THE UNIVERSITY OF CHICAGO

IDENTIFICATION OF NOVEL ONCOGENES ACTIVATED BY ENHANCER HIJACKING

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

COMMITTEE ON CANCER BIOLOGY

BY

ANQI YU

CHICAGO, ILLINOIS

DECEMBER 2024

This dissertation is dedicated to Yingzhen Pei.

Table of Contents

List of Figures viii

List of Tables x

Acknowledgements xi

Abstract xvi

Introduction 1

Cancer and oncogenes 1

Genomic instability and structural variations 3

Enhancers and their chromatin features 7

3D genome organization and gene expression regulation 9

Enhancer hijacking and the approaches to infer enhancer hijacking events 13

Oncogene activation in neuroblastoma 16

Questions remaining to be addressed 19

HYENA Detects Oncogenes Activated by Distal Enhancers in Cancer 21

Introduction 21

Materials and Methods 23

Datasets 23

HYENA algorithm 25

Benchmarking 28

Predicting 3D genome organization 29

In situ Hi-C and ATAC-Seq 29

Cell lines 30

TOB1-AS1 and luciferase overexpression 31

TOB1-AS1 transient knock-down using antisense oligonucleotides (ASOs) 32
RNA isolation and qRT-PCR 32
Transwell assay for cell invasion in vitro 33
Tumor metastasis in vivo 34
Ex vivo IVIS imaging 35
Tumor RNA sequencing and gene expression analysis 35
Code availability 36
Results 36
HYENA workflow 36
Benchmarking performances 39
Enhancer hijacking candidate genes in the PCAWG 42
Neo-TADs formed through somatic SVs 44
Oncogenic functions of TOB1-AS1 50
Discussion 54
Identifying Novel Oncogenes Detected by HYENA with CRISPR Activation Screening 59
Introduction 59
Methods 62
Cell culture 63
CRISPR activation screening 63
Sequencing data analysis 70
3D genome interaction prediction 70
Results 70
Cell proliferation screens confirmed putative oncogenes detected by HYENA 70

RCCD1 was predicted to have new enhancer-promoter interactions caused by SVs 71
Cell migration screen in two cancer cell lines revealed potential oncogenes 74
Discussion 75
Oncogenes Activated by Distal Enhancers in Neuroblastoma 78
Introduction 78
Methods 80
Datasets 80
SV calling and filtering 81
Predicting enhancer hijacking genes with HYENA 81
Results 82
SV calling for the 189 neuroblastoma samples 82
Enhancer hijacking candidates detected in GMKF neuroblastomas 84
EFTUD2 was a potential enhancer hijacking gene in high-risk neuroblastoma 92
Discussion 94
Further Discussion 97
Limitations of HYENA pipeline 98
Unaddressed questions on lncRNA TOB1-AS1 100
Future directions for studying enhancer hijacking events 101
Appendix 105
Supplementary Figures 105
Supplementary Tables 115
CRISPRa oligo library 115
MCF-7 proliferation screen significantly enriched genes with MAGeCK (D7 vs D0) 132

MCF-7 proliferation screen significantly enriched genes with MAGeCK (D14 vs D0)..... 132

MCF-7 migration screen significantly enriched genes with MAGeCK..... 133

PATU-8988T migration significantly enriched genes with MAGeCK..... 133

Sample information of the GMKF neuroblastomas included in the analysis..... 133

Reference..... 139

List of Figures

Figure 1. Diagram of enhancer hijacking events. 14

Figure 2. Outline of enhancer hijacking and HYENA algorithm. 37

Figure 3. Benchmarking HYENA. 40

Figure 4. Enhancer hijacking candidate genes in PCAWG. 43

Figure 5. TOB1-AS1 activated by various types of SVs in pancreatic cancer. 45

Figure 6. 3D genome structures in the *TOB1-AS1* locus in pancreatic cancer cell lines. 48

Figure 7. TOB1-AS1 promotes cell invasion and tumor metastasis. (Legends on next page) 51

Figure 8. Volcano plots of the enriched or depleted genes in MCF-7 proliferation screen. 71

Figure 9. RCCD1 gene expression and SVs near RCCD1. 72

Figure 10. 3D genome structures predicted by deep-learning based algorithm Orca. 73

Figure 11. Volcano plots of the genes in bottom chamber compared to top chamber in migration screens. 75

Figure 12. Landscape of 189 neuroblastoma cases. 82

Figure 13. The ratio of CNV-supported Manta SVs and the SV counts supported by SR and PR count combinations. 83

Figure 14. Gene expression levels of the five candidate genes detected in neuroblastoma. 84

Figure 15. The co-occurrence of recurrent CNVs, MYCN status, and SV status of candidate genes detected with HYENA default parameters. 86

Figure 16. MYCN expression is positively correlated with copy number, and SVs near MYCN reflected CNVs. 87

Figure 17. The co-occurrence of recurrent CNVs, MYCN status, and the SV status of selected candidate genes detected with HYENA without CNV information input and 3% recurrence rate cutoff. 90

Figure 18. EFTUD2 expression is not correlated with copy number in high-risk neuroblastoma. 93

Figure 19. Expression levels of MYCN, SLC6A18, EFTUD2, and SLC6A3 in high-risk samples. 94

Figure 20. Differentially expressed genes caused by TOB1-AS1 overexpression. 100

List of Tables

Table 1. HYENA default setting predicted oncogenes with 189 neuroblastoma samples 84

Table 2. Enhancer hijacking candidates in neuroblastoma with lower frequency requirement including high-copy genes 88

Table 3. Enhancer hijacking candidates in high-risk neuroblastoma including high-copy genes. 91

Acknowledgements

I would like to thank Dr. Lixing Yang, for his scientific insights, guidance, and advice about both research and life. For me, he is always calm and active. He is not like a mentor, but rather like a colleague. For the five years we work together, he has been teaching me how to solve scientific questions, how to face frustrations as well as unexpected issues, and to always be motivated to move to next steps. A quarter after I joined the lab, to begin some validation studies with cell lines, we built up our wet lab space together. Although we only have two benches for at most four people to work there, this moment was pretty rewarding for me, and was where my project started. I appreciate that Lixing gave me plenty of freedom and his trust so that I could learn techniques from other labs and perform experiments independently, as well as spend time on my career development processes. We met many frustrations these years, but he was always strong, keeping pursuing solutions to the questions. I will never forget our lab camping tours – although a lot of raining and mosquitoes – where we had beer and sincere conversations, together with laughter and tears. I feel honored to be his first student and work with him on open-ending questions for these years. I know we are growing together, and although there could be disagreement, he was always open and straightforward, so we always reached to conclusions that made sense to both of us. Facing stress, Lixing is never defeated, but faces challenges with rationale and a peaceful mindset. From him, I learned not only about science, but also a different attitude toward life.

I would like to thank my co-mentor, Dr. Mark Applebaum, and my thesis committee members, Dr. Oni Basu, Dr. Eileen Dolan, and Dr. François Spitz, for their constant guidance and support throughout my training process. Mark gave me a lot of constructive suggestions and insights when I was writing my thesis proposal, and as a physician and a scientist, he inspired me

by pointing out the clinical significance and how to ask important questions. Our conversations have been my fuel to keep driving forward along the research journey. Eileen, as an experienced mentor, listened to me and gave me her strong understanding and support when I had difficulties. PhD training is nothing easy. What is especially invaluable for me is that my committee members always recognize it, hear my thoughts, and share their experiences with me. During our committee meetings, they always asked key questions to my projects and proposed experiments, taught me different angles of data interpretation, and brought up new ideas. Without my committee, I could never reach where I am now.

I would like to thank my lab members and lab alumni. Dr. Ali Yesilkanal, a former postdoc who developed the pipeline of HYENA, was not only a senior scientist but also a patient teacher, a kind listener and a great artist. He gave me a lot of encouragement and understanding when we were talking about the frustrations of PhD training. Dr. Yang Yang, a postdoc in our lab, helped me a lot on coding skills and data management processes. Although we do not work on a same project, she is always open and patient to my questions. Dr. Xiaoming Zhong, a former postdoc, helped me a lot with the coding techniques and shared his programming experience with me, even after he left our lab, he never said no to our questions and requests. Will Philips, an MSTP student in the Yang lab, helped us with writing and brings in new energies into our lab.

I would like to thank Dr. Alex Muir and the Muir lab for their training during my first rotation and all the help for our project. Alex added me into the Muir lab's animal protocol, so that we could perform mouse experiments. I would also like to thank Dr. Juan J. Ápiz Saab, Grace Croley, Colin Sheehan, and Patrick Jonker. Juan trained me during my rotation with the Muir lab, and taught me a lot about cell culture, mouse experiments, and data analysis. Grace

Croley trained me about mouse work and CRISPR techniques. Also together with Colin and Alex, we discussed deeply about my CRISPR activation screen data and I really appreciate their suggestions and insights toward data interpretation and experimental design.

I would like to acknowledge many of the faculties in CCB. To Dr. Barbara Kee, who talked with me, gave her strong support when things got tough, and is always committed to building CCB into a better PhD training program. To Dr. Geoff Greene, who gave me helpful advice when I was applying for grad school, who carries the whole Ben May Department and found a lab space for us. To Dave Hosfield from the Greene Lab, who trained me how to perform molecular experiments and troubleshoot. To Dr. Marsha Rosner, who helped our lab with many useful protocols and to comply with all the lab regulation rules. To Dr. Kay Macleod, who served as the chair of CCB when I first came to UChicago in 2018, supported me when I was applying, and also who shared thoughtful insights about the functional studies of *TOB1-AS1*. To Dr. Steve Kron, who was the former chair of CCB, set high standard for our cohort, kept communicating with our cohort even during the pandemic, and initiated the reading class that was really helpful for me. To Dr. Lev Becker, who was my mentor when I worked here as a summer RA in 2018. The summer in the Becker Lab was when my story with CCB began, and was a precious experience for me where I learned how a scientific story begins with a single experiment, how to prepare presentations with clear logic flow, and the enthusiasm a scientist can possibly have.

I would also like to thank my friends from the neighbor labs and my friends outside of CCB. To Dr. Jingyun and Luan Xuejie Huang from Phoenix Miao Lab, Dr. Long Nguyen, Dr. Peter Yang, Wenchao Liu, and Maddy Henn from Marsha Rosner Lab, Dr. Lifeng Chen and Suman from Xiaoyang Wu Lab – from materials sharing to protocol troubleshooting, from

having lunch together to walking to mouse room together, from weekdays to weekends, my daily life becomes so colorful and vivid because of you. To Jingwen Xu, who dragged me out from my room and let me enjoy many beautiful days in Chicago. My special thanks to Yuqing Xue, for all of our time spending together on working-out, cooking, eating, shopping, watching movies, petting cats, and even doing nothing, which comforted me so much after all those dark moments I had in lab. To all my poker folks, Yucheng Deng, Jiaying Dong, Pinhan Chen, Ziqing Xu, Dongyue Xie, Wanrong Zhu, Yuguan Wang, Bo Yuan, Yihao Lu and Xiaoyuan Zhang, I cherish and appreciate our gathering time, when you not only inspired me with new thoughts about my future directions, but also greatly relieved me from the feeling of isolation during the pandemic.

I would like to thank my parents and my grandparents, for their unconditional love and support, for being strong enough to let me move to the US for better education, and for their understanding and encouragement when I was feeling low and lost. My family had no scientific background, but they would always listen to my explanation on research topics as well as my frustrations throughout PhD training, and always cheer me up with pure love and recognition.

I would like to thank my cats, Icee and Usagi, for their company through my PhD journey. Although they do not understand science, their warm and fluffy company and sticking around is all I ask for after long days and in cold nights.

Finally, I would like to thank Yingzhen Pei, my lifelong partner, my soulmate, my best friend, and my forever love. I appreciate that we understand each other so well and share common values toward important decisions. When I doubted myself, Yingzhen gave me confidence to love myself again. When I felt defeated, he brought me up and had my back. When I got overwhelmed, he always gave emotional supports as well as practical suggestions. He

taught me to believe in myself, do what I believe is right, and prioritize the long run. His love come from who we are and our hearts echoing, which we believe will never fade away.

Abstract

Genome instability is a hallmark of cancer, resulting in the accumulation of various types of alterations. 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 proto-oncogenes. 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 analyzed 1,146 tumors across 25 types of adult tumors and identify a total of 108 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 3D genome structure. We find that high expression of *TOB1-AS1* can promote cell invasion and metastasis. With CRISPR activation screens, we identified more potential oncogene candidates that can promote cancer cell growth or migration while confirming the known oncogenes. Applying HYENA to neuroblastoma samples, we identified 5 oncogene candidates activated by enhancer hijacking with default parameters and 58 candidates when gene copy information is excluded in the model. These genes may reveal new disease biology for neuroblastoma and potential new markers for risk level classification. In summary, our study highlights the contribution of genetic alterations in non-coding regions to tumorigenesis and tumor progression, and identified putative oncogenes activated by enhancer hijacking in multiple tumor types.

Introduction

Cancer and oncogenes

Cancer is a disease in which some cells grow uncontrollably and spread to, and eventually make damage to other parts of the body. It has been a leading cause of death for decades even after the widely used surgery, chemotherapy and radiotherapy, the invention of targeted therapies and the emergence of immunotherapies [1]. For many decades, the initiation and development of cancer have been investigated and summarized as more than ten hallmarks of cancer [2]. Among these hallmarks, six acquired capabilities (evading apoptosis, self-sufficiency in growth signals, insensitivity to anti-growth signals, sustained angiogenesis, limitless replicative potential and tissue invasion and metastasis) are driven by the activation of oncogenes and the inactivation of tumor suppressor genes [3].

Oncogenes are the genes whose activation drives cancer. The long-lasting questions of how oncogenes are activated and how their activities promote cancer have motivated generations of scientists to make groundbreaking discoveries. In 1970, the first transforming principle of Rous sarcoma virus (RSV) was physically identified [4], which started the decoding journey of molecular basis of oncogenesis. Six years later, the oncogenic *src* gene of RSV was found to be related to the cellular *src* gene in chicken [5]. This finding put oncogenes to a cellular matter, and eventually led to the discoveries of human oncogenes. In 1977 and 1979, the oncogenes now known as *MYC* and *ERBB/EGFR* were initially identified with biochemical approaches in avian acute leukemia virus genome [6, 7]. Other prominent oncogenes, like *ras*, were identified in murine tumor viruses, or *HER2*, by directly transfecting human tumor cell DNA into recipient cells, which is also considered as seminal experiments in the field [8]. Retrovirus oncogenes were just the beginning of the discoveries of a whole spectrum of oncogenes.

Oncogene activation is often through somatic genetic alterations. There are multiple types of alterations in cancer, including point mutations, small insertions and deletions less than 50 bp (indels), copy number variations (CNVs), as well as large DNA rearrangements. Point mutations are genetic mutations where single nucleotide bases are changed, inserted or deleted. Indels can cause frameshift, while point mutations can cause missense mutations, both of which often lead to changes in functions of the resulting proteins [9]. CNVs are an important type of genetic variations, affecting a greater segment of the genome than point mutations [10]. CNVs can lead to loss of tumor suppressor genes or gain of oncogenes. Large DNA rearrangements that are also called structural variations (SVs). They will be introduced in detail in later sections.

Although most mutations are passenger events, not leading to the initiation or promotion of cancer development, exploring driver events can help elucidate oncogenic pathways, provide potential therapy targets, and improve cancer treatment [11]. Countless efforts have been invested into the identification of oncogenes, and numerous point mutations have been identified to locate in and activate oncogenes. For example, *KRAS* is one of the best known oncogenes, and its mutations account for 20.4% of *KRAS* in non-small cell lung cancer (NSCLC) with the dominant substitution of G12C and for up to 67.6% of *KRAS* in pancreatic adenocarcinoma with G12D as the dominant mutant subtype [12]. To target this oncogene, sotorasib, a *KRAS* G12C inhibitor, was approved to be administrated to adult NSCLC patients and became a great breakthrough in target therapies [13]. Another example is epidermal growth factor receptor (EGFR), which has been identified as a biomolecular target for cancer since its discovery [14]. Aberrant activation of EGFR has been strongly associated to the etiology of several human epithelial cancers, and thus intense efforts have been made to inhibit EGFR activity by designing antibodies or small molecules [15]. The clinical approval of EGFR inhibitors such as afatinib,

dacomitinib, erlotinib, gefitinib, osimertinib in NSCLC and lapatinib in HER2-positive breast cancer has improved the prognosis greatly [16]. However, not all tumors carry druggable oncogenic mutations, so the investigation into more novel oncogenes, targetable mutations and new therapeutics is still a main focus in the field [17].

Genomic instability and structural variations

Cancer is known to be a disease involving dynamic alterations in the genome [18], so called genomic instability. Genomic instability is an evolving hallmark of cancer, providing the ground for cancer cells to develop a variety of abilities that drive uncontrolled growth and metastases in a multistep manner [2]. It is well established that the multistep process of oncogenesis is reflecting the accumulation of genetic and genomic alterations that transform normal cells into malignancies [3].

There are various forms of genomic instability [19]. Most cancers carry the form of chromosomal instability (CIN), where chromosome structure and copy number change at a high rate compared to normal cells. Although CIN is the major form of genomic instability, there are other forms described, including microsatellite instability [13], which is characterized as expansion or contraction of the number of oligonucleotide repeats present in microsatellite sequences [20], and forms of genomic instability that cause increased frequencies of base-pair mutations [21]. In hereditary cancers, these forms of genomic instability are associated with mutations in DNA repair genes. Well-documented examples include germline mutations in breast cancer susceptibility 1 (*BRCA1*), *BRCA2*, *RAD50*, Fanconi anemia genes, and some other genes functioning in DNA double strand break repair or DNA interstrand cross links [22, 23]. Although the germline mutations in such caretaker genes explain the presence of genomic

instability in inherited cancers, the molecular basis of genomic instability in sporadic cancers remains unclear [24].

Among different types of mutations caused by genomic instability, we are specifically interested in SV and their oncogenic consequences. The term SV (structural variant) refers to a spectrum of genomic rearrangements generally larger than 50 bp, including translocations, inversions, insertions, deletions, tandem duplications and other complex SVs. SVs are the major consequences of CIN [25]. Somatic SVs refer to the SVs that occur during the development processes of an organism, in contrast to germline genomic rearrangements that occur in reproductive cells, passed on from parents to offspring. SVs always involve breakage and rejoining of DNA fragments, so whole chromosomal gains and losses are not considered as SVs. Balanced SVs do not cause copy number changes, while SVs that have genomic imbalances can lead to CNVs [26].

