{copy_number} + \text{purity} + \text{age} + \text{sex} + PC_1 + PC_2 \ldots + PC_n$$

For each gene, all PCs were tested for associations with the SV status of that gene and those PCs that significantly correlate (Mann-Whitney test, $P<0.05$) with SV status were not used in regression.

Calculating empirical p values and model selection

Gene expression data were permuted 100 times by randomly shuffling expression values within the cohort. The normal-score regression was performed the same way on observed gene expression and permuted expression. *P* values for SV status from permuted expression were pooled as a null distribution. Then the *P* values for SV status from observed expression and the *P*-value null distribution were used to calculate empirical *P* values. One-sided *P* values were used since we were only interested in elevated gene expression. False discovery rates (FDRs) were calculated using Benjamini-Hochberg procedure. Genes with FDR less than 0.1 were considered candidate genes.

The above empirical *P* value calculation and candidate gene detection were performed iteratively with no PC and up to n PCs in the regression model. When different numbers of PCs were included in the model, the numbers of candidate genes varied. The regression model with the least number of PCs reaching 80% of maximum number of candidate genes in all regression models tested was selected as final model to avoid over fitting. For example, the sample size for PCAWG BRCA was 88, therefore, we tested from 0 to 9 PCs. Among these, the model including

8 PCs gave the highest number (82) of candidate genes. Therefore, the model including 7 PCs with 68 candidate genes was selected as final model since it had the lowest number of PCs reaching 80% of 82 candidate genes (**Supplementary Table S6**).

Testing eQTL-SV associations

Known germline eQTLs from the matching tissues were obtained from GTEx (**Supplementary Table S7**). The associations between germline genotypes of eQTLs and SV status of the 213 candidate genes in PCAWG cohort were tested. Genes with significant correlations ($P<0.05$) between their SV status and at least one eQTL were removed. The remaining genes were our final enhancer-hijacking candidate genes.

Benchmarking

Known enhancer hijacking target genes in PCAWG tumor types were selected to test the sensitivity of HYENA, CESAM and PANGEA. The genes included *MYC* in malignant lymphoma, *BCL2* in malignant lymphoma, *CCNE1* in stomach/gastric cancer, *TERT* in chromophobe renal carcinoma, *IGF2* in colorectal cancer, *IGF2* in stomach/gastric cancer, *IGF2BP3* in thyroid cancer, and *IRS4* in lung squamous cell carcinoma. The same SVs, CNVs, and SNVs were used as input for all three algorithms. For CESAM and PANGEA, upper-quantile normalized fragments per kilobase per million (FPKM-UQ) were normalized by tumor purity and gene copy number, and then used as gene expression inputs. CESAM was run using default parameters, and FDR of 0.1 was used to select significant genes. PANGEA requires predicted enhancer-promoter (EP) interactions based on ChIP-seq and RNA-seq data. The EP interactions were downloaded from EnhancerAtlas 2.0 (http://www.enhanceratlas.org/) (**Supplementary Table S8**). EP interactions from multiple cell lines of the same type were merged. PANGEA was run with default parameters as well and significant genes were provided by PANGEA (multiple testing adjusted *P* value <0.05). To test false positives for HYENA, CESAM and PANGEA, 20 random gene expression datasets for malignant lymphoma and breast cancer were generated by randomly shuffling sample IDs in gene expression data. HYENA, CESAM and PANGEA were run with random expression in the same way as above.

Predicting 3D genome organization

A 1 Mb sequence was extracted from the reference genome centered at each somatic SV breakpoint and was used as input for Akita [37] to predict the 3D genome organization. Two 500 kb sequences were merged according to the SV orientation to construct the sequence of rearranged genome. Akita was used to predict the genome organization for the rearranged sequence. High-resolution Micro-C data obtained from human H1-ESCs and HFF cells [43] were used to facilitate TAD annotation together with predicted genome organization. H3K27Ac and CTCF ChIP-seq data from PANC-1 cell line were downloaded from ENCODE data portal (https://www.encodeproject.org/).

Cell lines

HEK293T, PANC-1, and PATU-8988T cells were obtained from Dr. Alexander Muir (University of Chicago). All cell lines were cultured at 37°C/5% CO_2. HEK293T cells and PANC-1 cells were cultured in Dulbecco's Modified Eagle Medium (DMEM) (Gibco, 21041025) containing 10% fetal bovine serum (FBS) (Gibco, A4766) as per ATCC (American Type Culture Collection, USA) instructions (https://www.atcc.org/products/crl-3216, https://www.atcc.org/products/crl-1469). PATU-8988T cells were cultured with DMEM containing 5% FBS, 5% horse serum (Gibco, 26050088), and 2 mM L-glutamine as recommended by DSMZ (Deutsche Sammlung von Mikroorganismen and Zellkulturen, Germany) (https://www.dsmz.de/collection/catalogue/details/culture/ACC-162). All cell lines have been regularly monitored and tested negative for mycoplasma using mycoplasma detection kit (Lonza, LT07-218).

TOB1-AS1 and luciferase overexpression

A 1,351 bp *TOB1-AS1* cDNA (ENST00000416263.3) was synthesized by GenScript (New Jersey, USA) and subcloned into the lentiviral pCDH-CMV-MCS-EF1-Puro plasmid (SBI, CD510B-1). The cDNA sequence in the plasmid was verified by Sanger sequencing at University of Chicago Medicine Comprehensive Cancer Center core facility. The *TOB1-AS1* overexpression plasmid was amplified by transforming Stellar™ Competent Cells (Takara, 636763) with the plasmid as per instructions and isolated by QIAGEN HiSpeed Plasmid Midi Kit (QIAGEN, 12643). LucOS-Blast vector was obtained from Dr. Yuxuan Phoenix Miao (University of Chicago), cloned, and amplified as described above.

HEK293T cells were plated in T-25 flasks and grown to 75% confluence prior to transfection. For each T-25 flask, 240µl Opti-MEM (Gibco, 31985070), 1.6µg pCMV-VSV-G, 2.56µg pMDLg/pRRE, 2.56µg pRSV-Rev, 3.4µg *TOB1-AS1* overexpression vector and 22.8µl TransIT-LT1 Transfection Reagent (Mirus, MIR 2306) were mixed and incubated at room temperature for 30 minutes, then added to the plated HEK293T cells with fresh medium. The luciferase vector was packaged into lentivirus with the same method. Upon 48 hours of incubation, lentiviral supernatant was collected, filtered through 0.45-µmpolyvinylidene difluoride filter (Millipore), and mixed with 8µg/ml polybrene. PANC-1 or PATU-8988T cells at 60% confluence were transduced with the lentiviral supernatant for 48 hours followed by three rounds of antibiotic selection with 4µg/ml puromycin for *TOB1-AS1* overexpression and 10µg/ml blasticidin for the luciferase expression. *TOB1-AS1* expression was validated by quantitative reverse transcription polymerase chain reaction (qRT-PCR), and luciferase expression was validated by in vitro bioluminescence imaging in black wall 96-well plates (Corning, 3603). D-luciferin potassium salt (Goldbio, LUCK-100) solution with 0, 1.25, 2.5, 5 and 10µl 15mg/ml was added into the wells in a serial dilution way, and imaging was obtained after 5 minutes.

Finally, *TOB1-AS1* overexpression or empty pCDH transduced cell lines with luciferase co-expression were built for both PATU-8988T and PANC-1 cells.

RNA isolation and qRT-PCR

Cells were plated in 6-well plates and allowed to reach 80% confluence prior to RNA extraction. After cells being lysed in 300µl/well TRYzol™ (Invitrogen, 15596026), RNA samples were prepared following the manual of Direct-zol RNA Miniprep kit (RPI, ZR2052). Reverse transcription was performed using Applied Biosystems High-Capacity cDNA Reverse Transcription Kit (43-688-14) following manufacturer's instructions. Quantitative PCR (qPCR) was conducted on StepOnePlus Real-Time PCR System (Applied Biosystems, 4376600), using PowerUp SYBR Green Master Mix (A25742) following the manufacturer's instructions with a primer concentration of 300nM in 10µl reaction systems. Primers were ordered from Integrated DNA Technologies. Primer sequences used in this study are as follows:
TOB1 forward: 5' -GGCACTGGTATCCTG AAA AGCC- 3'
TOB1 reverse: 5' – GTGGCAGATTGCCACGAACATC- 3'
TOB1-AS1 forward: 5' -GGAGTGGTCAGGTGACTGATT- 3'
TOB1-AS1 reverse: 5' -ATTCCACTCCTGTTTGCAACT- 3'
GAPDH forward: 5' – ACCACAGTCCATGCCATCAC- 3'
GAPDH reverse: 5' -TCCACCACCCTGTTGCTGTA- 3'
Relative expression levels for *TOB1-AS1* and *TOB1* were calculated by the $2^{(-\Delta\Delta C_T)}$ method based on *GAPDH* expression.

Transwell assay for cell invasion in vitro

Transparent PET membrane culture inserts of 24-well plate (Falcon, 353097) were coated with Cultrex Reduced Growth Factor Basement Membrane Extract (BME) (R&D Systems, 3533-010-02) at 50µg per membrane (200µl of 0.25mg/ml BME stock per membrane) at 37°C for an hour. A total of 100,000 PANC-1 cells/well and 50,000 PATU-8988T cells/well were resuspended in serum-free, phenol-red free DMEM media and seeded into the coated inserts. Phenol red free DMEM of 500µl (Gibco, A1443001) with 10% FBS was added to the bottom of the wells and the cells were allowed to invade for 16 hours. Additional wells with 500µl serum free, phenol red free DMEM media without FBS in the bottom chamber were seeded with the same number of cells as indicated above as negative control. At the end of the assay, the membranes were stained with 500µl 4µg/ml Calcein AM (CaAM) (Corning, 354216) for an hour at 37°C. The cells that failed to invade were removed from the top chamber with a cotton swab and all inserts were transferred into 1x Cell Dissociation Solution (Bio-Techne, 3455-05-03) and shaken at 150rpm for an hour at 37°C. Finally, CaAM signal from the invaded cells was measured by a plate reader (Perkin Elmer Victor X3) at 465/535nm.