Genomic instability has a variety of oncogenic consequences, including facilitating tumor progression via multiple mechanisms such as the downregulation of damage surveillance mechanisms [27]. Even before the structure of DNA was defined, it started to be appreciated that the oncogenic consequences of genomic instability could be significant, as the theory that tiny microscopic bodies, chromosomes, were abnormally present in cancer cells was proposed [28, 29]. In the 1950s, it was appreciated that mutations could be the origin of the biological variation observed in cancer [30]. Decades ago, as gene cloning, chromosome banding, molecular cloning, and more approaches were developed, chromosomal translocations have guided the discoveries of many novel oncogenes [31, 32]. There are recurrent karyotypic abnormalities in multiple tumor types. The most famous case, Philadelphia chromosome, a derivative of chromosome 19 and 22, was identified in chronic myelogenous leukemia patients as the first consistent karyotype

abnormality in a human cancer [33]. As some activated oncogenes were identified in Philadelphia chromosome, more studies followed up, indicating that the molecular consequences of recurrent genomic rearrangements lead to oncogene activation, providing successful targets for drug therapies [34-37].

Although genomic instability and a large number of somatic mutations in cancers have been investigated for near a century, the analysis of DNA sequences was limited for decades due to the fact that the sequencing technology allowed only hundreds of nucleotides at a time. Development and improvement of high throughput sequencing technologies in the past 20 years brought an essential turning point in the field by enabling whole genome sequencing (WGS) and analysis at a large scale.

In 2001, the International Human Genome Sequencing Consortium and Celera Genomics published initial haploid drafts of the human genome separately [38, 39], providing an assembly of the reference genome. The groundbreaking studies on about 20 genomes from breast and colorectal cancers were followed five years later [40, 41]. Years later, larger-scale studies conducted by multi-institutional consortium, The Cancer Genome Atlas (TCGA), were published, providing expression profiles and genomic data from more than a thousand tumor and matched normal WGS pairs across more than 30 tumor types. As the sequencing cost goes down, recent large consortium studies, such as the metastatic tumor study from Hartwig Medical Foundation and the Pan-Cancer Analysis of Whole Genomes [42], further foster explorations into genetic and genomic alterations, as well as identification of recurrent driver mutations in cancer genomes by providing comprehensive tumor profiling with an enormous amount of data [43, 44]. Accessibility of these datasets significantly promoted cancer research, especially on the

key question of what are the impacts of somatic SVs at system level and the underlying mechanisms [45, 46].

There are varieties of consequences brought by SVs leading to oncogenesis, and the most extensively studied cases are SVs altering protein coding genes. Here are some scenarios: A) The duplications can amplify oncogenes and cause over-expression of oncogenes. They can be small-scale, including individual or a group of genes, to large-scale, causing genome duplication [42]. For example, *MYCN*, a MYC family member, is frequently amplified in about 25% of neuroblastomas and this is associated with poor prognosis [47]. Another well-studied example is *ERBB2*, also known as *HER-2* or *NEU*, which is amplified in 20-25% of primary breast cancers and at similar frequency amplified in ovarian cancers [8, 48]. B) The deletions can also cause loss of functions of tumor suppressor genes. For example, identified in 1994, *CDKN2A* gene located in chromosome 9p21 is the most frequently deleted genes in cancer [49]. Another frequently mutated tumor suppressor gene, *PTEN*, is located in chromosome 10q23, which is found to be commonly deleted in brain, prostate and bladder cancers [50]. C) SVs can produce oncogenic fusions, whose product proteins can drive cancer development. A most known case is BCR-ABL fusion found in chronic myeloid leukemia as the molecular product of Philadelphia chromosome [34], and the fusion protein became a therapeutic target to treat patients [51]. Another example is *ALK-RET* fusion in lung cancer. *ALK* gene activated by fusions to other genes with a recurrence of 3-6% in lung adenocarcinoma [52]. Tyrosine kinase inhibitors have become the standard drug treatment for advanced cases of lung adenocarcinoma harboring related mutations [53].

Gains of oncogenes, losses of tumor suppressors, and oncogenic gene fusions have been extensively studied, and related computational tools and experimental models have been well

developed. However, > 98% of human genome is non-coding, which means these parts do not translate into proteins. Existing studies largely underestimate the important roles of mutations located in non-coding regions as well as the regulatory functions of noncoding sequences.

Enhancers and their chromatin features

Opposed to *trans*-regulatory sequences that encode transcription factors (TFs) binding to *cis*-regulatory elements, *cis*-regulatory sequences regulate gene expression by binding to different TFs, and mutations affecting their activities are considered to be the most important cause of phenotypic divergence [54]. *Cis*-regulatory sequences can be discretized as *cis*-regulatory elements (CREs) that are composed of non-coding DNA containing binding sites TFs and other regulatory molecules needed to regulate gene transcription [55]. Promoters and enhancers are the best understood types of CREs [56].

Enhancers are DNA sequences containing multiple binding sites for a variety of TFs, and play important roles in the regulation of gene transcription [57]. Enhancers can regulate transcription independent of their location, distance or orientation related to the gene promoters. To achieve this, enhancers can interact with components of the mediator complex or transcription factor II D (TFIID) to help recruit RNA polymerase II (RNAPII) by extension [58]. In addition, activating genes in eukaryotes necessitates the loosening of the chromatin. Enhancer-bound TFs play a crucial role in this process by recruiting histone-modifying enzymes or ATP-dependent chromatin remodeling complexes. These actions modify the chromatin structure, enhancing the DNA accessibility to other proteins [59].

In the past twenty years, the technologies that can detect chromatin accessibility or map genome-wide epigenetic markers facilitate our understanding about how histone modifications

affect gene expressions, and also lead to new insights on the chromatin features of enhancers. Assay for transposase accessible chromatin with high-throughput sequencing (ATAC-Seq) assesses DNA accessibility with hyperactive Tn5 transposase, which inserts sequencing adapters into accessible regions of chromatin. In this approach, sequencing read coverages is used to infer regions of increased accessibility that might have more TF binding and be under active transcription [60]. ChIP-sequencing (ChIP-Seq) is a method to detect histone modification. It combines chromatin immunoprecipitation (ChIP) with next generation sequencing to identify DNA sequences binding specific TF [61]. Later, Cleavage Under Targets and Release Using Nuclease (CUT&RUN) was invented as a new chromatin profiling strategy where antibody-targeted controlled DNA cleavage releases protein-DNA complex supernatant for sequencing. It is an easier and higher-resolution method to detect TF or epigenetic marker binding on chromatin, alternative to ChIP-Seq [62]. These approaches and the genome-wide studies of histone modifications have greatly driven our understanding of the chromatin landscape of enhancers and then functional significance in gene expression regulation.

The distribution of histone modifications and some particular histone variants impact gene expression by directing the interaction of TFs and chromatin fiber [63]. Cyclic AMP-responsive element-binding (CREB) protein (CBP) and p300 are two proteins that have histone acetyltransferase activity and multiple functional domains to interact with other TFs and histone modifications [64]. After the extensive mapping of these proteins in different tissue types, it is well-known that there is a correlation between the presence of p300 and enhancer function, and cell type-specific occupancy of enhancers by CBP and/or p300 regulates distinct transcriptional programs in many cell types [65]. The maps of various histone modifications as well as transcription regulators like CBP and p300 have provided further insights of the distinct

chromatin features of different regulatory sequences [66]. Markers for active enhancers and promoters are well established. The presence of RNAPII and TBP-associated factor 1 (TAF1) can define active promoters, which are marked by marked by nucleosome-free, accessible regions with flanking histone H3 trimethylated at lysine 4 (H3K4me3). However, putative enhancers can be predicted by the presence of distant p300 binding sites and are highly enriched in H3K4me1, H3K4me2 and histone 3 acetylated at lysine 27 (H3K27ac). H3K27ac is now considered as the marker of functionally active enhancers [67]. By contrast, enhancers associated with H3K4me1 and H3K27me3 are linked to inactive genes.

It was reported that enhancers could interact with promoters in order to activate transcription. How enhancers find and interact with distant core promoters to trigger transcription, and the mechanisms that stabilize these interactions, are still under active investigation.

3D genome organization and gene expression regulation

Enhancers are key regulatory elements that control spatiotemporal gene expression programs by engaging in physical contacts with their cognate genes. This process is often through long-range chromosomal interactions, where gene promoters and enhancers can be hundreds of kilobases (kb) away [68]. In order to draw out their effect, enhancers are considered to be brought into close spatial proximity with target promoters through the formation of “chromatin loops”, and these loops build the three-dimensional (3D) organization of chromatin structure. Studies on 3D chromatin organization have suggested that chromosomes are hierarchically organized into large compartments composed of smaller domains called topologically associating domains (TADs) that are at sub-megabase scale, and the disruptions of normal TADs are frequently associated with diseases [69].

The technology to identify DNA fragments that interact closely within the three-dimensional (3D) space was initially employed in 1993 [70] and was further refined and broadened in 2002 [71], laying the groundwork for all chromosome conformation capture (3C) technologies. This includes Hi-C, which is a high-throughput variant of 3C. In most methods based on 3C, the initial step is crosslinking cells with formaldehyde. Subsequent procedures typically involve breaking down the chromatin into fragmentation of DNA using restriction enzymes or sonication. In standard 3C-based protocols, DNA digestion is followed by proximity-based ligation of adjacent DNA ends and determination of pair-wise interactions using either PCR or sequencing approaches. For next steps, different strategies are used to identify the chromatin interactions. The classical 3C method tests one pair of interacting loci at one time using quantitative PCR (qPCR). In the chromosome conformation capture-on-chip (4C), a second round of digestion and ligation is performed to increase resolution, followed by PCR with locus-specific primers to detect genome-wide interactions containing the locus of interest [72]. In 5C, the ligated and purified DNA is directly amplified using primers for all restriction fragments within a consecutive genomic region, usually hundreds of kilobases up to several megabases. The PCR products are sequenced and provide information about the ligation frequencies of all fragments within this region [73]. In Capture-C method, enrichment of interacting pairs can be done using biotin-labelled probes that are designed for restriction fragment ends of interest [74]. In Hi-C protocols, the restriction fragment ends are labelled using biotin, ligated products are enriched using streptavidin pull-down after sonication and interactions are interrogated in a genome-wide all-versus-all unbiased manner. Hi-C output can be in 1 kb resolution, showing the global 3D interaction map of genome. This method is now applied to single cells, providing information about 3D genomes in individual cell level [75]. Micro-C employs Micrococcal

nuclease to fragment the genome, which overcomes the resolution limit of restriction enzyme-based methods. Micro-C provides an improvement for 3C-based methods and resolves the fine-scale level of chromatin folding [76].

Chromatin Interaction Analysis with Paired-End-Tag sequencing (ChIA-PET) method is another emerging method for chromatin interactions at a global scale and higher resolution. In the ChIA-PET protocol, cells are first treated with formaldehyde to cross-link chromatin interactions, DNA segments bound by proteins are enriched by ChIP, and interacting DNA fragments are then captured by proximity ligation. The Paired-End Tag (PET) strategy is applied to the construction of ChIA-PET libraries, which are then sequenced. The results of ChIA-PET is a genome-wide map of the protein binding sites and chromatin interactions mediated by the protein of interest [77].

For visualization, results from Hi-C or other high-throughput technologies for 3D genome interactions are usually shown in heat-maps with plaid patterns. The plaid pattern reflects the interacting compartments in genome. Inter-compartmental domain interactions are stochastic, and their frequency or stability might rely on the quantity, affinity, and interaction capabilities of the involved proteins, which influence the cooperativity of these interactions. Active and inactive regions of the genome, known as A and B compartmental domains respectively, contain distinct sets of multivalent proteins. These proteins may interact with others within the same class, forming two separate phases that prevent interactions between A and B compartments. Phase separation of chromatin into droplets could regulate functions of compartmental domain interactions. The dynamics of droplet activity within a cell population might account for why active compartmental domains seem to interact with other active locus across the chromosome in Hi-C heat maps [78].

It is well appreciated that chromatin loops are important driving mechanisms of gene expression regulation. Enhancer-promoter interactions seem to be mostly constrained within a TAD. It was first shown in 2012 that experimentally induced contact between the mouse β-globin (*Hbb*) promoter and its locus control region enhancer ('forced chromatin looping') led to strong transcriptional activation of the *Hbb* gene, even without a key transcriptional activator GATA1 [79]. This study demonstrated that enhancer–promoter contacts is sufficient to induce transcription. Using forced chromatin looping target dCas9 fusion proteins to defined genomic loci, engineered chromatin loops can be experimentally achieved and induce gene activation in a reversible manner [80]. These studies suggested that forced chromatin looping may ultimately enable precision 3D genome rewiring with potential for therapeutic applications.

Another important question is what the formation mechanisms and processes of TADs are. The distinctive feature of regulatory or structural chromatin loops may be the stability of the loop, which might be increased by the binding of specific factors promoting loop formation. TAD boundaries are enriched for insulator proteins such as CCCTC-binding factor (CTCF) (detected at ∼76% of all boundaries), active transcription histone marks such as H3K4me3 and H3K36me3, nascent transcripts, housekeeping genes (present in ∼34% of TAD boundaries), and repeat elements [81]. Recent studies involving 76 DNA-binding proteins have pinpointed components of the cohesin complex, CTCF, Yin Yang 1 (YY1), and Zinc Finger Protein 143 (ZNF143) as being significantly enriched at the anchors of strong chromatin interactions [82]. Along with the mediator complex that is recognized for its pivotal role in connecting enhancers and promoters within 3D space, both CTCF and cohesin have been identified as critical for the formation of chromatin loops. They are suggested to work together as architectural proteins, linking either facultative or constitutive chromosome architecture with gene regulatory outcomes

[83]. During the formation of chromatin loops, the loop extrusion complex, such as cohesin complex with structural maintenance of chromatin protein 1 (SMC1) SMC3, SCC1 and SCC3 subunits, binds to chromatin and makes loop extend in both directions until a border element such as CTCF is encountered [84]. Given the crucial regulatory functions of 3D genome organization and the key roles of border elements in TAD formation, it is not surprising that the disruption of CTCF or CTCF binding sites will lead to abnormal gene expression and phenotypes. Loss of CTCF is lethal during embryonic development, and haploinsufficiency of CTCF results in intellectual disability, microcephaly and growth retardation [85]. Heterozygous CTCF-knockout mice render a high incidence of tumors, and mutations of specific CTCF binding sites show correlations with multiple cancer types in human [86]. In addition, changes in CTCF looping at specific genomic sites have effects on the expression of nearby genes [87]. Therefore, in cancer genomes, DNA rearrangement with breakpoints located in non-coding regions, is likely to disrupt CTCF binding sites and thus change 3D genome organization, which can cause abnormal gene activation or silence and contribute to oncogenesis.

Enhancer hijacking and the approaches to infer enhancer hijacking events

As described in the previous sections, cancer cells utilize a variety of mechanisms to activate proto-oncogenes to obtain selection advantages and survive. Enhancers play an important role in activating gene expression by recruiting TFs and transcriptional machinery to the promoters that locate in the same topologically associating domains (TAD). Since SVs can disrupt 3D genome organization, they may induce “enhancer hijacking” if an active enhancer is rearranged such that

it regulates genes that are not their original targets. This phenomenon can be induced by SVs in cancer, and it can happen when there is a breakpoint close to a gene promoter region (**Fig. 1**).

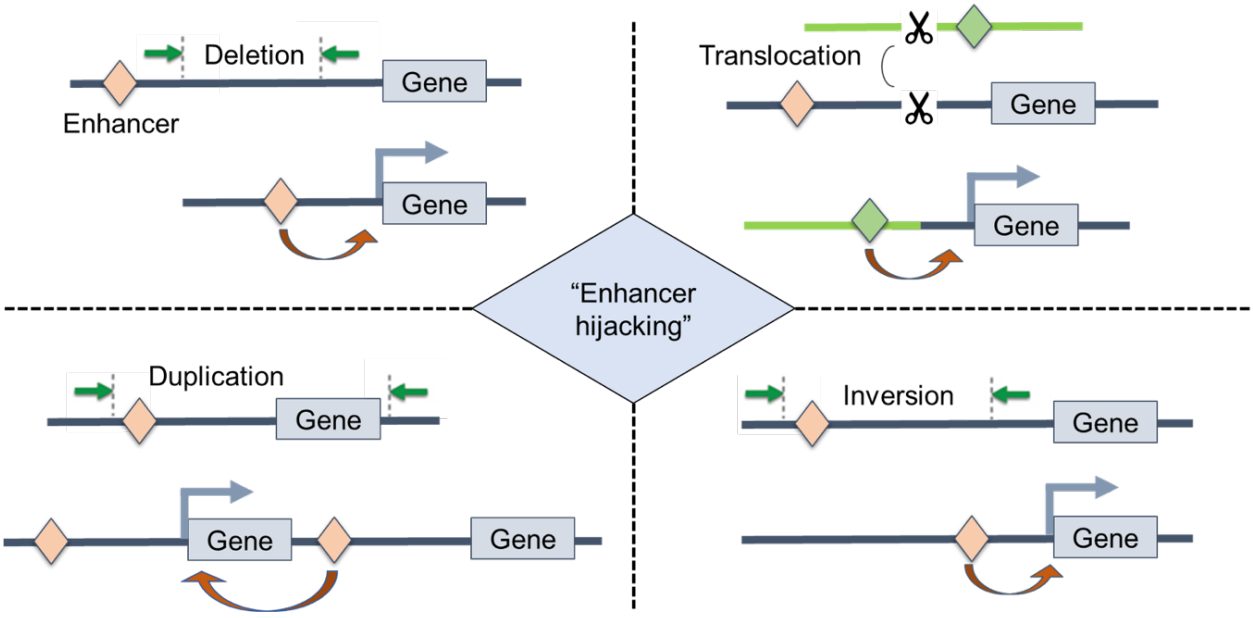

Figure 1. Diagram of enhancer hijacking events.
Enhancer hijacking events can happen when deletions, duplications, translocations or inversions bring distal enhancers to genes that are not actively transcribed in normal cells.

Decades ago, enhancer hijacking was first described for the activation of c-*myc* during B cell lymphomagenesis in mice [88]. In 2014, researchers demonstrated that GFI1 family oncogenes can be activated by somatic SVs in group 3 and group 4 medulloblastoma [89]. In 2017, it was reported that small insertions at the *LMO2* locus produces enhancer function and drive aberrant gene expression in MOLT4 T-lineage acute lymphoblastic leukemia (T-ALL) cells [90]. As a matter of fact, enhancer hijacking events are frequently occurring in multiple cancer types, including adult and pediatric cancers. Individual enhancer hijacking events have been appreciated in multiple cancer types: In neuroblastoma, activation of *MYCN* as a consequence of amplification can be driven by local or distal enhancers [91]. DNA rearrangements translocate active enhancer to activate NR4A3, a TF that then upregulates its target genes in acinic cell

carcinomas of the salivary glands and promotes oncogenesis [92]. *IRS4* in lung cancer and *IGF2* in colorectal cancer were identified as recurrent targets of enhancer hijacking in a pan-cancer study [93]. Rearrangements lead to enhancers mistargeting *CCNE1* and *IGF2* in primary gastric adenocarcinoma [94]. As more individual enhancer hijacking events and target oncogenes were identified, such events are better appreciated, and some computational tools have been developed to detect enhancer hijacking using next generation sequencing (NGS) data.