Tumor metastasis in vivo

All animal experiments for this study were approved by the University of Chicago Institutional Animal Care and Use Committee (IACUC) prior to execution. Male NSG mice were ordered from the Jackson Laboratory (strain#005557). Mice were injected intravenously through the tail vein with luciferase-expressing cells at 250,000 cells/mouse for PATU-8988T and at 400,000 cells/mouse for PANC-1 cells. Metastatic burden was measured weekly for 4 weeks via bioluminescence imaging using Xenogen IVIS 200 Imaging System (PerkinElmer) at the University of Chicago Integrated Small Animal Imaging Research Resource (iSAIRR) Facility. Each mouse was weighed and injected intra-peritoneally with D-luciferin solution at a concentration of 150μg/g of body weight, 14 minutes prior to image scanning ventral side up.

Code availability

The HYENA package is available at https://github.com/yanglab-computationalgenomics/HYENA.

Acknowledgements

We thank the Center for Research Informatics at the University of Chicago for providing the computing infrastructure, Matthew Stephens for helpful suggestions, Marsha Rosner for assistance in lentiviral experiments, and Ani Solanki for assistance in animal experiments. The work was supported by the Goldblatt Endowment (A.Y.), the National Institutes of Health grant K22CA193848 (L.Y.), R01CA269977 (L.Y.) and University of Chicago and UChicago Comprehensive Cancer Center (L.Y.).

Disclosure

The authors have no competing interests to declare.

Figure Legends

Figure 1. Outline of enhancer hijacking and HYENA algorithm. a, Mechanisms of gene activation by SVs. SVs can activate genes by recruiting distal active enhancers (top panel), by removing TAD boundaries and forming de novo enhancer-promoter interactions (middle panel), or by removing repressive elements that block gene expression (bottom panel). **b**, HYENA workflow. Green and purple boxes denote input and output files, respectively. Orange boxes denote intermediate steps. Numbers in parentheses represent default values of HYENA.

Figure 2. Benchmarking HYENA. a, Comparison of HYENA, CESAM, and PANGEA in detecting oncogenes known to be activated by enhancer hijacking in six tumor types from the PCAWG cohort. **b**, UPSET plot demonstrating candidate genes identified and shared among the three tools in five tumor types of PCAWG. Individual dots denote genes detected by one tool, and dots connected by lines denote genes detected by multiple tools. **c**, Number of genes detected by HYENA in two PCAWG tumor types using observed gene expression and randomized expression. Genes detected in random expression datasets are false positives.

Figure 3. Enhancer hijacking candidate genes in PCAWG. a, Candidate genes detected by HYENA in individual tumor types of PCAWG. Genes that are plotted twice (e.g. *TERT*) are detected in more than one cancer types. Genes labelled as red are known enhancer hijacking targets. **b**, Diverse types of candidate genes identified by HYENA in PCAWG. Numbers in parentheses denote sample size in the corresponding tumor types.

Figure 4. *TOB1-AS1* activated by various types of SVs in pancreatic cancer. a, Normalized expression of the *TOB1-AS1* in samples with (n=7) and without nearby SVs (n=67) in pancreatic cancers. Boxplot shows median values (thick black lines), upper and lower quartiles (boxes), 1.5× interquartile range (whiskers) and outliers (black dots). **b**, Circos plot summarizing intrachromosomal SVs (blue, n=5) and translocations (red, n=3) near *TOB1-AS1*. **c**, Diagrams depicting putative enhancer hijacking mechanisms that activate *TOB1-AS1* in one tumor with a 17:19 translocation (left panel) and another tumor with a large deletion (right panel). **d**, Predicted 3D chromatin interaction maps of the *TOB1-AS1* (left panel), *UQCRFS1* (middle panel), and the translocated region in tumor 9ebac79d-8b38-4469-837e-b834725fe6d5 (right panel). The downstream fragment of the chromosome 19 SV breakpoint was flipped in orientation and linked to chromosome 17. H3K27Ac and CTCF ChIP-seq data of PANC-1 cell line are shown at the bottom. **e**, Predicted 3D chromatin interaction maps of *TOB1-AS1* (left panel) and *KCNJ2* (middle panel) loci without deletion as well as the region after deletion in tumor 748d3ff3-8699-4519-8e0f-26b6a0581bff (right panel). **f**, *TOB1-AS1* overexpression in PATU-8988T (4 biological replicates) and PANC-1 (3 biological replicates) promoted in vitro cell invasion using Transwell assay. Each biological replicate was an independent experiment with 7 technical replicates per experimental group. The average fold change of cell invasion was calculated after the background invasion measured in the absence of any chemotactic agent being subtracted from each technical replicate. *P* values were calculated by two-sided student t test. **g**, *TOB1-AS1* overexpression in PANC-1 cells promoted in vivo tumor metastasis. Images of radiance in immunodeficient mice are shown on the left while the quantifications of radiance are shown on the right. Eight mice were used in both overexpression group and empty vector control. The images were analyzed by setting the regions of interest (ROIs) to mouse torsos and measuring

the average radiance level (in p/sec/cm^2/sr). Two-side student t test was used. *, **, *** represent P<.05, <.01, and <.001, respectively. Error bars in **f** and **g** indicate standard error of the mean.

References

1. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).

2. Krijger, P. H. L. & De Laat, W. Regulation of disease-associated gene expression in the 3D genome. *Nat. Rev. Mol. Cell Biol.* **17**, 771–782 (2016).

3. Symmons, O. *et al.* Functional and topological characteristics of mammalian regulatory domains. *Genome Res.* **24**, 390–400 (2014).

4. Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).

5. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).

6. Davis, C. F. *et al.* The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell* **26**, 319–330 (2014).

7. Bakhshi, A. *et al.* Cloning the chromosomal breakpoint of t(14;18) human lymphomas: clustering around Jh on chromosome 14 and near a transcriptional unit on 18. *Cell* **41**, 899–906 (1985).

8. Gostissa, M. *et al.* Long-range oncogenic activation of Igh-c-myc translocations by the Igh 3′ regulatory region. *Nature* **462**, 803–807 (2009).

9. Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458 (2016).

10. Gröschel, S. *et al.* A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell* **157**, 369–381 (2014).

11. Northcott, P. A. *et al.* Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428–434 (2014).

12. Weischenfeldt, J. *et al.* Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nat. Genet.* **49**, 65–74 (2016).

13. Northcott, P. A. *et al.* The whole-genome landscape of medulloblastoma subtypes. *Nature* **547**, 311–317 (2017).

14. He, B. *et al.* Diverse noncoding mutations contribute to deregulation of cis-regulatory landscape in pediatric cancers. *Sci. Adv.* **6**, eaba3064 (2020).

15. Kopp, F. & Mendell, J. T. Functional Classification and Experimental Dissection of Long Noncoding RNAs. *Cell* **172**, 393–407 (2018).

16. Lin, C.-P. & He, L. Noncoding RNAs in Cancer Development. *Annu. Rev. Cancer Biol.* **1**, 163–184 (2017).

17. Liu, S. J., Dang, H. X., Lim, D. A., Feng, F. Y. & Maher, C. A. Long noncoding RNAs in cancer metastasis. *Nat. Rev. Cancer 2021 217* **21**, 446–460 (2021).

18. Liu, Y. *et al*. Discovery of regulatory noncoding variants in individual cancer genomes by using cis-X. *Nat. Genet.* **52**, 811–818 (2020).

19. Wang, X. *et al.* Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes. *Nat. Methods* **18**, 661–668 (2021).

20. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

21. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).

22. Uhlen, M. *et al.* A pathology atlas of the human cancer transcriptome. *Science* **357**, eaan2507 (2017).

23. Zhang, X. *et al.* Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. *Nat. Genet.* **48**, 176–182 (2015).

24. Takeda, D. Y. *et al.* A Somatically Acquired Enhancer of the Androgen Receptor Is a Noncoding Driver in Advanced Prostate Cancer. *Cell* **174**, 422-432.e13 (2018).

25. Turner, K. M. *et al.* Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature* **543**, 122–125 (2017).

26. Wu, S. *et al.* Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature* **575**, 699–703 (2019).

27. Morton, A. R. *et al.* Functional Enhancers Shape Extrachromosomal Oncogene Amplifications. *Cell* **179**, 1330-1341.e13 (2019).

28. Li, Y., Ge, X., Peng, F., Li, W. & Li, J. J. Exaggerated false positives by popular differential expression methods when analyzing human population samples. *Genome Biol.* **23**, 79 (2022).

29. Ooi, W. F. *et al.* Integrated paired-end enhancer profiling and whole-genome sequencing reveals recurrent CCNE1 and IGF2 enhancer hijacking in primary gastric adenocarcinoma. *Gut* **69**, 1039–1052 (2020).

30. Yun, J. W. *et al.* Dysregulation of cancer genes by recurrent intergenic fusions. *Genome Biol.* **21**, 166 (2020).