Enhancer hijacking events can be inferred from genomic and transcriptomic data, or from chromatin conformation data, such as Hi-C seq data. There are several algorithms detecting enhancer hijacking genes based on large consortium datasets, including CESAM (*cis* expression structural alteration mapping) [93] and PANGEA [95]. CESAM is a framework that can infer cancer-related gene overexpression caused by CRE reorganization by integrating somatic copy number alterations (SCNAs), gene expression data and information on TADs. It applied linear regression model, adjusting for confounders like the total number of SCNAs and principal components, to relate TAD-binned SCNA breakpoints with gene expression changes to detect enhancer hijacking events [93]. PANGEA can identify recurrent noncoding mutations including SNVs, small indels, CNAs and SVs that disrupt enhancer/promoter sequences or their interactions. It employs weighted elastic net to perform regression analysis to find the impacts of these noncoding mutations on gene expressions, and thus identify mutations that influence gene expression [95]. The drawbacks of these algorithms are related to the regression models based on linear regression, in which outliers can impair the performances. And PANGEA requires the annotation of tissue-specific promoter-enhancer pairs, which are not available for many tumor types.

On the other hand, tools like Cis-X [96] and NeoLoopFinder [97] can infer enhancer hijacking events with individual sample data instead of cohorts of data. The cis-X framework is designed to integrate WGS, RNA-seq data and functional genomics data like ChIP-seq from individual tumor genomes for the analysis of gene regulation. It focuses on identifying cis-activated genes using key indicators such as allele-specific expression and unusually high gene expression levels [96]. NeoLoopFinder can identify enhancer-hijacking events directly from genome-wide chromatin interaction experiments such as Hi-C [97]. The limitation for these tools mainly resides in the fact that chromatin conformation or functional genomics data are still limited for most tumor patients or cohort studies, so identifying recurrent events might be a major challenge. Thus, tools that leverage large-scale whole-genome and transcriptome sequencing data would be more effective in detecting oncogene activation driven by SVs.

Oncogene activation in neuroblastoma

Neuroblastoma, a tumor that derives from primitive sympathetic neural precursors, is among the most common childhood solid tumors and is the most common cancer diagnosed during infancy, accounting for approximately 8% of all childhood cancers and 15% of childhood cancer mortality. Neuroblastoma displays great clinical and genetic heterogeneity [98]. It can be classified into distinct risk groups based on well-defined criteria (imaging stage, age at the time of diagnosis, histology, differentiation, amplification of *MYCN*, diploidy and 11q aberration) [99]. Patients with non-high-risk neuroblastoma, low- and intermediate-risk categories, represent nearly half of all newly diagnosed cases. Those patients usually do not need intensive treatments to cure the tumor, and some children (especially young infants with small tumors) might not need to be treated at all because some of these neuroblastomas will mature or disappear

automatically [100]. High-risk neuroblastoma has a much worse prognosis with the overall survival of 50%, and patients with high-risk neuroblastoma have to take intense multi-modal treatment, including chemotherapy, radiation therapy, surgery resection, autologous stem cell transplantation, immunotherapy, and the differentiating agent (to make the tumor cells differentiate and less aggressive) [101]. Therefore, it is urgent to identify novel actionable targets for further improvement of existing treatments.

A number of oncogenes have been reported to drive high-risk neuroblastoma, including *ALK*, *MYCN*, and *TERT*. Many efforts have been put into identifying cancer driving mutations and understanding the oncogenic mechanisms.

Genetic predispositions have been extensively studied to identify risk related mutations. A study mapping for single nucleotide polymorphisms (SNPs) that have linkage with neuroblastoma predisposition first identified the linkage signal on the short arm of chromosome 2 (2p23-2p24), which included *MYCN*, but no sequence mutations were found. *ALK* was identified as the major familial neuroblastoma predisposition gene [102]. With genome-wide association study (GWAS), common SNPs at 6p22 within the predicted genes *FLJ22536* and *FLJ44180* were identified to be associated with neuroblastoma [103]. Following on those studies, a large-scale high-risk neuroblastoma study revealed that 6p22 and 2q35 SNPs were associated with aggressive neuroblastoma [104]. Besides SNP genotypes, CNVs represent a substantial part of genetic diversity that can upregulate oncogenes and is associated with risk level. Somatic copy gain and high-level amplification of the *ALK* locus have been identified as recurrent genomic aberration in neuroblastomas, suggesting multiple mechanisms can activate this gene, contributing to neuroblastoma development [105].

MYCN is an oncogene found to be amplified in ~25% of neuroblastoma patients [47]. It is a gene homologous to v-*myc* but distinct from *MYC* in human neuroblastoma [106]. As a MYC family protein, MYCN expression is especially high in early developmental stages, and is important for the morphology of nervous system, with a direct role in blocking differentiation pathways and maintaining pluripotency [107]. The amplification of *MYCN* is a high-risk marker, and maintains an undifferentiated and aggressive phenotype, leading to poor prognosis [108]. The advances of NGS and inter-institutional collaboration have deepened our understanding of neuroblastoma biology and risk classification. Recent studies have demonstrated the association between genomic status and clinical outcome [109]. Compared to adulthood cancers, pediatric tumors usually have fewer point mutations and small indels, indicating that genomic rearrangements play important roles in oncogenesis. It has been demonstrated that *MYCN* can be upregulated as a result of CNVs (copy gain), or by extrachromosomal circular DNA (ecDNA) amplicons [91].

However, mutations related to protein coding regions could not explain all the oncogene activation cases. As we have discussed in previous sections, the non-coding regions in the genome contribute substantially to gene expression regulation, and their rearrangements can lead to oncogene activation by positioning enhancers to gene promoter regions. It has been reported that focal enhancer amplification or genomic rearrangements leading to enhancer hijacking could result in the activation of *MYC*, and drive a subset of high-risk neuroblastoma [110]. In addition, MYCN amplification frequently happens in extrachromosomal DNA (ecDNA). The exploration of *MYCN* amplicons and the epigenetic markers within the ecDNA sequences revealed that there were co-amplifications of proximal enhancers or distal chromosomal fragments harboring enhancers together with *MYCN*, suggesting the crucial role of enhancer hijacking events to drive

MYCN high-expression [91]. Furthermore, the genomic rearrangements connecting with distal super enhancers can also activate *TERT* and lead to aggressive tumor phenotypes in high-risk neuroblastoma [111]. The emerging role of enhancer hijacking has been more and more identified to explain oncogene activation process, but it is still needed to use unbiased bioinformatic approaches to define driver SV events in different groups, especially high-risk neuroblastomas, to investigate novel oncogenes and biomarkers to improve our knowledge of prognosis and treatments.

Questions remaining to be addressed

Given all the advances in the field of enhancer hijacking, there are still many questions awaiting to be addressed.

(1) A highly sensitive and reliable bioinformatic tool that uses whole-genome and transcriptome sequencing data is required to unbiasedly explore novel oncogenes activated by enhancer hijacking.

(2) The putative oncogenes and how they are regulated by rearranged regulatory sequences remain to be demonstrated after the analysis by computational methodologies. This can be achieved by individual gene functional studies with experiments, or by a comprehensive screening. Epigenetic markers and chromatin conformation information are capable and available to study the interactions between enhancers and gene promoters.

(3) The oncogenic functions of less-investigated non-coding genes and their associated mechanisms need to be elucidated.

(4) Unbiased studies that can infer putative novel oncogenes from gene activation mechanisms are still challenging, especially in pediatric tumors like neuroblastoma. This type of

analysis will provide genetic biomarkers and promising targets to guide patient classification and precise low-toxicity treatments, thus having great significance in neuroblastomas and pediatric brain tumors which are known to have considerable heterogeneity and be in need of efficient therapies for some patient groups.

(5) How the enhancer hijacking genes drive cancer together with other driving mutations, and what are the recurrence of the enhancer hijacking events in large patient populations need to be explored with larger-scale studies but not limited to a single patient cohort.

HYENA Detects Oncogenes Activated by Distal Enhancers in Cancer

This chapter is adapted from a published study: Anqi Yu, Ali E Yesilkanal, Ashish Thakur, Fan Wang, Yang Yang, William Phillips, Xiaoyang Wu, Alexander Muir, Xin He, Francois Spitz, Lixing Yang, HYENA detects oncogenes activated by distal enhancers in cancer, *Nucleic Acids Research*, 2024;, gkae646, https://doi.org/10.1093/nar/gkae646.

Introduction

At the mega-base-pair scale, linear DNA is organized into topologically associating domains (TADs) [81], and gene expression is regulated by DNA and protein interactions governed by 3D genome organization. Enhancer-promoter interactions are mostly confined within TADs [112-114]. Non-coding somatic single nucleotide variants (SNVs) in promoters and enhancers have been linked to transcriptional changes in nearby genes and tumorigenesis [115]. Structural variations (SVs), including deletions, duplications, inversions, and translocations, can dramatically change TAD organization and gene regulation [116] and subsequently contribute to tumorigenesis. Previously, we discovered that *TERT* is frequently activated in chromophobe renal cell carcinoma by relocation of distal enhancers [117], a mechanism referred to as enhancer hijacking (**Fig. 2A**). In fact, many oncogenes, such as *BCL2* [118], *MYC* [119], *TAL1* [120], *MECOM/EVI1* [121], *GFI1* [89], *IGF2* [94], *PRDM6* [122], and *CHD4* [95], can be activated through this mechanism. These examples demonstrate that genomic architecture plays an important role in cancer pathogenesis. However, the vast majority of the known enhancer hijacking target oncogenes are protein-coding genes, and few non-coding genes have been reported to promote diseases through enhancer hijacking. Here, we refer to non-coding genes as all genes that are not protein-coding. They include long non-coding RNAs (lncRNAs), pseudogenes, and other small RNAs such as microRNAs, small nuclear RNAs (snRNAs), small

nucleolar RNAs (snoRNAs), etc. They are known to play important roles in many biological processes [123], and some are known to drive tumorigenesis [124]. In this study, we will focus on identifying oncogenes, including oncogenic non-coding genes, activated by enhancer hijacking.

Several existing algorithms can detect enhancer hijacking target genes based on patient cohorts, such as CESAM [93] and PANGEA [95]. These two algorithms implemented 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 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 the annotation of tissue-specific promoter-enhancer pairs, which are not readily available for many tumor types. Cis-X [96] and NeoLoopFinder [125] 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, which remain very limited. Furthermore, the identification of recurrent mutational events that result in oncogenic activation requires large patient cohorts. Therefore, tools that use whole-genome and transcriptome sequencing data, which are available at much larger sample sizes, would be more useful in identifying SV-driven oncogene activation. Finally, no non-coding oncogenes 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 [126] 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. Among the 108 putative oncogenes detected by HYENA, we studied the oncogenic functions of a lncRNA, *TOB1-AS1*, and demonstrated that it is a regulator of cancer cell invasion *in vitro* and tumor metastasis *in vivo*.

Materials and 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 RNA-Seq 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, 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 [126]. Somatic and germline SNVs, somatic copy number variations (CNVs), SVs, and tumor purity were detected by multiple algorithms and consensus calls were made. Genome coordinates were based on the 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). *TOB1* and *TOB1-AS1* expression data in CCLE pancreatic cancer cell lines were downloaded from DepMap Public 22Q2 version (https://depmap.org/portal/download/all/). Gene expression data of the Cancer Genome Atlas (TCGA) PAAD cohort (TCGA.PAAD.sampleMap/HiSeqV2_PANCAN) and International Cancer Genome Consortium (ICGC) PACA-CA cohort for 45 samples of which "analysis-id" were labeled as "RNA" were downloaded from Xena Data Hubs (https://xenabrowser.net/datapages/) and ICGC data portal (https://dcc.icgc.org/projects/PACA-CA) respectively.

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.

The raw sequencing data for Hi-C and ATAC-Seq were available through NCBI Sequence Read Archive (SRA) with accession number PRJNA1036282. The raw sequencing data for mouse xenograft tumor RNA-Seq were available through NCBI SRA with accession number PRJNA1011356.

HYENA algorithm

First, small tandem duplications (<10 kb) were discarded since they are unlikely to produce new promoter-enhancer interactions. The remaining SVs were mapped to the flanking regions (500 kb 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 carrying SVs in at least 5% of 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 the *quantile.normalize()* function from the *preprocessCore* R package to enhance cross-sample comparison. For each gene, samples were ranked based on their 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

A 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 the sample size of the cohort and up to 20 if the sample size was more than 200. The regression model was as shown 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. A similar strategy was used to detect eQTLs in normal tissues [127].

Calculating empirical P values and model selection

Gene expression data were permuted 1,000 times by randomly shuffling expression values within the cohort. For tumor types with more than 10,000 genes to test (**Supplementary Table S1),** only 100 permutations were performed to reduce run time. The normal-score regression was performed in 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 the Benjamini-Hochberg procedure. Genes with FDR less than 0.1 were considered candidate genes. For example, in MALY, there were 1,863 genes reaching 5% SV frequency and 1,863 *P* values were obtained in each permutation. After 1,000 permutations, 1,863,000 *P* values were generated and

should represent the null distribution very well. Empirical *P* values were calculated using these 1,863,000 permuted *P* values.

The above empirical *P* value calculation and candidate gene detection were performed iteratively with no PCs 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 lowest number of PCs reaching 80% of the maximum number of candidate genes in all regression models tested was selected as the final model to avoid over fitting. For example, the sample size for PCAWG UCEC was 42; therefore, we tested from 0 to 4 PCs. Among these, the model including 4 PCs gave the highest number (4) of candidate genes. Therefore, the model including 4 PCs with 4 candidate genes was selected as the final model (**Supplementary Table S2**).

In our normal-score regression, we essentially attempt to model variations in gene expression. Including confounding factors will improve performance. Tumor purity, gene copy number, patient age, and sex are factors known to affect gene expression. Therefore, they were included in the regression model. Unobserved variations may include tumor subtype, tumor stage, patient ethnicity, smoking status, alcohol consumption, and other unknown factors that may alter gene expression. Since HYENA was designed for wide applications, we did not require users to provide information on tumor subtype, tumor stage, patient ethnicity, smoking status, alcohol consumption, etc. Principle component analysis is a linear decomposition of gene expression variations. Therefore, including PCs in a regression model was suitable for removing systematic variations and could better model the effects of SV status. However, some enhancer hijacking target genes are master transcription factors, such as *MYC*, and have a profound impact on the gene expression of multiple pathways. Hence, it is possible that some PCs capture the

activities of transcription factors. If these transcription factors were activated by somatic SVs, the PCs would be correlated with SV status. Including these PCs would diminish our ability to detect the effects of SV status. Therefore, we excluded these PCs from the regression model.

Testing eQTL-SV associations

Known germline eQTLs from the matching tissues were obtained from GTEx (**Supplementary Table S3**). The associations between germline genotypes of eQTLs and SV status of the candidate genes in the PCAWG cohort were tested using a Chi-squared test. Genes with significant correlations (P<0.05) between their SV status and at least one eQTL were removed. The remaining genes were our final candidate enhancer-hijacking target 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 adenocarcinoma, *TERT* in chromophobe renal carcinoma, *IGF2* in colorectal cancer, *IGF2* in stomach/gastric adenocarcinoma, *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 S4**). 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 [128] to predict the 3D genome organization. Two 500 kb sequences were merged according to the SV orientation to construct the sequence of the rearranged genome fragments. 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 [129] were used to facilitate TAD annotation together with predicted genome organization. H3K27Ac and CTCF ChIP-Seq data from the PANC-1 cell line were downloaded from the ENCODE data portal (https://www.encodeproject.org/). SV breakpoints were provided to Orca [130] to predict 3D genome structures through its web interface (https://orca.zhoulab.io/).

In situ Hi-C and ATAC-Seq

Ten million cells of Panc 10.05, PANC-1, PATU-8988S, and PATU-8988T cell lines were collected to construct Hi-C libraries [82]. The Hi-C libraries were sequenced on Illumina NovaSeq X Plus platform with 1% phix. About 2 billion reads were obtained from Panc 10.05, PATU-8988S, and PATU-8988T, and 1 billion reads were obtained from PANC-1. The paired-end reads were aligned to chromosomes 1-22, X, Y and M by bwa-mem. SVs were identified by

EagleC [131] at 5 kb, 10 kb and 50 kb resolutions. The non-redundant SVs in **Supplementary Table S5** were combined for the three resolutions. Chromatin loops were identified by NeoLoopFinder [97]. A probability threshold of 0.95 was used, and default values were used for all other parameters. Fifty thousand cells of Panc 10.05, PATU-8988S, and PATU-8988T cell lines were harvested to construct ATAC-Seq libraries [132]. The libraries were sequenced using Illumina NovaSeq. About 60 million reads were generated from each library. The paired-end reads were aligned to the reference genome by hisat2. Hi-C and ATAC-Seq read coverages were generated by deepTools with 10 bp bin-size, RPGC normalization, and an effective genome size of 2,864,785,220.

Cell lines

HEK293T, PANC-1, and PATU-8988T cells were obtained from Dr. Alexander Muir (University of Chicago). Panc 10.05 was purchased from ATCC (American Type Culture Collection, USA) (https://www.atcc.org/products/crl-2547) and PATU-8988S was purchased from DSMZ (https://www.dsmz.de/collection/catalogue/details/culture/ACC-204). 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), and Panc 10.05 cells were cultured in RPMI-1640 medium (Gibco, 11875093) containing 10% FBS, as per ATCC instructions (https://www.atcc.org/products/crl-3216, https://www.atcc.org/products/crl-1469, https://www.atcc.org/products/crl-2547). PATU-8988T and PATU-8988S 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). The cell lines were passaged 2-3 times a week. All cell lines have been regularly monitored and tested negative for mycoplasma using a 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 as serial dilutions, 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.

TOB1-AS1 transient knock-down using antisense oligonucleotides (ASOs)

Three Affinity Plus® ASOs were synthesized by Integrated DNA Technologies (IDT), with two targeting *TOB1-AS1* and one non-targeting negative control. The ASO sequences were:

Non-targeting ASO (NC): 5' -GGCTACTACGCCGTCA- 3'

TOB1-AS1 ASO1: 5' -GCCGATTTGGTAGCTA- 3'

TOB1-AS1 ASO2: 5' -CTGCGGTTTAACTTCC- 3'

The ASOs were transfected into PATU-8988S and Panc 10.05 cells with Lipofecatmine™ 2000 (Invitrogen, 11668019) using reverse-transfection method according to IDT protocol (https://www.idtdna.com/pages/products/functional-genomics/antisense-oligos) with a final ASO concentration of 9 nM. Cells were transfected in 6-well plates and incubated for 48 hours to reach 60% confluence before RNA extraction or Transwell assay.