31. Richter, J. *et al.* Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. *Nat. Genet.* **44**, 1316–1320 (2012).

32. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).

33. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).

34. Panebianco, F. *et al.* THADA fusion is a mechanism of IGF2BP3 activation and IGF1R signaling in thyroid cancer. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 2307–2312 (2017).

35. Zapatka, M. *et al.* The landscape of viral associations in human cancers. *Nat. Genet. 2020*

523 **52**, 320–330 (2020).

36. Neuveut, C., Wei, Y. & Buendia, M. A. Mechanisms of HBV-related hepatocarcinogenesis. *J. Hepatol.* **52**, 594–604 (2010).

37. Fudenberg, G., Kelley, D. R. & Pollard, K. S. Predicting 3D genome folding from DNA sequence with Akita. *Nat. Methods* **17**, 1111–1117 (2020).

38. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680 (2014).

39. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* **15**, 2038–2049 (2016).

40. Franke, M. *et al.* Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016).

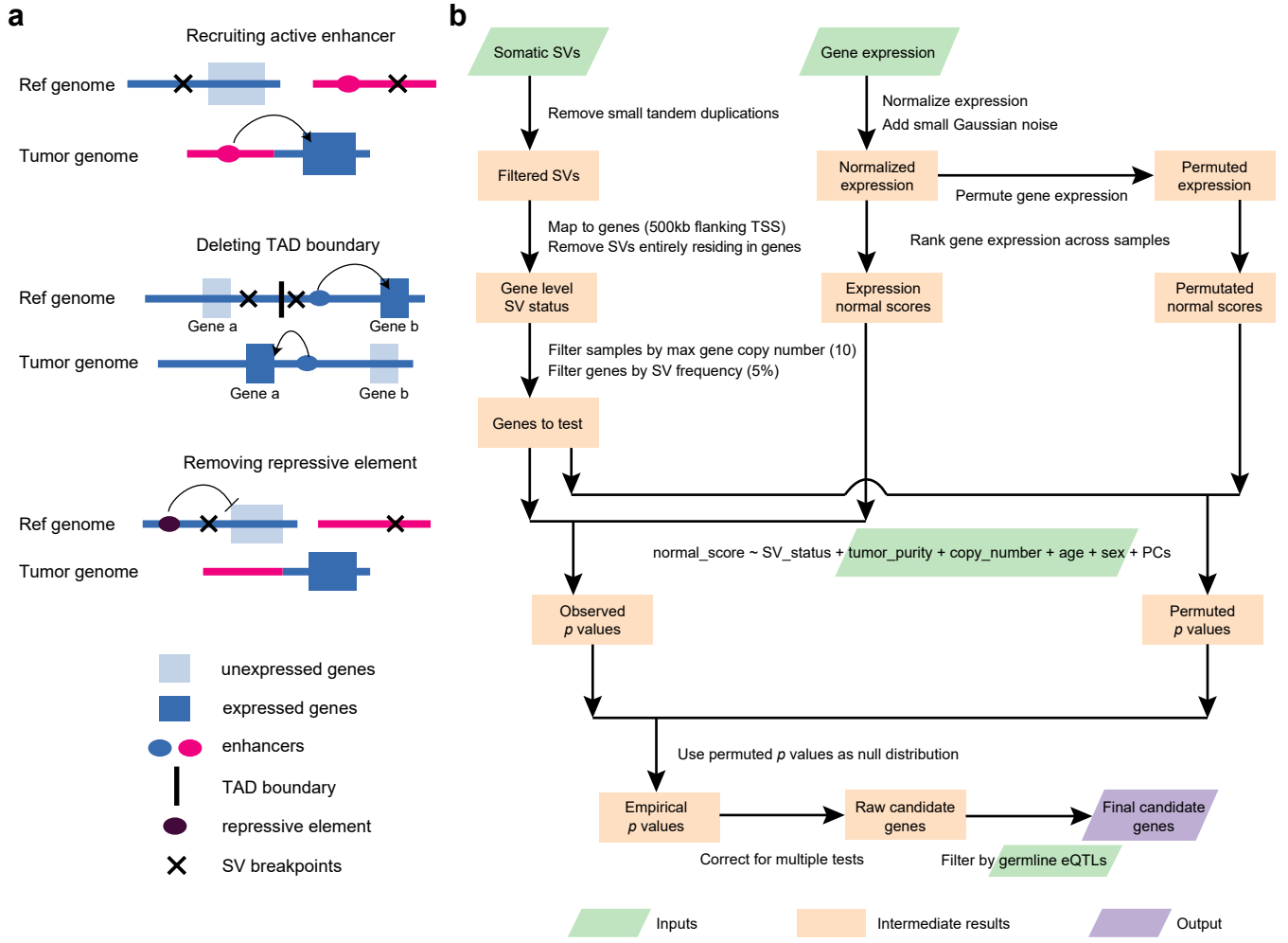
41. Melo, U. S. *et al.* Complete lung agenesis caused by complex genomic rearrangements with neo-TAD formation at the SHH locus. *Hum. Genet.* **140**, 1459–1469 (2021).

42. de Bruijn, S. E. *et al.* Structural Variants Create New Topological-Associated Domains and Ectopic Retinal Enhancer-Gene Contact in Dominant Retinitis Pigmentosa. *Am. J. Hum. Genet.* **107**, 802–814 (2020).

43. Krietenstein, N. *et al.* Ultrastructural Details of Mammalian Chromosome Architecture. *Mol. Cell* **78**, 554-565.e7 (2020).

44. Yao, J. *et al.* Long noncoding RNA TOB1-AS1, an epigenetically silenced gene, functioned as a novel tumor suppressor by sponging miR-27b in cervical cancer. *Am. J. Cancer Res.* **8**, 1483 (2018).

45. Shangguan, W. *et al.* TOB1-AS1 suppresses non-small cell lung cancer cell migration and invasion through a ceRNA network. *Exp. Ther. Med.* **18**, (2019).

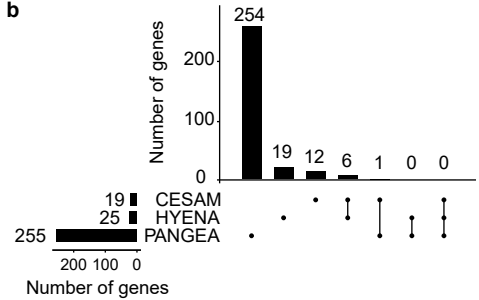

a
Recruiting active enhancer
Ref genome
Tumor genome
Deleting TAD boundary
Ref genome
Gene a
Gene b
Tumor genome
Gene a
Gene b
Removing repressive element
Ref genome
Tumor genome
unexpressed genes
expressed genes
enhancers
TAD boundary
repressive element
SV breakpoints
b
Somatic SVs
Gene expression
Remove small tandem duplications
Normalize expression
Add small Gaussian noise
Filtered SVs
Normalized expression
Permuted expression
Permute gene expression
Map to genes (500kb flanking TSS)
Remove SVs entirely residing in genes
Rank gene expression across samples
Gene level SV status
Expression normal scores
Permutated normal scores
Filter samples by max gene copy number (10)
Filter genes by SV frequency (5%)
Genes to test
normal_score ~ SV_status + tumor_purity + copy_number + age + sex + PCs
Observed p values
Permuted p values
Use permuted p values as null distribution
Empirical p values
Raw candidate genes
Final candidate genes
Correct for multiple tests
Filter by germline eQTLs
Inputs
Intermediate results
Output

a

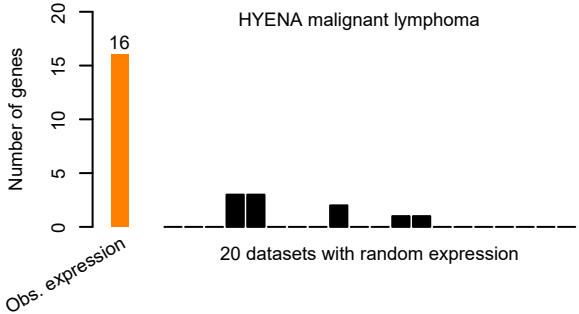
		HYENA	CESAM	PANGEA
MYC	malignant lymphoma	●	●	–
BCL2	malignant lymphoma	●	●	–
CCNE1	stomach/gastric cancer	–	–	–
TERT	chromophobe renal cell carcinoma	●	●	–
IGF2	colorectal cancer	●	–	–
IGF2	gastric/stomach cancer	–	–	–
IGF2BP3	thyroid cancer	●	–	NT
IRS4	lung squamous cell carcinoma	–	–	–