RNA isolation and qRT-PCR

Cells were plated in 6-well plates and allowed to reach 80% confluence, or transfected by ASOs as described above, prior to RNA extraction. After cells lysis in 300μl/well TRYzol™

(Invitrogen, 15596026), RNA samples were prepared following the Direct-zol RNA Miniprep kit manual (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 as an endogenous control.

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, 50,000 PATU-8988T cells/well, 50,000 Panc 10.05 cells/well, or 50,000 PATU-8988S cells were resuspended in serum-free, phenol-red free

DMEM medium 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 medium without FBS in the bottom chamber were seeded with the same number of cells as indicated above as a negative control. At the end of the assay, the membranes were stained with 500µl 4µg/ml Calcein AM (CaAM) (Corning, 354216) for one 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). For tail vein inoculation, mice were injected intravenously through the tail vein with luciferase-expressing at 400,000 cells/mouse for PANC-1 cells in cold phosphate buffered saline (PBS) (Gibco, 10010-023). For orthotopic inoculation, mice were injected with 200,000 PANC-1 cells/mouse into the pancreas under general anesthesia. Cells were resuspended in cold PBS containing 5.6mg/mL Cultrex Reduced Growth Factor BME (R&D Systems, 3533-010-02). Primary tumor and metastatic tumor burdens were measured weekly for 4 and 6 weeks for tail vein injection models and orthotopic models, respectively, 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.

Ex vivo IVIS imaging

Ex vivo imaging was done for the PANC-1 orthotopic injection mice after 8 weeks of orthotopic inoculation. Mice were injected intra-peritoneally with D-luciferin solution at a concentration of 150 μg/g of body weight immediately before euthanasia. Immediately after necropsy, mice were dissected, and tissues of interest (primary tumors, livers and spleens) were placed into individual wells of 6-well plates covered with 300 μg/mL D-luciferin. Tissues were imaged using Xenogen IVIS 200 Imaging System (PerkinElmer) and analysis was performed (Living Image Software, PerkinElmer) maintaining the regions of interest (ROIs) over the tissues as a constant size.

Tumor RNA sequencing and gene expression analysis

RNA was isolated from mouse subcutaneous tumors (six *TOB1-AS1* overexpression and six control mice) after 6 weeks of PANC-1 cell subcutaneous injection using Direct-zol RNA Miniprep kit (RPI, ZR2052). Quality and quantity of the RNA was assessed using Qubit. Sequencing was performed using the Illumina NovaSeq 6000. About 40 million reads were sequenced per sample. The pair-end reads were aligned to mouse genome (mm10) and human genome (hg19) with hisat2, and the reads mapped to mouse or human genomes were disambiguated using AstraZeneca-NGS disambiguate package. Gene counts were generated with

htseq-count. Differential gene expression was analyzed using DESeq2. Differentially expressed genes were defined as genes with a FDR smaller than 0.1 and a fold change greater than 1.5.

Code availability

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

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 (RNA-Seq), 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 [133]. Furthermore, gene expression networks in cancer are widely rewired [134], and outliers of gene expression are common.

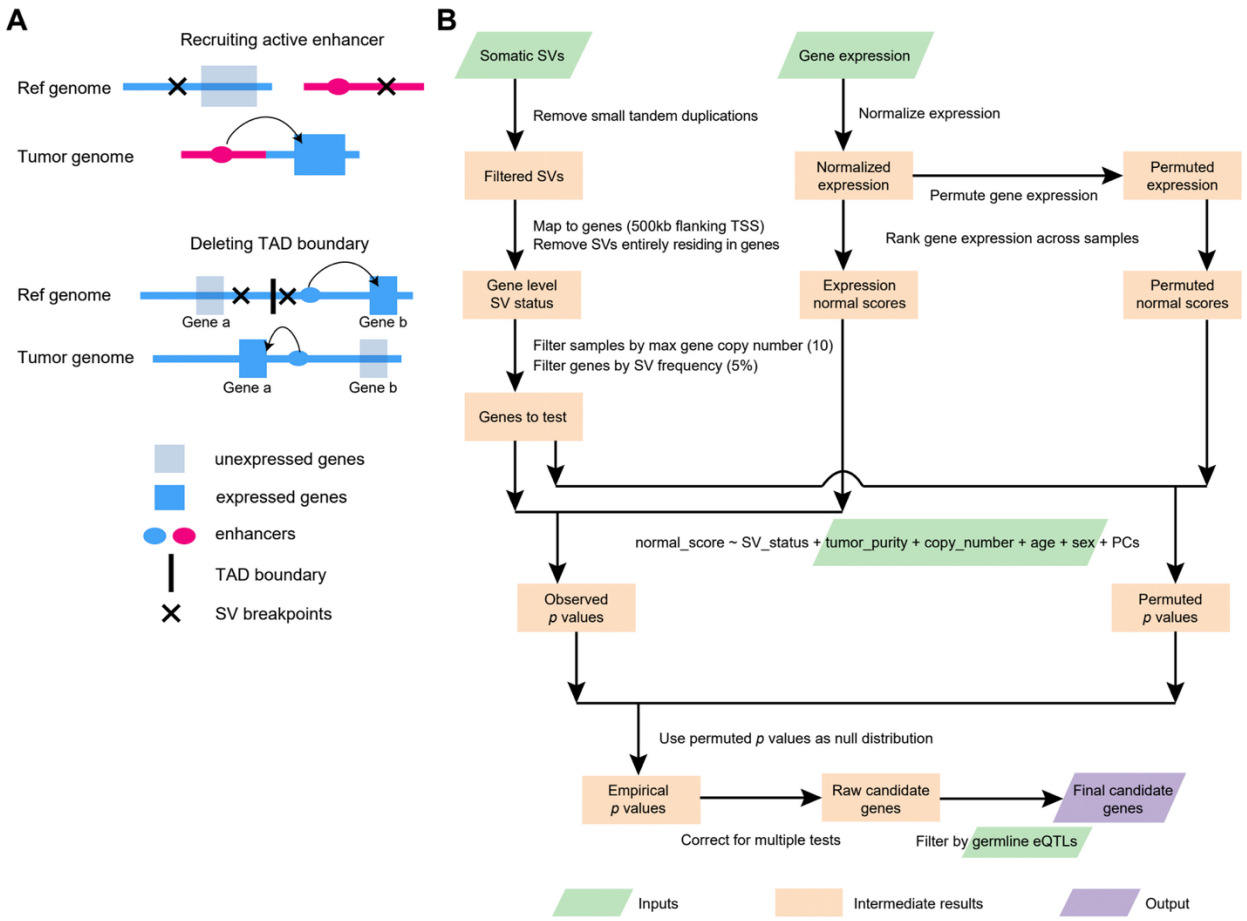

Figure 2. Outline of enhancer hijacking and HYENA algorithm.
A, Mechanisms of gene activation by SVs. SVs can activate genes by recruiting distal active enhancers (top panel) and by removing TAD boundaries and forming de novo enhancer-promoter interactions (bottom panel). **B**, HYENA workflow. Green and purple boxes denote input and output files, respectively. Orange boxes denote intermediate steps. Numbers in parentheses represent the default parameters of HYENA.

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 SVs within 500 kb of transcription start sites (**Fig. 2B**). Although promoter-enhancer interaction may occur as far as several mega-bases, mega-base-level long-range interactions are extremely rare. In addition, although duplicated enhancers can upregulate genes [135, 136], 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. 2A**). 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) [137], and ecDNA can promote accessible chromatin [138] with enhancer rewiring [139]. 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 positive associations would be detected [140]. 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 ranked the genes based on quantile-normalized expression and transformed the ranks to the quantiles of the 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 tested 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 to 1000 times (**Supplementary Table S1 Column E**) 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 testing 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. We compared HYENA's performance to two other algorithms—CESAM and PANGEA. All three algorithms were run on the same somatic SVs and gene expression data from six types of adult tumors profiled by the PCAWG (**Supplementary Table S1**): malignant lymphoma (MALY), stomach/gastric adenocarcinoma (STAD), chromophobe renal cell carcinoma (KICH), colorectal cancer (COAD/READ), thyroid cancer (THCA), and lung squamous cell carcinoma (LUSC) [21], because known enhancer hijacking genes have been reported in these tumor types (see details below). Note that PANGEA depends on promoter-enhancer interactions predicted from cell lines and such data were not available for thyroid tissue. Therefore, thyroid cancer data were not analyzed by PANGEA. To compare the performance of HYENA to the other algorithms, we used the following three strategies.

First, we used eight known enhancer hijacking target genes including *MYC* [119], *BCL2* [118], *CCNE1* [94], *TERT* [117], *IGF2* [93, 94] (in two tumor types), *IGF2BP3* [141] and *IRS4* [93] to test sensitivities. The 8 positive control genes were selected based on our literature review for genes that are both well-known as oncogenes and that are activated by distal enhancers due to restructured 3D genome organization. Out of the eight genes, HYENA detected four (*MYC*, *BCL2*, *TERT*, and *IGF2BP3*) (**Fig. 3A** and **Supplementary Fig. S1A**), CESAM detected three (*MYC*, *BCL2*, and *TERT*), and PANGEA did not detect any (**Fig. 3A**). In the five tumor types analyzed by all three algorithms, HYENA identified a total of 25 candidate genes, CESAM

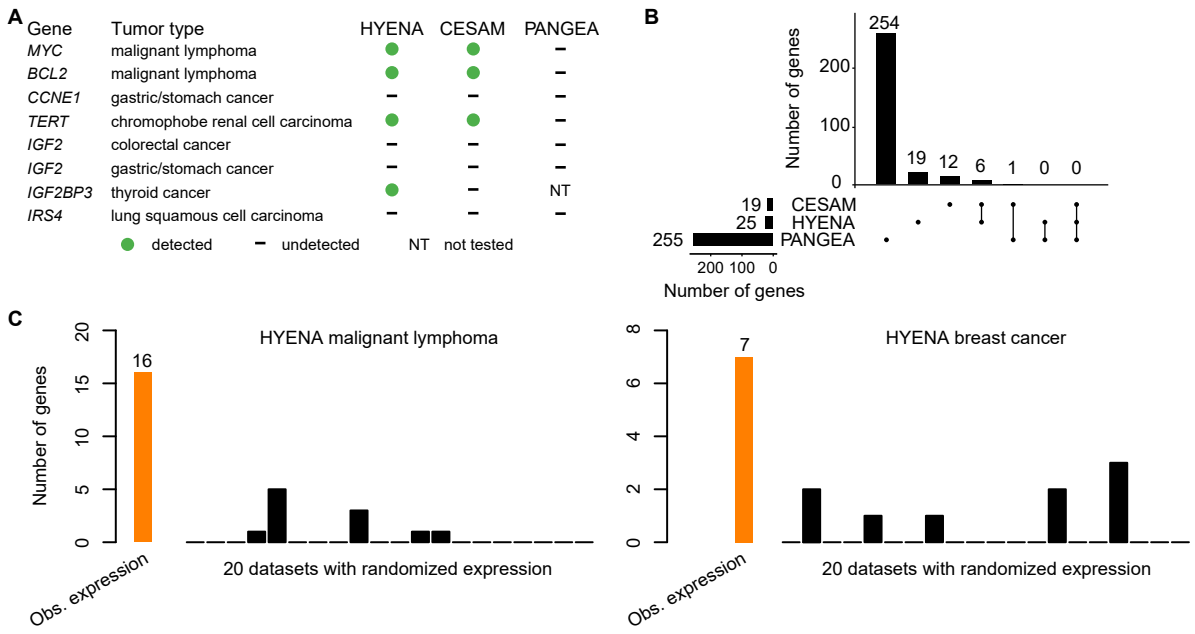

Figure 3. Benchmarking HYENA.
A, Comparison of HYENA, CESAM, and PANGEA in detecting oncogenes known to be activated by enhancer hijacking in six tumor types from the PCAWG cohort. **B**, UPSET plot demonstrating candidate genes identified and shared among the three tools in five tumor types of PCAWG. The numbers of candidate genes predicted by three algorithms are shown on the bottom left (19, 25, and 255). On the bottom right, individual dots denote genes detected by one tool, and dots connected by lines denote genes detected by multiple tools. The numbers of genes detected are shown above the dots and lines. For example, the dot immediately on the right of "PANGEA" shows there are 254 candidate genes detected only by PANGEA but not CESAM and HYENA. The left most line connecting two dots indicates that there are six genes detected by both CESAM and HYENA but not by PANGEA. **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.

identified 19, whereas PANGEA identified 255 genes (**Fig. 3B**, **Supplementary Tables S6**, **S7**, and **S8**). Six genes were detected by both HYENA and CESAM, while PANGEA had little overlap with the other algorithms (**Fig. 3B**). The ability of the algorithms to detect known target genes seems to be sensitive to sample size. Both *IGF2* and *IRS4* were initially discovered by CESAM as enhancer hijacking target genes using CNV breakpoints profiled by microarray with much larger sample sizes (378 colorectal cancers and 497 lung squamous cell carcinomas) [93]. In the PCAWG, there were far fewer samples with both WGS and RNA-Seq data available (51 colorectal cancers and 47 lung squamous cell carcinomas). Neither *IGF2* nor *IRS4* was detected

by any algorithms. *IGF2* reached the 5% SV frequency cutoff required by HYENA, however its FDR did not reach the significance cutoff (**Supplementary Fig. S1B**). In stomach/gastric adenocarcinoma, *IGF2* and *CCNE1* were identified as enhancer hijacking target genes in a cohort of 208 samples [94]. Neither of these genes was detected by any algorithms because there were only 29 stomach tumors in the PCAWG. Therefore, known target genes missed by HYENA were likely due to the small sample size. In summary, HYENA had the best sensitivity of the three algorithms.

Second, we also expect immunoglobulin genes to be detected as enhancer hijacking candidates in B-cell lymphoma due to V(D)J recombination. 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. Of the 16 genes detected by HYENA in malignant lymphoma (B-cell derived Burkitt lymphomas [142]), there were two immunoglobulin light chain genes from the lambda cluster (*IGLC7* and *IGLJ7*) and an immunoglobulin-like gene *IGSF3* (**Supplementary Table S6**). CESAM detected 11 genes, one of which was an immunoglobulin gene (*IGLC7*) (**Supplementary Table S7**). In contrast, PANGEA detected 30 candidate genes, but none were immunoglobulin genes (**Supplementary Table S8**). These data further support HYENA as the algorithm with the best sensitivity among the three algorithms.

Third, to evaluate the specificity of the algorithms, we ran each algorithm on 20 datasets generated by randomly shuffling gene expression data in both MALY and breast cancer (BRCA). 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 S6, S7, and S8**). In the 20 random gene expression datasets for malignant lymphoma, HYENA detected an average of 0.55 genes per dataset (**Fig. 3C**), and CESAM detected an average of 0.5 genes per dataset, whereas PANGEA detected an average of 40 genes per dataset (**Supplementary Fig. S2**). In breast cancer with observed gene expression, HYENA, CESAM, and PANGEA detected 7, 9, and 2,309 candidate genes, respectively (**Supplementary Tables S6, S7, and S8**). In 20 random gene expression datasets for breast cancer, HYENA, CESAM, and PANGEA detected 0.45, 0.9, and 2,296 genes on average (**Fig. 3C** and **Supplementary Fig. S2**). 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. S2**). 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 genes. Although the performances of CESAM were similar to HYENA, the genes detected by HYENA and CESAM in the six benchmarking tumor types had little overlap (**Fig. 3B**). We performed extensive validation on one gene detected only by HYENA.

Enhancer hijacking candidate genes in the PCAWG

We used HYENA to analyze a total of 1,146 tumors across 25 tumor types in the PCAWG with both WGS and RNA-Seq data. When each tumor type was analyzed individually, we identified 108 candidate enhancer hijacking target genes in total (**Supplementary Tables S1** and **S6**), four of which were known enhancer hijacking targets (**Fig. 4A**). *TERT* was detected in kidney cancers both from the US cohort (KICH) and the European cohort (RECA) which further

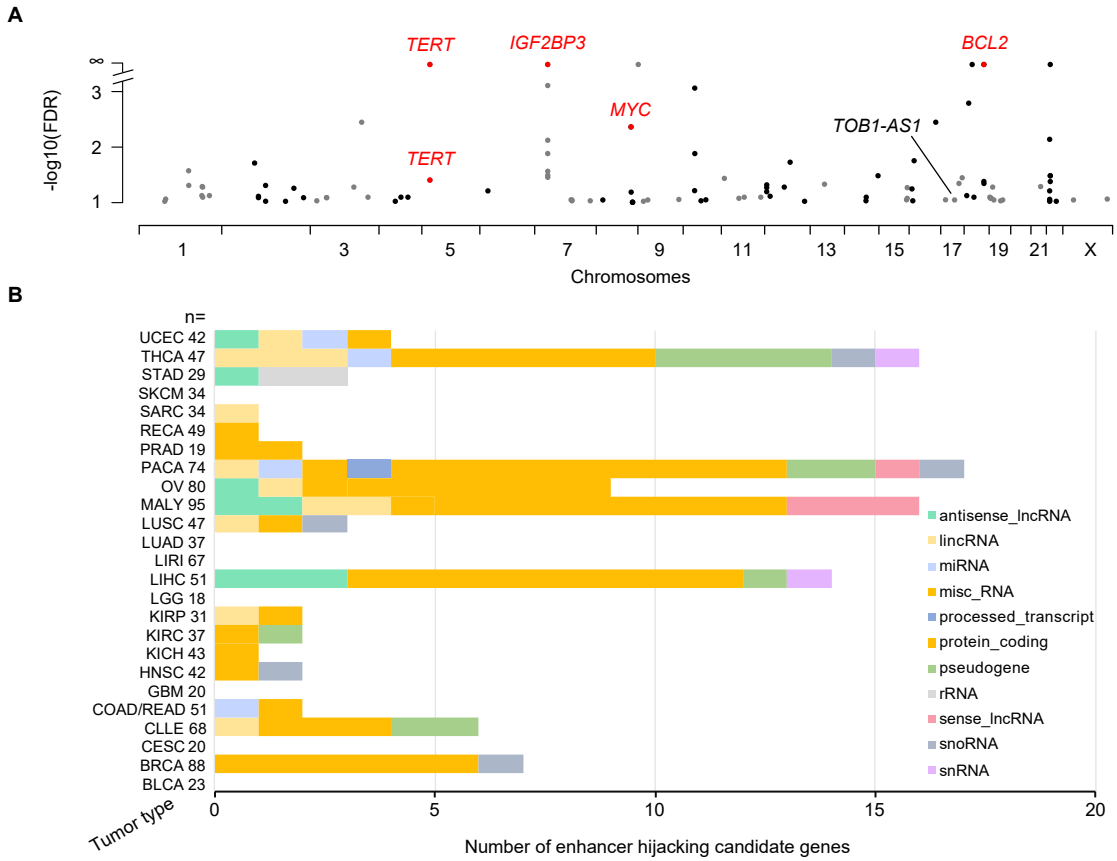

Figure 4. Enhancer hijacking candidate genes in PCAWG.
A, Candidate genes detected by HYENA in individual tumor types of PCAWG. *TERT* is plotted twice since it is detected in two cancer types. Genes labeled as red are known enhancer hijacking targets. **B**, Diverse types of candidate genes identified by HYENA in PCAWG. Numbers after tumor type names denote sample size in the corresponding tumor types.

demonstrated the reproducibility of HYENA. All other candidate genes were only detected in one tumor type, highlighting the high tumor type specificity of the findings. The number of genes detected in each tumor type differed dramatically (**Fig. 4B**) and was not associated with the level of genome instability (**Supplementary Fig. S3**). No genes were detected in bladder cancer (BLCA), cervical cancer (CESC), glioblastoma multiforme (GBM), or low-grade glioma (LGG), probably due to their small sample sizes. Pancreatic cancer (PACA) had the greatest number of candidate genes. There were two liver cancer cohorts with comparable sample sizes—LIHC from the US and LIRI from Japan. Interestingly, a total of 14 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 cancer in Japan [143], and virus integration into the tumor genome can result in oncogene activation [144]. In Chronic Lymphocytic Leukemia (CLLE), a total of six genes were detected, and three were immunoglobulin genes from both the lambda and kappa clusters (**Supplementary Table S6**). Given that sample size and genome instability can only explain a small fraction of the variations of enhancer hijacking target genes detected in different tumor types, the landscape of enhancer hijacking in cancer seems to be mainly driven by the underlying disease biology. The candidate protein-coding genes were enriched for oncogenes annotated by Cancer Gene Census [145] and OncoVar [146] (**Supplementary Table S6**, P=0.001 and 0.039 respectively by one-sided Fisher's exact test). Intriguingly, out of the 108 candidate genes, 54 (50%) were non-coding genes including lncRNAs and microRNAs (**Fig. 4B**).