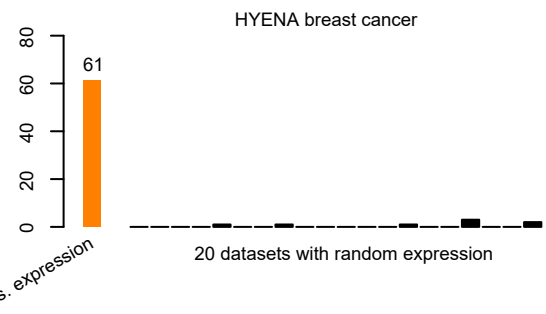
● detected – undetected NT not tested

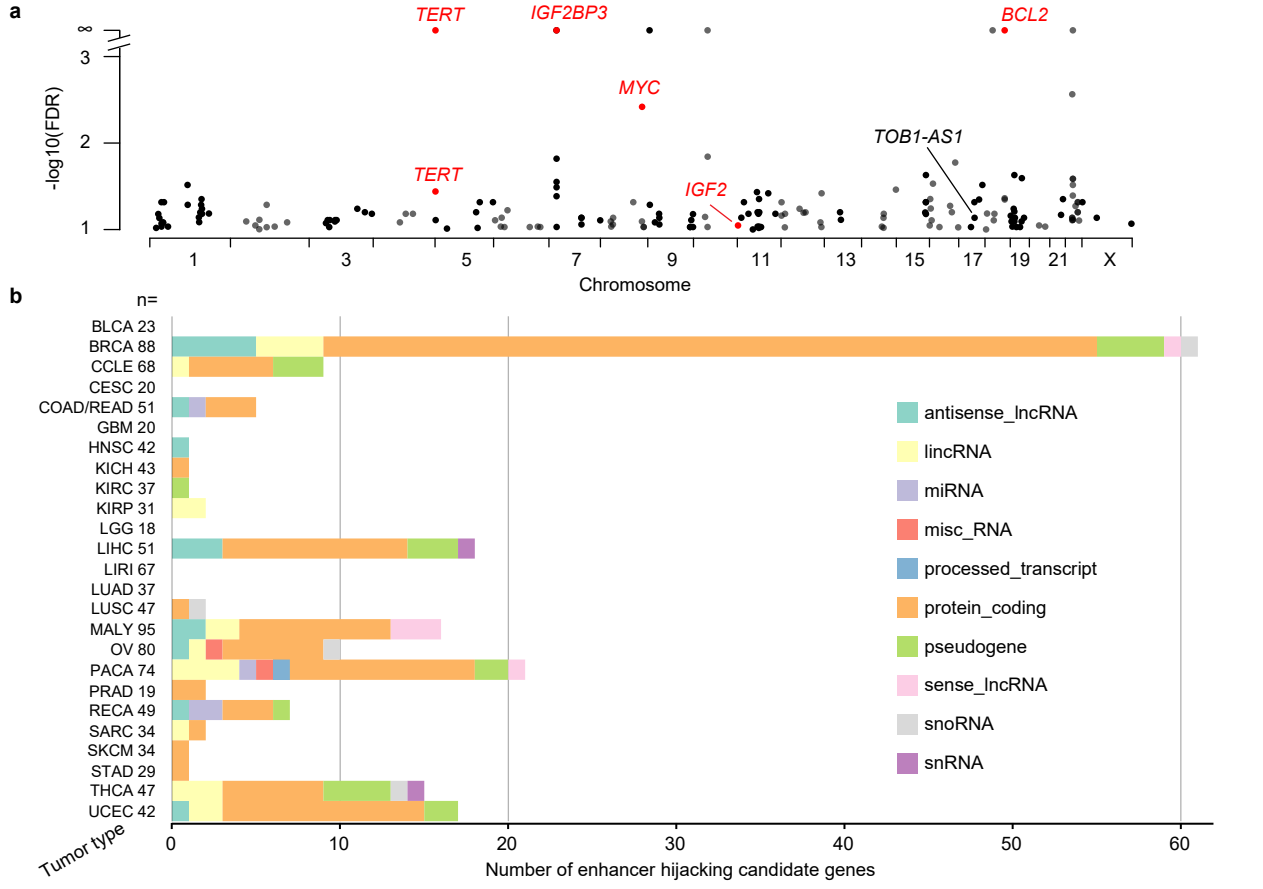
b

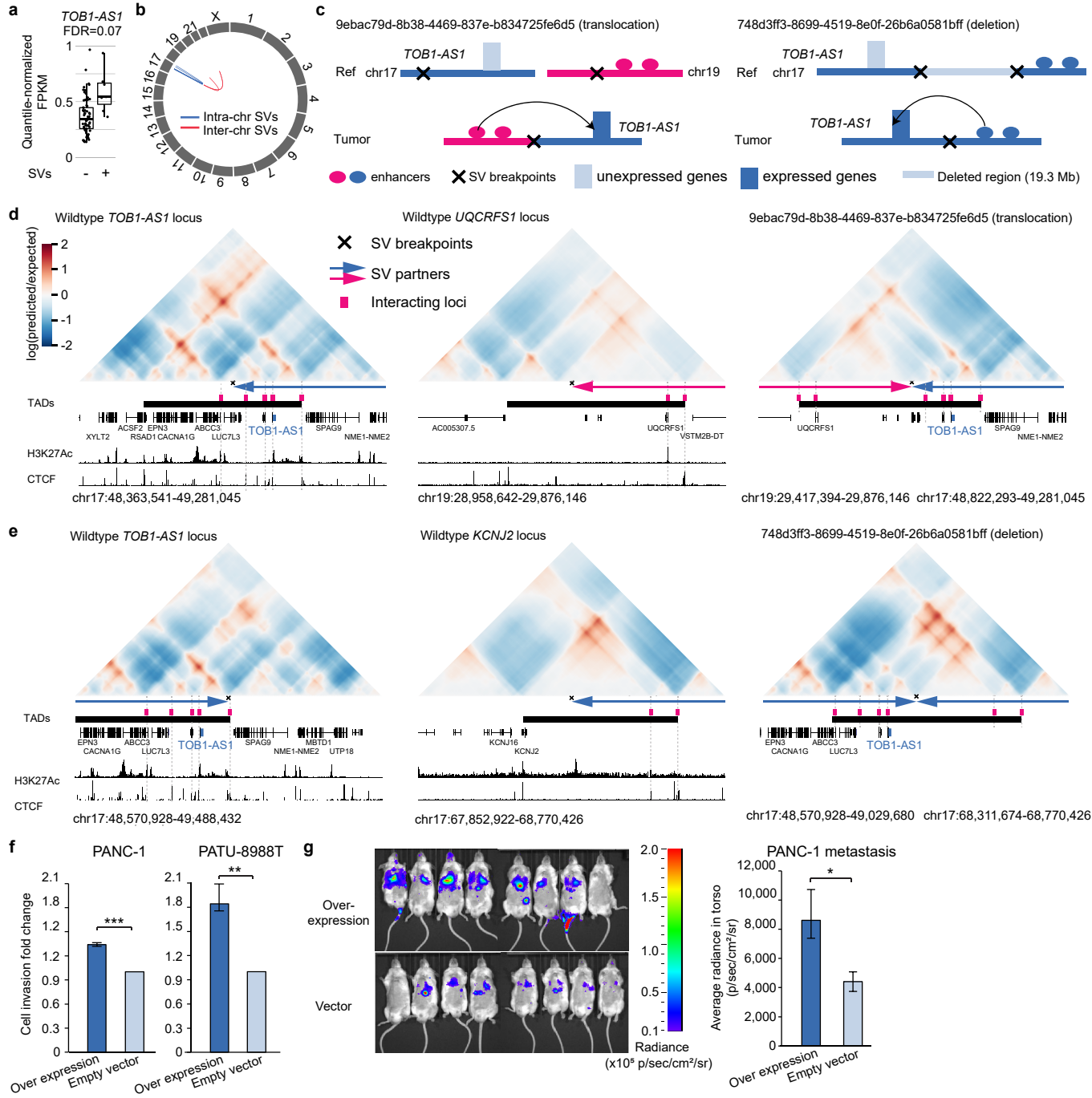


c






a
TERT
IGF2BP3
BCL2
MYC
TOB1-AS1
TERT
IGF2
-log10(FDR)
∞
3
2
1
1
3
5
7
9
11
13
15
17
19
21
X
Chromosome
b
n=
BLCA 23
BRCA 88
CCLE 68
CESC 20
COAD/READ 51
GBM 20
HNSC 42
KICH 43
KIRC 37
KIRP 31
LGG 18
LIHC 51
LIRI 67
LUAD 37
LUSC 47
MALY 95
OV 80
PACA 74
PRAD 19
RECA 49
SARC 34
SKCM 34
STAD 29
THCA 47
UCEC 42
Tumor type
antisense_lncRNA
lincRNA
miRNA
misc_RNA
processed_transcript
protein_coding
pseudogene
sense_lncRNA
snoRNA
snRNA
0
10
20
30
40
50
60
Number of enhancer hijacking candidate genes


a
TOB1-AS1
FDR=0.07
Quantile-normalized FPKM
SVs
b
Intra-chr SVs
Inter-chr SVs
c
9ebac79d-8b38-4469-837e-b834725fe6d5 (translocation)