Neo-TADs formed through somatic SVs

Next, we focused on the most frequently altered candidate non-coding enhancer-hijacking target gene in pancreatic cancer: *TOB1-AS1* (**Fig. 5A**), a lncRNA. *TOB1-AS1* was not detected as a candidate gene by either CESAM (**Supplementary Table S7**) or PANGEA (**Supplementary Table S8**) using the same input data. Seven (9.6%) out of 74 tumors had some form of somatic SVs near *TOB1-AS1* including translocations, deletions, inversions, and tandem duplications (**Fig. 5B** and **Supplementary Table S9**). For example, tumor 9ebac79d-8b38-4469-837e-b834725fe6d5 had a translocation between chromosomes 17 and 19 (**Fig. 5C**). The breakpoints were upstream of *TOB1-AS1* and upstream of *UQCRFS1* (**Fig. 5D**). In tumor

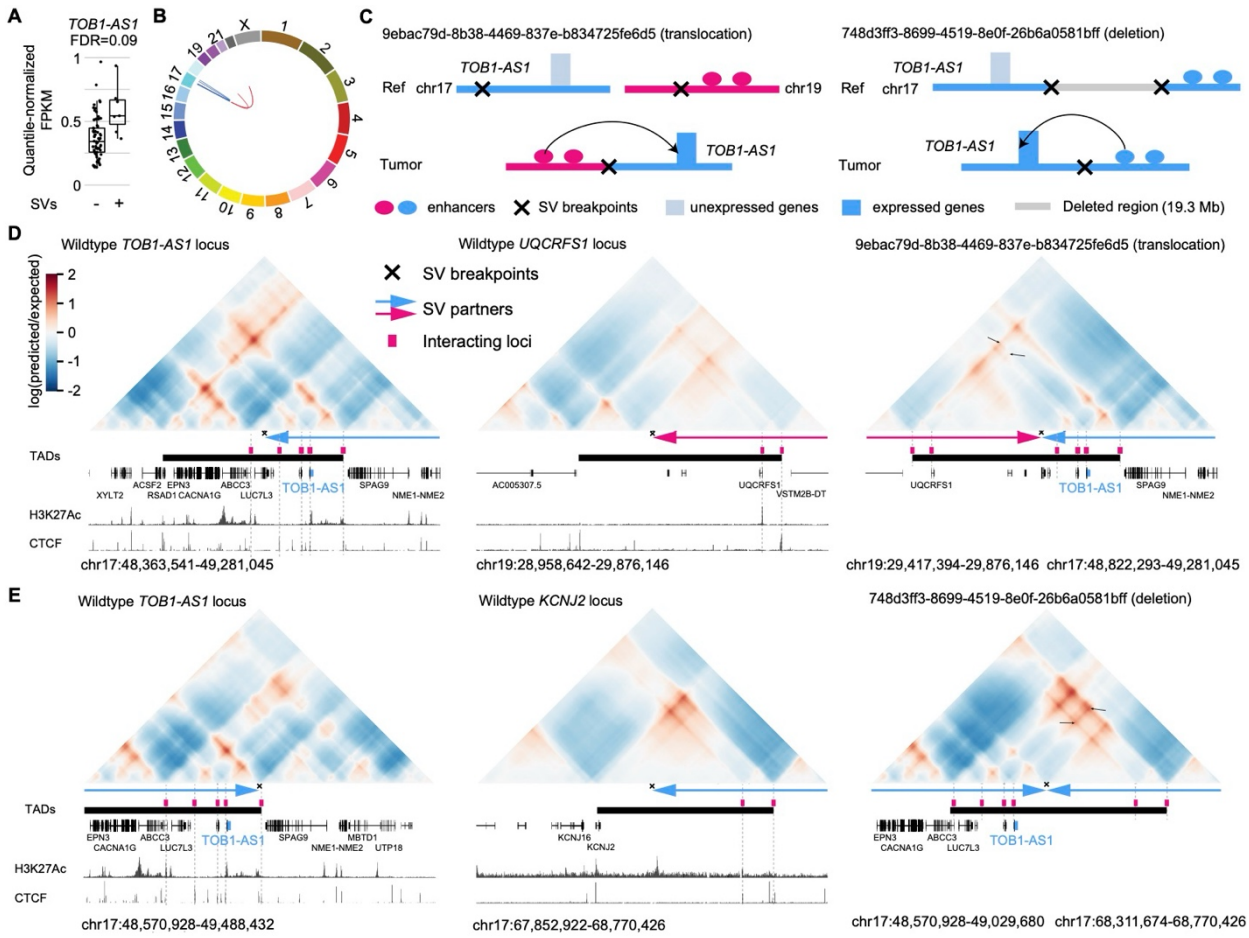

Figure 5. *TOB1-AS1* activated by various types of SVs in pancreatic cancer.
A, Normalized expression of *TOB1-AS1* in samples with (n=7) and without (n=66) nearby SVs in pancreatic cancers. The boxplot shows median values (thick black lines), upper and lower quartiles (boxes), and 1.5× interquartile range (whiskers). Individual tumors are shown as 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 *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. The expected level of 3D contacts depends on the linear distance between two genomic locations. Longer distances correlate with fewer contacts. Akita predicts 3D contacts based on DNA sequences. The heatmaps are showing the ratio between predicted and expected contacts. The darkest red represents regions having 100 times more contacts than expected given the distance between the regions. **E**, Predicted 3D chromatin interaction maps of *TOB1-AS1* (left panel) and *KCNJ2* (middle panel) loci without deletion as well as the same region following deletion in tumor 748d3ff3-8699-4519-8e0f-26b6a0581bff (right panel).

748d3ff3-8699-4519-8e0f-26b6a0581bff, there was a 19.3 Mb deletion which brought *TOB1-AS1* next to a region downstream of *KCNJ2* (**Fig. 5C** and **5E**).

We used Akita [128], a convolutional neural network that predicts 3D genome organization, to assess 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 across different cell-types [81]. TAD boundaries are defined locally by the presence of binding sites for CTCF, a ubiquitously expressed DNA-binding protein [81, 82], and TAD formation arises from the stalling of the cohesin-extruded chromatin loop by DNA-bound CTCF at these positions [84]. For this reason, it is expected that upon chromosomal rearrangements, normal TADs can be disrupted, and new TADs can form by relocation of TAD boundaries. This assumption has been validated with direct experimental evidence from examining the "neo-TADs" associated with SVs at different loci [147-149]. The wildtype *TOB1-AS1* locus had a TAD between a CTCF binding site in *RSAD1* and another one upstream of *SPAG9* (**Fig. 5D** and **Supplementary Fig. S4**). There were TADs spanning *UQCRFS1* and downstream of *KCNJ2* in the two partner regions (**Fig. 5D**, **5E,** and **Supplementary Fig. S4**). 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* (**Fig. 5D)**. 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* (**Fig. 5E**). In both cases, within these predicted neo-TADs, Akita predicted strong chromatin interactions involving several CTCF binding sites and H3K27Ac peaks between *TOB1-AS1* and its two SV partners (**Fig. 5D** and **5E** black arrows in the right panels), indicating newly formed promoter-enhancer interactions. In the vicinity of the *TOB1-AS1* locus, *TOB1-AS1* was the only

gene with significant changes in gene expression. Similar neo-TADs could be observed in two additional tumors (**Supplementary Fig. S5**). In two tumors harboring tandem duplications of *TOB1-AS1* of 317 kb and 226 kb, the *TOB1-AS1* TADs were expanded (**Supplementary Fig. S6A**). However, not all SVs near *TOB1-AS1* led to alterations in TAD architecture; for example, in tumor a3edc9cc-f54a-4459-a5d0-097879c811e5, *TOB1-AS1* was predicted to remain in its original TAD after a 4 Mb tandem duplication (**Supplementary Fig. S6B**). 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*. We then used another deep-learning algorithm called Orca [130] to predict 3D genome structure based on DNA sequences. Orca-predicted 3D genome architectures were very similar to Akita predictions (**Supplementary Fig. S7**) in neo-TAD formation due to SVs in the *TOB1-AS1* locus.

To further study the 3D genome structure of the *TOB1-AS1* locus, we performed high-resolution in situ Hi-C sequencing for four pancreatic cancer cell lines. Among these, two cell lines (Panc 10.05 and PATU-8988S) had high expression of *TOB1-AS1*, whereas the other two (PANC-1 and PATU-8988T) had low expression (**Fig. 6A**). At the mega-base-pair scale, three cell lines (Panc 10.05, PATU-8988S, and PATU-8988T) carried several SVs (black arrows in **Fig. 6B**). In Panc 10.05, a tandem duplication (chr17:43,145,000-45,950,000) was observed upstream of *TOB1-AS1* (**Fig. 6B** black arrow in the left most panel and **Supplementary Table S10**). However, the breakpoint was too far away (2 Mb) from *TOB1-AS1* (chr17:48,944,040-48,945,732) and unlikely to regulate its expression. A neo chromatin loop was detected by NeoLoopFinder [125] near *TOB1-AS1* (chr17:34,010,000-48,980,000) driven by a deletion (chr17:34,460,000-47,450,000) detected by EagleC [131] (**Supplementary Fig. S8A**, **Supplementary Tables S5** and **S10**). The deletion breakpoint was also too far away (1.5 Mb)

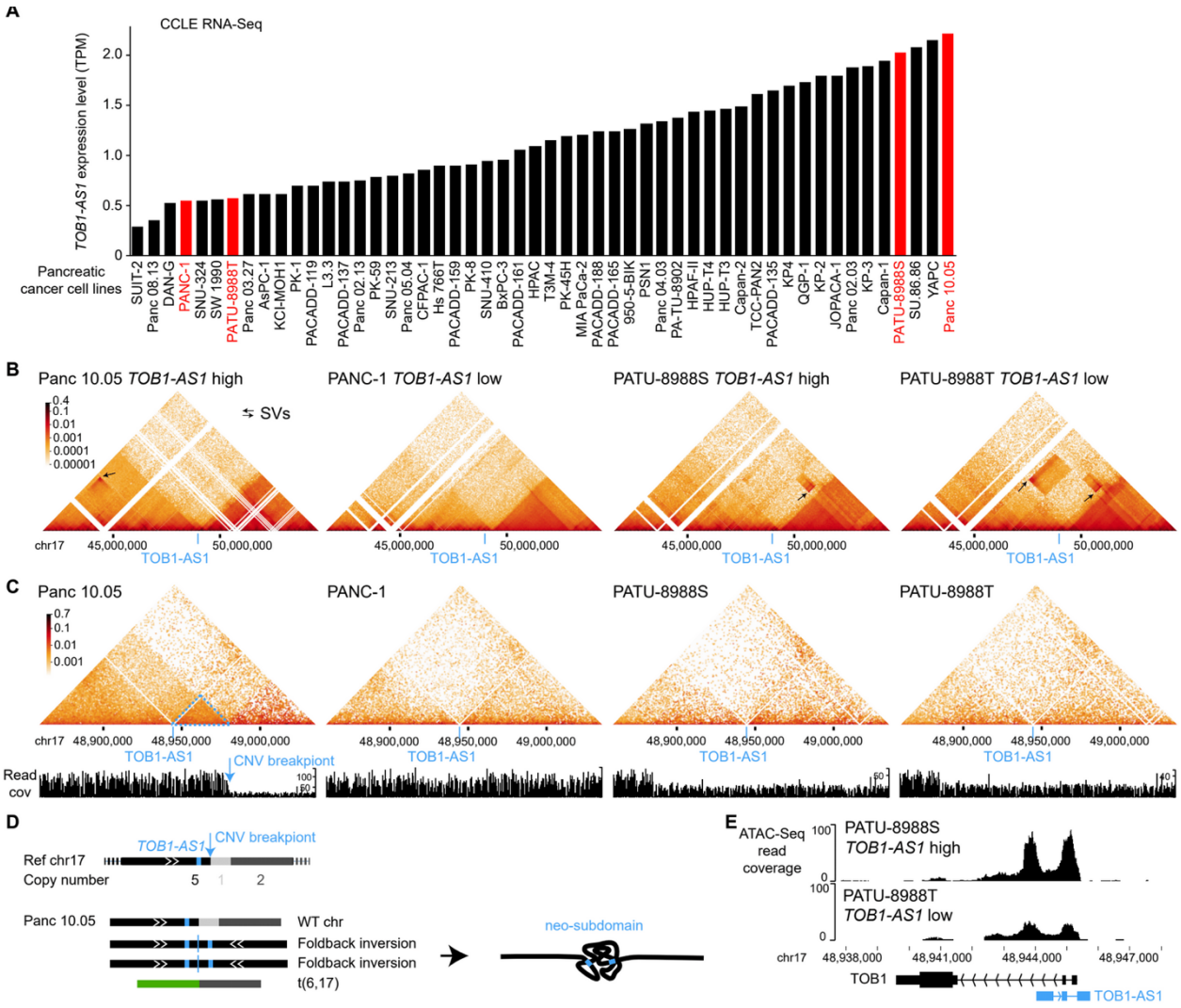

Figure 6. 3D genome structures in the *TOB1-AS1* locus in pancreatic cancer cell lines. **A**, *TOB1-AS1* expression in pancreatic cancer cell lines in CCLE. The cell lines in red are selected for further studies. **B** and **C**, 3D genomic interactions in four pancreatic cancer cell lines. Black arrows represent SVs with off-diagonal interactions. The locations of *TOB1-AS1* are marked by blue lines. In Panc 10.05, the blue arrow points to the CNV breakpoint and the dashed blue triangle represents the neo-subdomain formed due to the foldback inversion. **D**, The reference chromosome 17 and derived chromosomes in Panc 10.05. The chromosomes are not to scale. *TOB1-AS1* is shown as small blue boxes in the chromosomes. **E**, Open chromatin measured by ATAC-Seq in PATU-8988S and PATU-8988T at the *TOB1-AS1* locus.

from *TOB1-AS1* and unlikely to regulate its expression. No other SVs or neo chromatin loops were detected near *TOB1-AS1* (**Supplementary Tables S5** and **S10**). Interestingly, there was a CNV breakpoint (chr17:48,980,000) 36 kb downstream of *TOB1-AS1* in Panc 10.05 (**Fig. 6C** left most panel) which was also the boundary of the neo chromatin loop. In the high copy region

(upstream of the CNV breakpoint), heterozygous SNPs were present with allele ratios of approximately 4:1 (**Supplementary Fig. S9A**), whereas in the low copy region (downstream of the CNV breakpoint), all SNPs were homozygous (**Supplementary Fig. S9B**). These data suggested that the DNA copy number changed from five copies to one copy at the CNV breakpoint. The gained copies must connect to some DNA sequences since there should not be any free DNA ends other than telomeres. Given that no off-diagonal 3D genome interactions were observed at chr17:48,980,000, we considered the possibility that the high copy region was connected to repetitive sequences or to sequences that were not present in the reference genome. If so, reads mapped to the high copy region should have an excessive amount of non-uniquely mapped mates or unmapped mates. However, this was not the case (**Supplementary Fig. S10**). The only possible configuration was a foldback inversion in which two identical DNA fragments from the copy gain region were connected head to tail (**Fig. 6D** bottom left panel). As a result, in Panc 10.05, there was a wildtype chromosome 17, two foldback-inversion-derived chromosomes, and a translocation-derived chromosome (**Fig. 6D** bottom left panel and **Supplementary Fig. S8B**). Foldback inversions are very common in cancer. If DNA double strand breaks are not immediately repaired, following replication, the two broken ends of sister chromatids can self-ligate head to tail and sometimes result in dicentric chromosomes [150, 151]. Algorithms, such as hic-breakfinder [152] and EagleC [131], rely on off-diagonal 3D genomic interactions in the Hi-C contact matrix to detect SVs. However, foldback inversions do not form any off-diagonal interactions since the two connected DNA fragments have the same coordinates, so they are not detectable by existing algorithms. The 3D genome structure of the *TOB1-AS1* locus in Panc 10.05 was quite distinct from the other three cell lines (**Fig. 6C**). The region immediately involved in the foldback inversion had homogeneous 3D interactions (**Fig.**

6C dashed blue triangle in the left most panel) suggesting that a neo-subdomain was formed (**Fig. 6D** right panel). The high expression of *TOB1-AS1* in Panc 10.05 was likely a combined effect of the copy gain and the neo-subdomain. In PATU-8988S and PATU-8988T, a shared SV (chr17:48,880,000-52,520,000) near *TOB1-AS1* was detected (**Fig. 6B** two right panels) since the two cell lines were derived from the same pancreatic cancer patient [51]. This shared SV could not regulate *TOB1-AS1* because it pointed away from *TOB1-AS1* (**Supplementary Fig. S11**). No other SVs were found near *TOB1-AS1* in these two cell lines. The high expression of *TOB1-AS1* in PATU-8988S was likely due to transcriptional regulation since the promoter of *TOB1-AS1* in PATU-8988S was more accessible than that in PATU-8988T (**Fig. 6E**). This result was consistent with a handful of patient tumors that had high expression of *TOB1-AS1* without any SVs (**Fig. 6A**).

Taken together, our results demonstrated that *TOB1-AS1*, a candidate enhancer hijacking gene detected by HYENA, is 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 [153, 154]. However, HYENA predicted it to be an oncogene in pancreatic cancers. To test the potential oncogenic functions of *TOB1-AS1* in pancreatic cancer, we performed both in vitro and in vivo experiments. We surveyed pancreatic cancer cell line RNA-Seq data from Cancer Cell Line Encyclopedia (CCLE) and identified that the commonly transcribed isoform of *TOB1-AS1* in pancreatic cancers was ENST00000416263.3 (**Supplementary Fig. S12**). 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* (**Fig. 6A** and

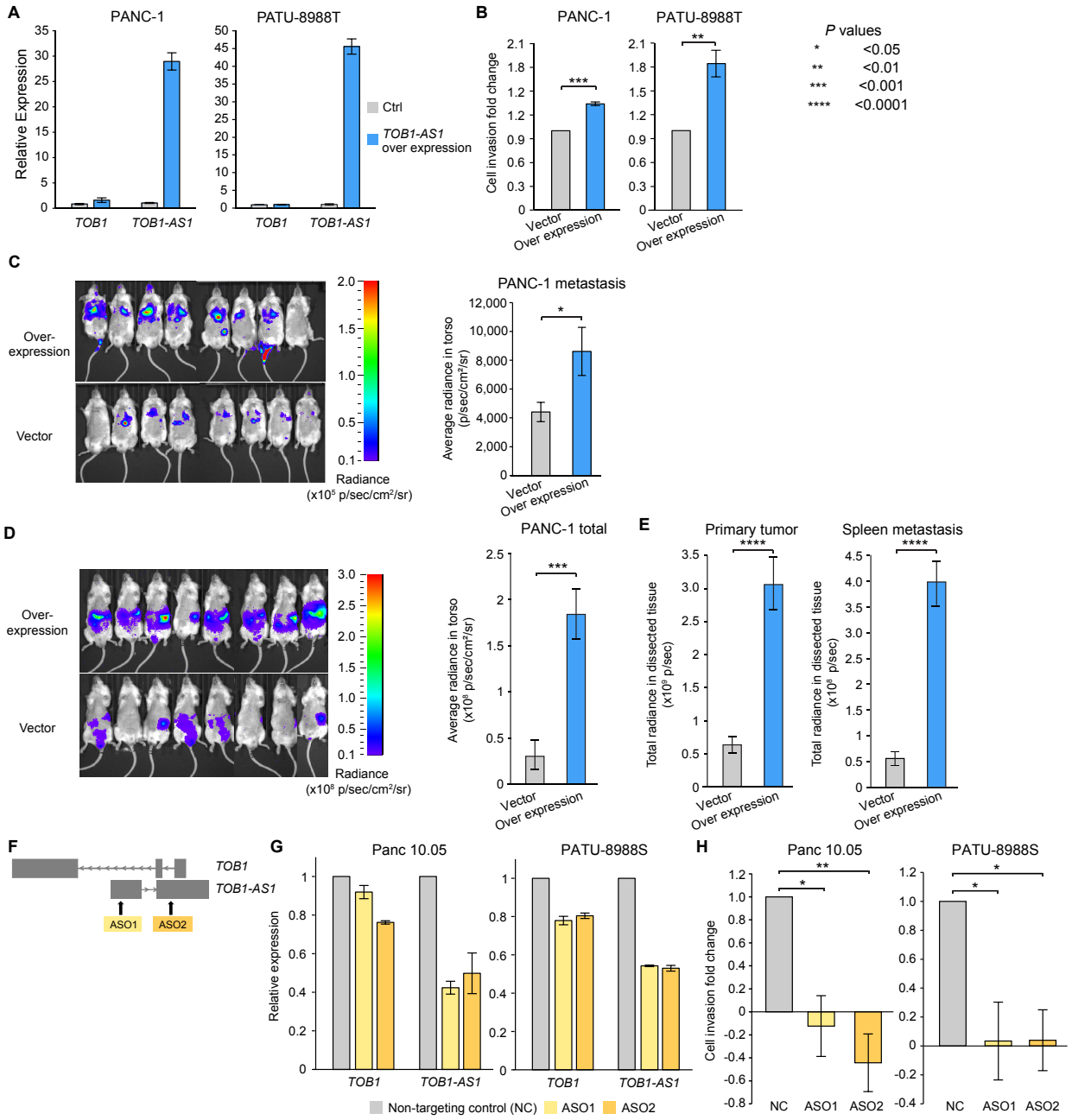

Figure 7. *TOB1-AS1* promotes cell invasion and tumor metastasis. (Legends on next page)

Supplementary Fig. S13A). In both cell lines, overexpression of *TOB1-AS1* (**Fig. 7A**) promoted in vitro cell invasion (**Fig. 7B**). 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. 7C**). Six weeks after orthotopic injection, mice carrying *TOB1-AS1*

overexpressing PANC-1 cells showed exacerbated overall tumor burden (**Fig. 7D**), elevated primary tumor burden, and elevated metastatic burden in the spleen (**Fig. 7E** and **Supplementary Fig. S13B**). Liver metastasis was not affected (**Supplementary Fig. S13C**). In addition, we knocked down *TOB1-AS1* in two other pancreatic cancer cell lines Panc 10.05 and PATU-8988S, both of which had high expression of *TOB1-AS1* (**Fig. 6A** and **Supplementary Fig. S13A**), using two antisense oligonucleotides (ASOs) (**Fig. 7F**). *TOB1-AS1* expression was reduced by approximately 50% by both ASOs (**Fig. 7G**). Knockdown of *TOB1-AS1* substantially suppressed cell invasion in vitro (**Fig. 7H**). Note that PATU-8988T and PATU-8988S were derived from the same liver metastasis of a pancreatic cancer patient, and they had drastic differences in *TOB1-AS1* expression (**Fig. 6A** and **Supplementary Fig. S13A**). It was reported that PATU-8988S can form lung metastases in vivo with tail vein injection of nude mice,

Figure 7. *TOB1-AS1* promotes cell invasion and tumor metastasis.
A, *TOB1-AS1* and *TOB1* relative expression levels in PATU-8988T and PANC-1 cells transduced with *TOB1-AS1* overexpression vector (n=3) or control vector (n=3). **B**, *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 was subtracted from each technical replicate. *P* values were calculated by two-sided student t test. **C**, *TOB1-AS1* overexpression in PANC-1 cells promoted in vivo tumor metastasis in the tail vein injection model. **D**, *TOB1-AS1* overexpression in PANC-1 cells exacerbated in vivo tumor growth and spontaneous metastasis in the orthotopic tumor model. 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 the overexpression group and the 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). **E**, Primary tumor burden and spleen metastatic burden were higher in the mice that were orthotopically injected with *TOB1-AS1* overexpression PANC-1 cells. The bar plots show quantified total radiance with a set area (in p/sec). **F**, Targeting *TOB1-AS1* by two ASOs. **G**, *TOB1-AS1* knockdown in Panc 10.05 and PATU-8988S cells transduced with ASO1 (n=3), ASO2 (n=3) or non-targeting control ASO (NC) (n=3). **H**, *TOB1-AS1* knockdown suppressed Panc 10.05 (3 biological replicates) and PATU-8988S (3 biological replicates) cell invasion in vitro. Cell invasion fold change calculation is the same as in **B**. Two-sided student t test was used. Error bars in all panels indicate standard error of the mean.

whereas PATU-8988T cannot form any metastases in any organ [155]. By altering the expression of *TOB1-AS1*, we were able to reverse the cell invasion phenotypes in these two cell lines (**Fig. 7B** and **7H**). These results suggested that *TOB1-AS1* has an important function in regulating cell invasion.

It is possible that *TOB1-AS1*, as an anti-sense lncRNA, transcriptionally regulates the expression of the sense protein-coding gene *TOB1*. However, we did not find consistent correlations between *TOB1-AS1* and *TOB1* expression in different pancreatic cancer cohorts and pancreatic cancer cell lines (**Supplementary Fig. S13D**). Hence, it is unlikely that *TOB1-AS1* functions through transcriptional regulation of *TOB1*. Although knocking down *TOB1-AS1* resulted in down regulation of *TOB1* expression, this is an expected result given that the ASOs also targeted the introns of *TOB1* (**Fig. 7F**). The decrease in *TOB1* expression was relatively mild at 10-20% (**Fig. 7G**). Overexpression of *TOB1-AS1* did not have a major impact on *TOB1* expression (**Fig. 7A**). Therefore, the oncogenic functions of *TOB1-AS1* that we observed in vitro and in vivo are likely independent of *TOB1*. To gain further insights into the pathway that *TOB1-AS1* is involved in and its downstream targets, we performed RNA-Seq on PANC-1-generated mouse tumors with *TOB1-AS1* overexpression and found that the most significantly differentially expressed gene was *CNNM1* (**Supplementary Fig. S13E**). No significantly enriched pathway was detected. *CNNM1* is a cyclin and CBS domain divalent metal cation transport mediator and is predicted to be involved in ion transport [156]. How *TOB1-AS1* promotes cell invasion and tumor metastasis and whether *CNNM1* plays a role require further study.

Our results showed that the lncRNA *TOB1-AS1* is oncogenic and has a pro-metastatic function in pancreatic cancer, and that 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 regulatory regions of the genome and caused shuffling of regulatory elements, altering 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, and HYENA aims to identify oncogenes recurrently activated by somatic SVs since these events are under positive selection. Therefore, sample size is a major limiting factor. Of the eight ground truth cases, HYENA only detected four (**Fig. 3A**); undetected genes were likely due to small sample size. However, genes detected in individual tumors by tools such as cis-X and NeoLoopFinder may not be oncogenes, and recurrent events would be required to identify candidate oncogenes.

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 sex, age, and principal components of gene expression. In addition, it is also possible that the high gene expression caused somatic SVs. Open chromatin and double helix regions unwound 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. Although *TOB1-AS1* has been reported as a tumor suppressor in several tumor types [153, 154], it promotes cell invasion and metastasis in pancreatic cancer, which suggests that the functions of lncRNA *TOB1-AS1* depend on cell lineage. Furthermore, most enhancer hijacking candidate genes detected by HYENA are only found in one tumor type. This further supports the tumor-type-specific roles of these potential oncogenes.

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) [128] 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. S4**) in genome organization. For example, the left boundary of the TAD at the *UQCRFS1* locus was different between HFF and H1hESC (**Supplementary Fig. S4A**). Nonetheless, the translocation between chromosomes 17 and 19 removed the left boundary and merged the right side of the *UQCRFS1* TAD with the *TOB1-AS1* TAD (**Fig. 5D**). Therefore, the cell-type difference likely does not have a major impact on our results.

HYENA includes multiple parameters including the SV mapping window. In the analysis, SV breakpoints were mapped to individual genes if located within 500kb up- or down-stream of the gene TSS, with the assumption that most enhancer-promoter interactions happen within this range. However, this window might not be suitable to detect all the interactions between gene promoters and enhancers. To adjust this window to a proper range for each sample cohort, it would be helpful to have some known enhancer hijacking events in the corresponding

tumor type, so users can adjust the range to where HYENA can identify the known genes as significant. In this way, the SV mapping window would better help the discovery of new enhancer hijacking genes in a specific sample cohort.

HYENA is a discovery platform based on computational analysis, which means not only the candidate gene functions need further research, but the SVs and rearranged enhancers need to be validated in experimental models as well to confirm the analysis results. Epigenetic studies should be done to confirm the activated enhancers at SV partner regions, and the new enhancer-promoter interactions should be supported by 3D genome architecture data in patient samples carrying the SV, or in a model where SVs can be engineered.

Random indels induced by Cas9 with single sgRNAs are usually not enough to generate a desired SV. Engineering a large DNA fragment could be achieved by Cas9 reprogrammed with dual sgRNAs, which would generate two concurrent double-strand breaks (DSBs) in a genome. With the participation of cellular DNA repair proteins, the four DSB ends generated by the two Cas9 cleavages are randomly ligated, resulting in DNA fragment deletion or inversion when concurrent DSBs occur on single chromosomes and DNA fragment duplication or translocation when the DSBs are on different chromosomes [157]. However, other than to engineer a deletion, to engineer other types of SVs with CRISPR/Cas9 has very low target efficacy in human cell lines [158], making it challenging to apply this technology to enhancer hijacking studies. An easier approach would be having a cancer cell line that has high expression of the gene of interest and carries the SV of interest. In this way, the first step would be investigating the gene activation mechanism in the cell line. Technologies like Hi-C can detect the genome interactions related to the gene, and the effects of the SV can be experimentally tested. If there is a stronger interaction induced by the SV, the next step would be to identify the enhancers hijacked. Besides

showing the epigenetic markers for the enhancer of interest, deleting the enhancer region, deleting the gene promoter region, or inserting a TAD boundary (e.g. a CTCF binding site) between the enhancer and the gene with CRISPR technologies can be helpful for confirming the enhancer hijacking event. If the gene expression level is significantly reduced when we disrupt the enhancer-promoter interactions, the enhancer hijacking event can be validated.

The ultimate goal for this study is to identify novel oncogenes as therapeutic targets or biomarkers for patient prognosis. After the identification of an oncogene, the question of how to target it to treat cancers follows. A straight-forward and common way to develop a target therapy is to design a drug based on the structures of proteins. If a specific mutation induces protein structural changes in cancer cells but not in normal cells, small molecules can be developed and screened to generate drug candidates. However, in the context of enhancer hijacking, the genes usually do not have a recurrent mutation in gene body that can be targeted, making it challenging to design a therapy. An alternative strategy for drug discovery is to directly modulate disease-associated enhancers. One class of proteins that is of particular interest in the context of enhancers is the bromo- and extra-terminal (BET) family [159]. Previous studies showed that JQ1, a hieno-triazolo-1,4-diazepine, which displaces BET bromodomains from chromatin by competitively binding to the acetyl lysine recognition pocket, could significantly suppress *MYCN*-amplified neuroblastoma growth [160]. It showed that bromodomain inhibition downregulated MYCN transcriptional programs in neuroblastoma, providing a new framework of targeting transcriptional machineries instead of specific proteins. To target the candidate enhancer hijacking genes, inhibition of the identified enhancer-promoter interactions might be a feasible approach to provide clinical benefits. With a specific hijacked enhancer, editing the enhancer sequence using CRISPR might be another choice. Currently, the therapy has been

pioneered in transfusion-dependent β-thalassemia and sickle cell disease [161]. Since there are currently few clinical trials that use CRISPR/Cas9 edited cells as treatment and even fewer that target enhancers, it might take longer for CRISPR technology to accumulate pre-clinical evidence in cancer treatment [160].

Identifying Novel Oncogenes Detected by HYENA with CRISPR Activation Screening

Introduction

Enhancer hijacking, as a cancer driving event caused by structural variants (SVs), has been more and more explored and identified in multiple tumor types. The progresses in bioinformatics using whole-genome sequencing (WGS), RNA-sequencing (RNA-Seq) and other sequencing technologies to profile chromatin conformations in human genome have achieved effective predictions of individual or recurrent enhancer hijacking events that drive oncogenesis [89, 96, 122]. Our previous work has presented a sensitive and reliable tool to infer novel oncogene candidates activated by genomic rearrangements and reported 108 putative oncogenes that included both coding and noncoding genes [162]. However, most of these reports included only a few oncogenes with their cancer driving functions validated in tumor models, and most of the validated genes are protein-coding genes, leaving a substantial number of candidate genes untested. Therefore, a comprehensive study that can investigate the oncogenic functions of these candidate genes is needed.

With powerful computational tools and extensive studies, many important cancer genes and how they promote cancer development have been demonstrated, but such studies are mainly limited in coding genes. The human genome contains both coding and noncoding genes, many of which are crucial for the intricate processes involved in cancer development. Thousands of unique non-coding RNA (ncRNA) sequences exist within cells. Over the past decade, research has transformed our understanding of ncRNAs from being considered 'junk' transcriptional products to recognizing them as functional regulatory molecules involved in various cellular processes, such as chromatin remodeling, transcription, post-transcriptional modifications, and

signal transduction [124, 163]. The networks in which ncRNAs operate can influence numerous molecular targets, driving specific cellular responses and determining cell fates. As key regulators of physiological programs, ncRNAs play significant roles in both developmental and disease contexts [164-166]. Therefore, gaining a deeper understanding of the cancer driving functions ncRNAs offers a unique opportunity to design more effective therapeutic interventions.

Unlike coding genes, relatively few ncRNA genes have been shown to be regulated by enhancer hijacking events. We have reported that a long non-coding RNA (lncRNA) *TOB1-AS1*, which was activated by enhancer hijacking in patients, could promote cancer cell invasion and tumor metastasis in pancreatic tumor models [162]. Other non-coding oncogenes have been identified in multiple cancer types. Examples such as SAMMSON in melanoma and lncGRS-1 in glioma have garnered attention as drug targets due to the strong and specific sensitivity of tumor cells to their inhibition via antisense oligonucleotide (ASO) therapies [167, 168]. *PVT1* and *MALAT1* are frequently overexpressed or amplified in lung tumors, and their manipulation affects cell growth and invasiveness both *in vitro* and *in vivo*, making them promising therapeutic targets [169]. Other examples include *LINC00680*, which acts by binding to GATA6 [170], and *LINC00511*, which promotes non-small cell lung cancer (NSCLC) by binding to the chromatin-modifying enzyme EZH2 and repressing tumor suppressor genes such as p57 and *LATS2* [171]. As the emerging roles of non-coding oncogenes have been more and more studied, it is imperative to explore other cancer driving ncRNAs along with coding genes that play a role in cancer.

Programmable nucleases have emerged as a powerful technology for genetic perturbation, capable of precisely recognizing and cleaving target DNA. In particular, the RNA-guided endonuclease Cas9, derived from the microbial CRISPR (clustered regularly interspaced

short palindromic repeat) immune system, has proven to be a powerful tool for precise DNA modifications [172, 173]. Cas9 is directed to specific genomic targets by short RNAs that form Watson-Crick base pairs with the DNA, making Cas9 easily retargetable. Cas9 creates precise double-strand breaks (DSBs) at target sites, which are repaired through either homology-directed repair (HDR) or, more commonly, non-homologous end-joining (NHEJ) [174]. HDR repairs the DSB accurately using a homologous DNA template, while NHEJ is error-prone and introduces insertions or deletions (indels). When Cas9 targets a coding region, loss-of-function mutations can occur due to frameshifting indels that produce a premature stop codon, leading to nonsense-mediated decay of the transcript or the creation of a non-functional protein. These characteristics make Cas9 ideal for genome editing applications [175].