748d3ff3-8699-4519-8e0f-26b6a0581bff (deletion)
Ref
Tumor
chr17
chr19
TOB1-AS1
enhancers
SV breakpoints
unexpressed genes
expressed genes
Deleted region (19.3 Mb)
d
Wildtype TOB1-AS1 locus
Wildtype UQCRFS1 locus
9ebac79d-8b38-4469-837e-b834725fe6d5 (translocation)
log(predicted/expected)
SV breakpoints
SV partners
Interacting loci
TADs
H3K27Ac
CTCF
chr17:48,363,541-49,281,045
chr19:28,958,642-29,876,146
chr19:29,417,394-29,876,146 chr17:48,822,293-49,281,045
e
Wildtype TOB1-AS1 locus
Wildtype KCNJ2 locus
748d3ff3-8699-4519-8e0f-26b6a0581bff (deletion)
chr17:48,570,928-49,488,432
chr17:67,852,922-68,770,426
chr17:48,570,928-49,029,680 chr17:68,311,674-68,770,426
f
PANC-1
PATU-8988T
Cell invasion fold change
Over expression
Empty vector
g
Over-expression
Vector
Radiance (x10⁵ p/sec/cm²/sr)
PANC-1 metastasis
Average radiance in torso (p/sec/cm²/sr)