In addition to generating loss-of-function mutations and indels, Cas9 can modulate transcription without altering the genomic sequence by fusing catalytically inactive Cas9 (dCas9) to transcriptional activation or repression domains [176]. CRISPR activation (CRISPRa) and CRISPR interference (CRISPRi) are achieved by direct fusion or recruitment of activation and repression domains, such as VP64 and KRAB, respectively [177, 178]. CRISPRa, in particular, offers a significant improvement as a screening platform over other activation methods. Previously, gain-of-function screens were primarily limited to cDNA overexpression libraries, which faced challenges like incomplete representation, overexpression beyond physiological levels and endogenous regulation, lack of isoform diversity, and high construction costs. CRISPRa addresses these limitations by activating gene transcription at the endogenous locus, requiring only the synthesis and cloning of RNA guides, which makes it much more cost-effective and align with our goal of studying a group of candidate genes' effects in cancer cells.

Most CRISPR-based screens have focused on the protein-coding genome, typically excluding ncRNA loci, and there are more knock-out (KO) or knock-down (KD) screens compared to activation screens [179]. Despite this, these studies offer insights into the principles of coding genome function by integrating screen data with a rich foundation of literature, including knowledge of physical and functional interaction networks. Although genetic screens targeting ncRNAs are beginning to emerge, the functional knowledge of these molecules primarily comes from studying individual ncRNAs. Genome-wide screens that incorporate data from both the coding and ncRNA genomes are rare but have been conducted in complex contexts such as cell differentiation and cancer cell proliferation, migration as well as drug resistance [180, 181]. Such comprehensive genome-wide approaches provided valuable data resources to uncover principles of normal tissue and cancer development, but there is not a study specially focused on enhancer hijacking genes.

Here we perform CRISPRa screens within a breast cancer cell line, MCF-7, and a pancreatic cancer cell line, PATU-8988T, to study the impacts of the upregulated transcription of the putative oncogenes detected by HYENA in cancer cells, to mimic the scenario of oncogene activation caused by relocated enhancers. We found that the known oncogenes *RCCD1* and *POLR2F*, as well as a number of non-coding genes could drive cancer cells to proliferate or migrate at a faster speed. By *in silico* analysis, we demonstrated that a known oncogene, *RCCD1*, was activated in PCAWG breast cancer patients by rearranged enhancers and disturbed 3D-genome interactions.

Methods

Cell culture

PATU-8988T cells were obtained from Dr. Alexander Muir (University of Chicago). MCF-7 cells were obtained from Dr. Marsha Rosner (University of Chicago). All cell lines were cultured at 37°C/5% CO_2. PATU-8988T cells were cultured with Dulbecco's Modified Eagle Medium (DMEM) containing 5% fetal bovine serum (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). The PATU-8988T cells were seeded at 0.5 x 10^6 cells/80 cm^2 and split the confluent culture 1:5 to 1:10 every 3-5 days using trypsin/EDTA. MCF-7 cells were cultured with Eagle's Minimum Essential Medium (EMEM) (ATCC 50-238-2632) containing 10% FBS and 0.01 mg/ml human recombinant insulin (Sigma Aldrich 91077C). A subcultivation ratio of 1:3 to 1:6 was done 2 to 3 times a week according to the recommendation of ATCC (American Type Culture Collection, USA) (https://www.atcc.org/products/htb-22).

All cell lines have been regularly monitored and tested negative for mycoplasma using a mycoplasma detection kit (Lonza, LT07-218).

CRISPR activation screening

Part of these methods were adapted from the publications by Joung *et al.* 2017 [182], and the methods used by Dr. Alexander Muir's lab.

CRISPR library design

The guide oligo design was done by CRISPick by Broad Institute (https://portals.broadinstitute.org/gppx/crispick/public) with Human GRCh37 reference genome, CRISPRa mechanism, SpyoCas9 and gene ID (or gene sequences for non-coding genes that could not be found with gene ID). 3 oligos were designed for each gene, and the genes that could

not be targeted properly were removed from the library. There were 112 protein coding genes, 15 antisense genes and 44 other non-coding genes in the library. The genes were predicted to be enhancer hijacking genes by HYENA 0.5.3 from PCAWG database. The IG genes or IG pseudogenes were not included in the library. 31 non-targeting guides were also included in the library, along with 9 guides targeting 3 positive control genes (*CCND1*, *ERBB2*, *PIK3CA*). In total, there were 553 guide oligos in the pool.

"TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCG" was added to the 5' end, and "TTTTAGAGCTAGGCCAACATGAGGATCACC" was added to the 3' end to the designed guide oligos to generate the customized CRISPR activation library.

The library was synthesized by Twist Biosciences (https://www.twistbioscience.com/). A full list of the library was in Appendix.

CRISPR library PCR

The 25µl reaction included 12.5µl NEBNext High Fidelity PCR Master Mix (NEB M0541S) to make a final concentration of 1x, pooled oligo library template at a final concentration of 0.04ng/µl, primers (Fwd: 5' -GTAACTTGAAAGTATTTCGATTTCTT GGCTTTATATATCTTGTGGAAAGGAC GAAACACC- 3', Rev: 5' - ATTTTAACTTGCTAGGCCCTGCAGAC ATGGGTGATCCTCATGTTGGCCTAGC TCTAAAAC- 3') at a final concentration of 0.5 µM each, and pure water to reach the final volume of 25µl. Cycling conditions were set as following: cycle 1) 98 °C 30 s; cycle 2-21) 98 °C 10 s, 63 °C 10 s, 72 °C 15 s; cycle 22) 72 °C, 2 min.

The PCR product was pooled and purified with QIAquick PCR Purification Kit (Qiagen 28104) according to the manufacturer's directions. The purified product was run on a gel along with a 50-bp ladder (Thermo Fisher Scientific 10416014): cast a 2% (wt/vol) agarose

gel in TBE buffer (Thermo Fisher Scientific 15581028) with SYBR Safe DNA dye (Thermo Fisher Scientific S33102). Run half of the oligo library in the gel at 15 V cm^{-1} for 45 min. Gel was extracted to get the purified PCR product (140bp) using the QIAquick Gel Extraction Kit (Qiagen 28704) according to the manufacturer's directions.

Library cloning

Restriction digest of plasmid backbone with the restriction enzyme Esp3I (BsmBI), which cuts around the single guide (sgRNA) target region. The plasmid backbone lenti_SAMv2_Puro was from Dr. Alexander Muir and available at AddGene (Plasmid 75112). After running a gel and extracted, the linear backbone was ready for Gibson Assembly with the oligo pool. The Gibson Assembly reactions were set up by each 20µl Gibson reaction according to the reaction ratios, including Gibson Assembly Master Mix 2× 10µl, digested library plasmid backbone from 330ng, sgRNA library insert 50ng and UltraPure water up to 20µl. After isopropanol precipitation, the plasmid library was electroporated into 100µl MegaX DH10B cells (Invitrogen C640003) at 2.0 kV, 200 ohms, 25 µF, for maxi-prep.

Next-generation sequencing of the amplified sgRNA library

To amplify the sgRNA cassette, PCR was done with the plasmid DNA as input. Reactions

were prepared on ice with 25 µL NEBNext High Fidelity PCR Master Mix (NEB M0541S) to make a final concentration of 1x, pooled oligo library template at a final concentration of 0.4ng/µl, primers (Fwd: a pool of ten forward primers for sequencing purpose listed below, Rev: 5' -CAAGCAGAAGACGGCATACGAGATTC GCCTTGGTGACTGGAGTTCAGACGTG TGCTCTTCCGATCTGCCAAGTTGATAA CGGACTAGCCTT- 3') at a final concentration of 0.25 µM each, and pure water to reach the final volume of 50µl. Cycling conditions were set as

following: cycle 1) 98 °C 3 min; cycle 2-21) 98 °C 10 s, 63 °C 10 s, 72 °C 25 s; cycle 22) 72 °C, 2 min. All PCR products were pooled and mixed thoroughly by pipetting. Illumina NextSEQ 500 was used for sequencing by the Genomics Core at the University of Chicago. The sample is low complexity and low nucleotide diversity (a CRISPR library with less than 600 different guides). A 20% PhiX control was applied to improve library diversity. 80 cycles of read 1 (forward) and 8 cycles of index 1 was used.

Primer sequences (5'-3') (Rev primers have barcodes bolded):

NGS-Lib-Fwd-1 AATGATACGGCGACCACCGAGATCTA CACTCTTTCCCTACACGACGCTCTTCC GATCTTAAGTAGAGGCTTTATATATCT TGTGGAAAGGACGAAACACC

NGS-Lib-Fwd-2 AATGATACGGCGACCACCGAGATCTA CACTCTTTCCCTACACGACGCTCTTCC GATCTATCATGCTTAGCTTTATATATC TTGTGGAAAGGACGAAACACC

NGS-Lib-Fwd-3 AATGATACGGCGACCACCGAGATCTA CACTCTTTCCCTACACGACGCTCTTCC GATCTGATGCACATCTGCTTTATATAT CTTGTGGAAAGGACGAAACACC

NGS-Lib-Fwd-4 AATGATACGGCGACCACCGAGATCTA CACTCTTTCCCTACACGACGCTCTTCC GATCTCGATTGCTCGACGCTTTATATA TCTTGTGGAAAGGACGAAACACC

NGS-Lib-Fwd-5 AATGATACGGCGACCACCGAGATCTA CACTCTTTCCCTACACGACGCTCTTCC GATCTTCGATAGCAATTCGCTTTATAT ATCTTGTGGAAAGGACGAAACACC

NGS-Lib-Fwd-6 AATGATACGGCGACCACCGAGATCTA CACTCTTTCCCTACACGACGCTCTTCC GATCTATCGATAGTTGCTTGCTTTATA TATCTTGTGGAAAGGACGAAACACC

NGS-Lib-Fwd-7 AATGATACGGCGACCACCGAGATCTA CACTCTTTCCCTACACGACGCTCTTCC GATCTGATCGATCCAGTTAGGCTTTAT ATATCTTGTGGAAAGGACGAAACACC

NGS-Lib-Fwd-8 AATGATACGGCGACCACCGAGATCTA CACTCTTTCCCTACACGACGCTCTTCC GATCTCGATCGATTTGAGCCTGCTTTA TATATCTTGTGGAAAGGACGAAACAC C

NGS-Lib-Fwd-9 AATGATACGGCGACCACCGAGATCTA CACTCTTTCCCTACACGACGCTCTTCC GATCTACGATCGATACACGATCGCTTT ATATATCTTGTGGAAAGGACGAAACA CC

NGS-Lib-Fwd-10 AATGATACGGCGACCACCGAGATCTA CACTCTTTCCCTACACGACGCTCTTCC GATCTTACGATCGATGGTCCAGAGCTT TATATATCTTGTGGAAAGGACGAAAC ACC

NGS-Lib-SAM-Rev-1 CAAGCAGAAGACGGCATACGAGAT **TCGCCTTG** GTGACTGGAGTTCAGACGTG TGCTCTTCCGATCTGCCAAGTTGATAA CGGACTAGCCTT

NGS-Lib-SAM-Rev-2 CAAGCAGAAGACGGCATACGAGAT **ATAGCGTC** GTGACTGGAGTTCAGACGTG TGCTCTTCCGATCTGCCAAGTTGATAA CGGACTAGCCTT

NGS-Lib-SAM-Rev-3 CAAGCAGAAGACGGCATACGAGAT **GA AGAAGT** GTGACTGGAGTTCAGACGTG TGCTCTTCCGATCTGCCAAGTTGATAA CGGACTAGCCTT

NGS-Lib-SAM-Rev-4 CAAGCAGAAGACGGCATACGAGAT **ATTCTAGG** GTGACTGGAGTTCAGACGTG TGCTCTTCCGATCTGCCAAGTTGATAA CGGACTAGCCTT

NGS-Lib-SAM-Rev-5 CAAGCAGAAGACGGCATACGAGAT **CGTTACCA** GTGACTGGAGTTCAGACGTG TGCTCTTCCGATCTGCCAAGTTGATAA CGGACTAGCCTT

NGS-Lib-SAM-Rev-6 CAAGCAGAAGACGGCATACGAGAT **GTCTGATG** GTGACTGGAGTTCAGACGTG TGCTCTTCCGATCTGCCAAGTTGATAA CGGACTAGCCTT

Lentiviral transduction

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 lenti_SAM_v2 library plasmid 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 lenti-MS2 vector was from Dr. Alexander Muir and available at AddGene (Plasmid 118699) and was packaged into lentivirus with the same method. Upon 48 hours of incubation, lentiviral supernatant was collected, filtered through 0.45-µm polyvinylidene difluoride filter (Millipore), and mixed with 8µg/ml polybrene. MCF-7 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 CRISPRa sgRNA library and 10μg/ml blasticidin for the MS2 component expression. The MOI for sgRNA library was 0.25 to make sure the transduced cells only carried one sgRNA (one lentivirus molecule) for each cell.

Proliferation and migration screen

For proliferation, at D0 0.1M/dish of cells (MCF-7) were plated into 10cm cell culture dishes and allowed to grow. Total cell count passaged could maintain a coverage >1,000X (defined as the number of cells divided by the number of unique library sequences). Cells were harvested at 7 and 14 days for gDNA extraction.

For migration screen, at D0 0.5M cells (MCF-7 or PATU-8988T) were divided and seeded in the upper part of 5 transwell inserts (0.1 M cells/transwell). The upper part of transwell inserts was filled with media lacking FBS, and the lower part with media containing 10% FBS. After 48h the PATU-8988T cells (1 week for MCF-7 cells) in the upper part of the chamber (impaired migration) and lower part (accelerated migration) were trypsinized and plated separately for growing for another 72h, after this time, cells were counted and collected for gDNA extraction. Control cells (D0) for both cell lines that did not undergo the migration assay were harvested at the same time as a reference population.

Genomic DNA extraction and sequencing library preparation

Genomic DNA was extracted with Zymo Quick-gDNA MidiPrep (Zymo Research D3100) as per the manufacturer's instructions. For PCR amplification for the sequencing purposes, the reactions and primers were the same as the sequencing for the library coverage. 6 Rev primers were used for sequencing (representing 6 barcodes).

Sequencing data analysis

For sgRNA library sequencing and for the screened cells, the fastq files from sequencing were trimmed with 'seqtk trimfq' (https://github.com/lh3/seqtk), and the sgRNAs was counted with count_spacer.py (https://github.com/fengzhanglab/Screening_Protocols_manuscript). All the sequencing libraries showed perfect matched reads >85%, undetected guides < 0.5% and skew ratio < 5.

For enrichment analysis, MAGeCK was applied for statistical analysis [183] (https://sourceforge.net/p/mageck/wiki/Home/).

3D genome interaction prediction

A 1 Mb sequence was extracted from the reference genome centered at each somatic SV breakpoint and was used as input for Orca [130] to predict the 3D genome organization with the same dataset from the previous chapter in this dissertation. SV breakpoints were provided to Orca to predict 3D genome structures through its web interface (https://orca.zhoulab.io/).

Results

Cell proliferation screens confirmed putative oncogenes detected by HYENA

The cell proliferation screen was applied to a breast cancer cell line MCF-7 with the previous version of HYENA detected putative oncogenes (Methods). MCF-7 was derived from the pleural effusion of a 69-year-old Caucasian metastatic breast cancer (adenocarcinoma), expressing the WNT7B oncogene and carrying PIK3CA gain-of-function mutation.

We collected the cells from D0, D7 and D14 after the antibiotic selection, and performed NGS to sequence the sgRNAs in each group of cells (Methods). Note that the lncRNA *TOB1-AS1* that we reported to accelerate cancer cell invasion in pancreatic cancer, was enriched in D7

cells but not D14 cells (**Fig. 8**), suggesting the pro-cell growth ability of *TOB1-AS1* was not strong enough to promote cell growth after the cells reached a specific confluence, but it could significantly drive cancer cell proliferation when the cells were seeded sparsely. The result also suggested about the different roles of the same cancer driving gene might have in different tumor contexts. We noticed that there was an oncogene, *RCCD1*, enriched in the D14 cells (**Fig. 8**), along with two other genes enriched – *AC021876.4* and *RPL31P59*. Both are annotated as pseudogenes.

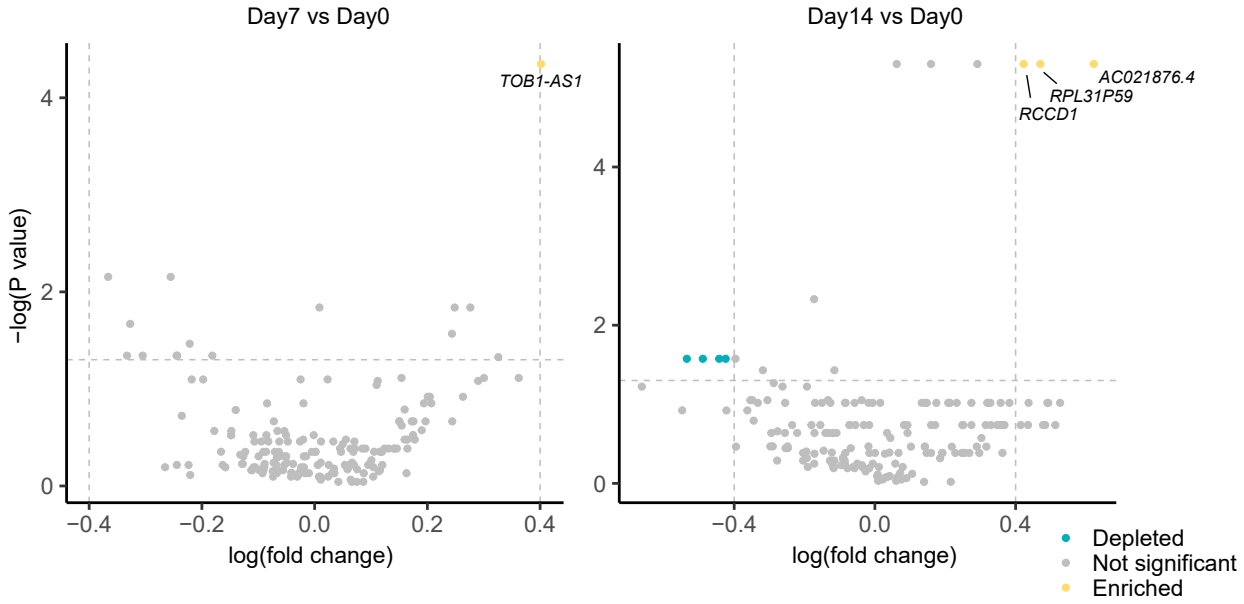

Figure 8. Volcano plots of the enriched or depleted genes in MCF-7 proliferation screen. Yellow and blue dots represent significantly (P value < 0.05) enriched and depleted genes with log(fold-change) larger than 0.4 and smaller than -0.4, respectively. Grey dots represent all other genes. Grey dash lines represent -log(P value) of -log(0.05) (horizontal), log(fold change) of 0.4 (vertical, right) and -0.4 (vertical, left). The significantly enriched genes were also labeled with gene symbols.

RCCD1 was predicted to have new enhancer-promoter interactions caused by SVs

RCCD1 (Regulator of chromosome condensation domain-containing protein 1) is recognized as a partner of the histone H3K36 demethylase KDM8 in chromosome segregation [184], has been identified as a potential driver of breast cancer in a recent transcriptome-wide

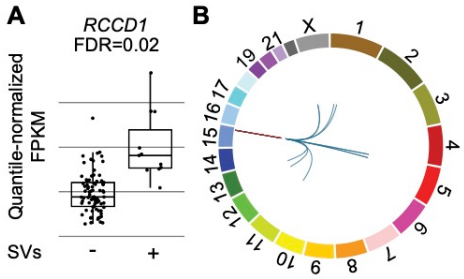

Figure 9. *RCCD1* gene expression and SVs near *RCCD1*.
A, Normalized expression of *RCCD1* in samples with (n=11) and without (n=66) nearby SVs in breast cancers. The boxplot shows median values (thick black lines), upper and lower quartiles (boxes), and 1.5× interquartile range (whiskers). Individual tumors are shown as black dots. **B**, Circos plot summarizing intrachromosomal SVs (blue, n=8) and translocations (red, n=3) near *RCCD1*.

association study [185]. A recent study reveals that RCCD1 is present in the mitochondrial matrix, where it interacts with the mitochondrial contact site/cristae organizing system and mitochondrial DNA (mtDNA), playing a crucial role in regulating mtDNA transcription, oxidative phosphorylation, and reactive oxygen species production [186]. Reported by Peng *et al*., RCCD1 is upregulated under hypoxic conditions, leading to reduced reactive oxygen species generation and decreased apoptosis, which supports cancer cell survival. It was demonstrated that RCCD1 promotes breast cancer cell proliferation in vitro and accelerates breast tumor growth in vivo. RCCD1 is overexpressed in breast carcinomas, and its expression levels are associated with more aggressive breast cancer phenotypes and poorer patient survival [186]. In addition, it has been shown that RCCD1 is overexpressed and associated with accelerated cancer cell proliferation and metastasis in lung adenocarcinoma and non-small cell lung cancer [187, 188], and initial evidence suggests that the oncogenic effect of RCCD1 stems from its regulatory role in cytoskeletal microtubule stability and TGF-β-induced epithelial-mesenchymal transition [174]. Given a number of the studies recognizing *RCCD1* as an oncogene in breast and lung cancers, it remains to be elucidated how this gene is activated in breast cancer.

In the HYENA results, 11 (12.5%) out of 77 tumors had some form of somatic SVs near *RCCD1* including translocations, deletions, inversions, and tandem duplications (**Fig. 9B**, **Supplementary Table S9** of the previous chapter). In one tumor with SV near *RCCD1*, based on the 3D genome interaction prediction, a translocation between chromosome 1 and 15 rearranged the regulatory sequences on *ADGRL2* gene body to *RCCD1*, and induced new chromatin interactions potentially activating *RCCD1* (**Fig. 10**). The results suggested *RCCD1* may be upregulated by enhancer hijacking in breast cancer.

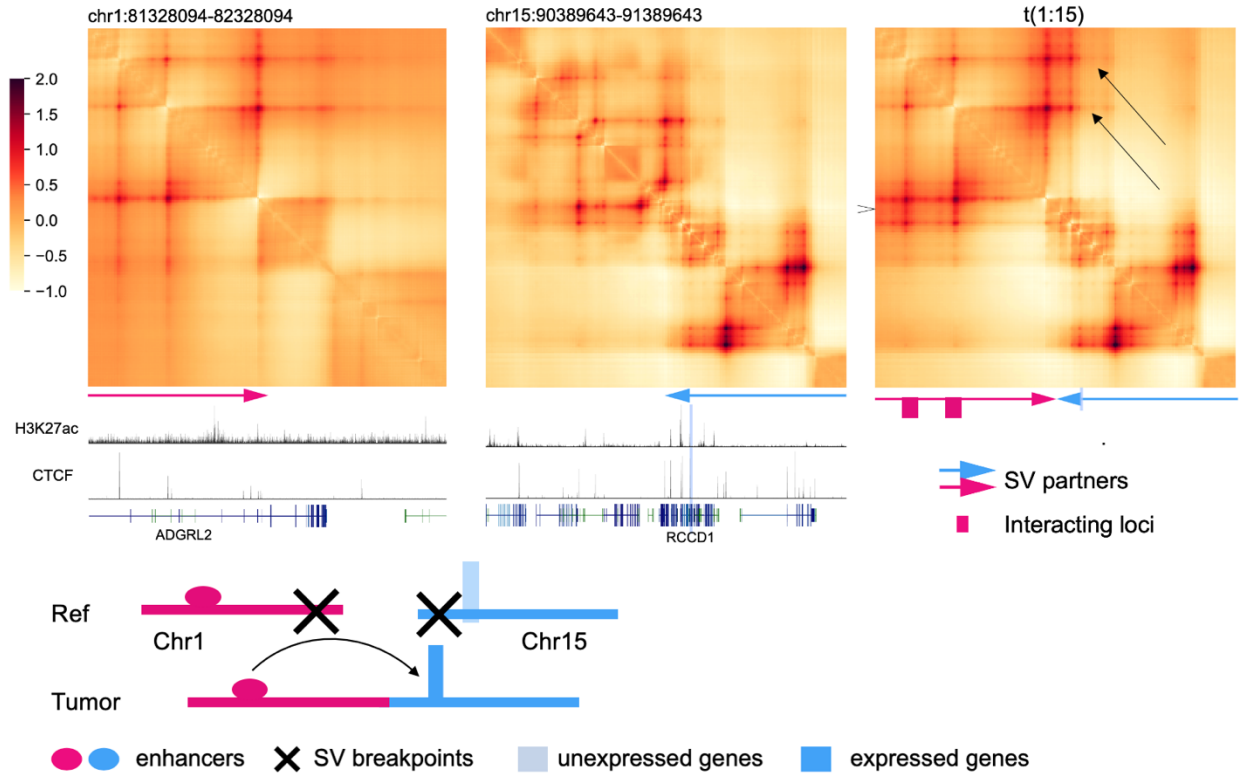

Figure 10. 3D genome structures predicted by deep-learning based algorithm Orca. Upper part shows the predicted 3D chromatin interaction maps of the chromosome 1 SV partner (left panel), chromosome 15 SV partner (middle panel), and the translocated region in the translocation between chromosome 1 and 15 (t1:15) (right panel). H3K27Ac and CTCF ChIP-Seq data of MCF-7 cell line are shown below the interaction maps. Lower part shows the diagram representing the proposed model of how this translocation activated *RCCD1*.

Cell migration screen in two cancer cell lines revealed potential oncogenes

The cell migration screens were performed using two cell lines, PATU-8988T, a pancreatic cancer cell line that originally had weak migration potence [155], and MCF-7, a breast cancer cell line with limited migration ability [189]. There was one gene enriched in the migrated cells in each of the two screens (**Fig. 11**).

In the MCF-7 result, the only significant gene that promoted cell migration was *POLR2F* (**Fig. 11**). It encodes the sixth largest subunit of RNA Polymerase II complex. Studies showed that *POLR2F*, together with two other genes, was significantly overexpressed in colorectal carcinoma tissues compared to normal tissues, and specifically its overexpression correlated with early disease occurrence and relapse [190]. In addition, *POLR2F* has been reported to be upregulated in other cancer types including gastric cancer [191], triple negative breast cancer [192], prostate cancer [193] and glioblastoma [194]. Those studies confirmed that *POLR2F* is playing a role in cancer development, relapse and drug resistance, associated with patient survival.

In our migration screen, the breast cancer cells with activated *POLR2F* showed more migration ability, suggesting that this gene might be able to promote cancer cell migration through unknown mechanisms related to transcription that is crucial to sustain their growth and survival. Combining with that it was predicted by HYENA to be an enhancer hijacking gene, its activation might be associated with transcription activation by distal enhancers.

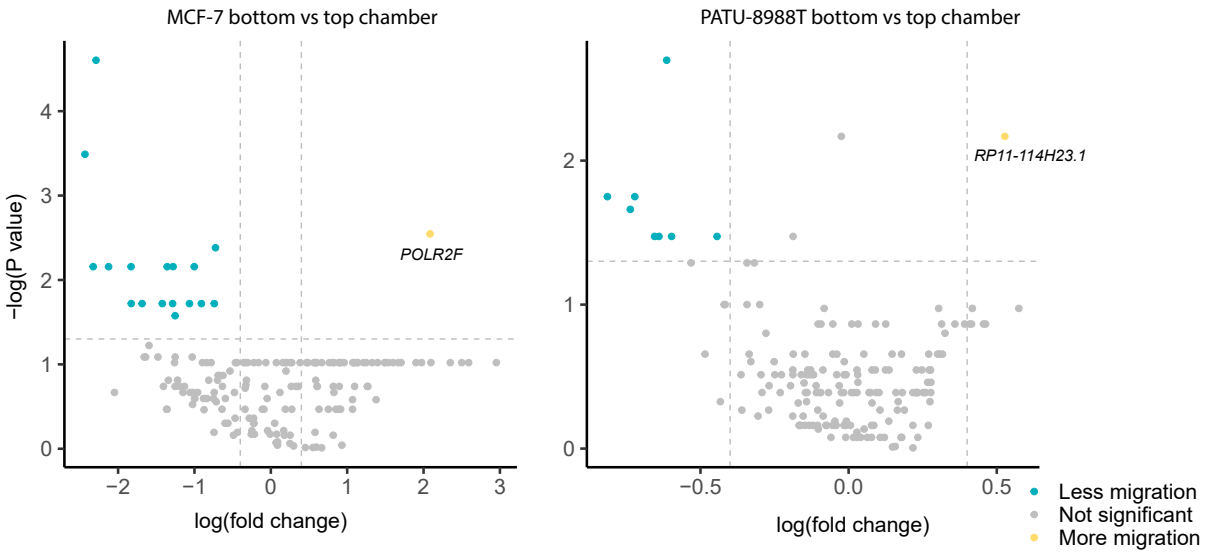

Figure 11. Volcano plots of the genes in bottom chamber compared to top chamber in migration screens.
Yellow and blue dots represent significantly (P value < 0.05) enriched and depleted genes in bottom chamber compared to top chamber with log(fold-change) larger than 0.4 and smaller than -0.4, respectively. Grey dots represent all other genes. Grey dash lines represent -log(P value) of -log(0.05) (horizontal), log(fold change) of 0.4 (vertical, right) and -0.4 (vertical, left). The significantly enriched genes were also labeled with gene symbols.

Discussion

In this study, we built a customized library for HYENA candidate genes and performed CRISPRa screens to identify the genes that can promote cancer cell proliferation or migration after transcription activation, to mimic the context of enhancer hijacking, where oncogenes were activated by distal enhancers. We found *RCCD1*, a gene reported to accelerate breast tumor growth [186], can promote MCF-7 proliferation after activation (**Fig. 8**); *POLR2F*, a subunit of RNA Pol II complex known to be overexpressed in multiple cancer types and involved in mechanisms of cisplatin resistance in gastric cancer [190-192, 194], can promote cell migration (**Fig. 8**). The results confirmed the capability of HYENA to predict oncogenes and suggested that

gene activation by CRISPRa could be a good approach to study the consequences and effects of oncogene activation.

However, there are still drawbacks in this study. First of all, in the library design, we included three well-studied oncogenes *CCND1*, *ERBB2*, *PIK3CA* as positive control. However, only one sgRNA targeting *ERBB2* was enriched in the migration screen, while other genes or sgRNAs were not enriched. This indicated that the readout measurement (proliferation and migration) or the cell line model selection (MCF-7) were likely not suitable for the aim of detecting genes' cancer-promoting abilities. Since MCF-7 is a cancer cell line that originally grows fast (about 30 hours) due to multiple mutations and oncogene activation [189, 195], the further increase of proliferation rate induced by potential oncogene activation is hard to distinguish or very marginal by traditional 2D cell culture approach. In addition, the activation of the three oncogenes in the library is possibly unable to enrich the cells carrying them, because *PIK3CA* is already upregulated in MCF-7 cells. To address this flaw in the study design, to grow a normal immortalized cell line or a cancer cell line that has a longer doubling time in 3D culture would make more sense. For example, MCF10A will be a better model because it is an epithelial cell line that undergoes growth arrest in Matrigel and forms acini [196, 197]. MCF10A is an extensively used model to investigate cell transformation and is known to be transformed by the expression of *ERBB2* and *PIK3CA* [198, 199]. This model fits better for our proposed aim to detect potential oncogenes. Besides, an in vivo screen may also be helpful to investigate the effects of gene activation in tumor growth or metastasis giving the context that models tumor development in the body. Note that genes function differently in different cancer types, so data interpretation should be done with caution. Non-enriched genes in one model could be cancer drivers in another model.

Second, the sgRNA library need to be improved. As HYENA was updated to a new version, the putative oncogene list was also significantly updated. One future direction is to design a library that contains a new list of candidate genes and positive control oncogenes that are not expressed in the model cell line, to further improve the readout. To use CRISPR/dCas9 to target non-coding genes is more frequently applied in recent screening studies, but many were targeting lncRNAs and the libraries went through stringent filtering for better targeting effects [164, 180, 181]. In our HYENA results, there were small RNAs and pseudogenes which are hard to target. Therefore, our library might not serve the goal of targeting those genes efficiently as desired. To address this issue, in the future a Perturb-Seq can be applied, to further identify individual gene targets, gene signatures, and cell states affected by individual sgRNAs and their genetic interactions [200].

Last but not least, to thoroughly investigate gene functions and oncogenic mechanisms, experiments that test individual gene's functions using KD and overexpression are required. Although here we performed screens to identify potential oncogenes detected by HYENA, we did not perform individual functional validations due to the limitations in our expertise and resources. A future direction should be exploring the enriched genes in our screen data one by one to demonstrate their pro-cancer abilities and underlying mechanisms. It is important to choose the models that align with the study aims and understand the tissue specific context of gene functions.

Oncogenes Activated by Distal Enhancers in Neuroblastoma

Introduction

Neuroblastoma is among the most common childhood solid tumors and displays great clinical and genetic heterogeneity [201]. Neuroblastomas can be classified into distinct groups of risk levels based on age as well as radiographic, histologic and cytogenetic factors [99]. The advances of next-generation sequencing (NGS) and inter-institutional collaboration have deepened our understanding of neuroblastoma biology and risk classification [98]. In addition to well-defined criteria (imaging stage, age, histology, differentiation, amplification of *MYCN*, diploidy status and 11q aberration), recent studies have demonstrated the association between genomic status and clinical outcome [202]. Although intense multi-modal treatment has been incorporated into clinical practice, survival rate of high-risk patients is still as poor as 50% [101], suggesting that our knowledge of pathogenetic mechanisms and potential risk factors are still far from enough.

The activation mechanisms of oncogenes are important for understanding the tumorigenic mechanisms and designing drugs targeting the actionable cancer drivers [120]. In turn, inferring putative oncogenes based on activation mechanisms becomes an efficient approach to identify novel oncogenes [203]. Well-defined mechanisms of oncogene activation include point mutations happening in coding regions causing gain of function, amplifications and gene fusions that express fused driver proteins [204, 205]. However, both experimental and bioinformatic studies based on these patterns omit the mutations located in noncoding regions as well as the regulatory functions of noncoding sequences that widely distribute in human genome [44]. Epigenomic and genomic studies in neuroblastomas have revealed that the rearrangements of enhancers could explain aberrantly expressed oncogenes like *MYC*, *MYCN* and *TERT* [91, 110,

111], suggesting the unignorable oncogenic roles of such events called enhancer hijacking. Mis-regulation of *cis*-regulatory sequence (CRE) activities or enhancer-promoter interactions have been shown to activate oncogenes in multiple tumor types, and distinct CREs in some cases activate same oncogenes, rendering tumor cells selective advantages and leading to oncogenesis [114].

Genomic instability is a hallmark of cancer, and structural variants (SVs) that widely spread in cancer genomes can heavily affect enhancer-promoter interactions by different mechanisms, including disruption or repositioning of CREs near genes [95], formation of cryptic promoters and disruption of topologically associating domain (TAD) organization affecting long-range enhancer-promoter interactions [94, 115, 148]. As pediatric cancer genomes have less point mutations and small indels [206], the impact of SVs can be stronger than what has been observed in adult cancers. Previous studies mainly focus on copy number variants (CNVs) and how they activate well-known oncogenes like *MYC*N [111, 207]. However, a systematic exploration of oncogenes activated by enhancer hijacking needs to be done in different groups of neuroblastomas and pediatric brain tumors, to discover previously unknown oncogenes as well as better understand the cancer driving functions of genomic rearrangements.

Genome-wide large-scale projects in pediatric cancers such as the Gabriella Miller Kids First Pediatric Research Program (GMKF) have provided unprecedented resources for us to integrate expression profiles, mutation effects and pathway enrichment to study cancer genomes. Many publications have drawn mutational landscapes based on properties and consequences of somatic mutation [122, 126, 204]. In addition, unbiased computational tools like HYENA which can utilize whole genome sequencing (WGS), RNA sequencing (RNA-Seq), CNV profiles, and clinical information to identify the association between putative oncogenes activation and SV

breakpoints nearby, are helpful to predict novel driver genes activated by enhancer hijacking events [93, 96, 162]. This type of analysis will provide genetic biomarkers and promising targets to guide patient classification and precise low-toxicity treatments, thus having great significance in neuroblastoma which are known to have considerable heterogeneity and lack in efficient therapies for high-risk patient groups.

In this project, we analyzed 189 neuroblastoma samples that have RNA-Seq, CNV, and normal-matched WGS data. We detected somatic SVs with two algorithms and applied HYENA pipeline to explore putative oncogenes activated by genomic rearrangements. We identified five putative oncogenes, including *TERT*, that had no more than 10 copies and were upregulated when carrying SV breakpoints nearby. When loosening the parameters, we detected 58 oncogene candidates in all samples and 26 candidates in high-risk samples. Our study provides insights into the novel oncogenes activated by enhancer hijacking in neuroblastoma and the putative oncogenes could potentially serve as therapeutic targets in the future.

Methods

Datasets

This study used data generated by the Gabriella Miller Kids First Pediatric Research Program (GMKF). We limited our study to 189 neuroblastoma cases from which both WGS data and RNA-Seq data were available for tumor samples and WGS data was available for normal samples. The cohort was composed of 97 low-risk, 44 intermediate-risk, and 48 high-risk neuroblastoma cases. More detailed information on the sample distribution can be found in **Figure 12**.

WGS bam files (both normal and tumor), gene expression fragments per kilobase of million mapped (FPKM) data from RNA-Seq, CNV data, and somatic SVs called by Manta [208] were downloaded from the data portal of GMKF (https://kidsfirstdrc.org/help-center/cavatica-cloud-platform/). All were mapped with reference genome hg38. Clinical information including gender, risk level, age at diagnosis was from INRG (https://commons.cri.uchicago.edu/pcdc/) with the help of Dr. Mark A. Applebaum and with the consent from Dr. Susan L. Cohn.

SV calling and filtering

Manta called somatic SVs were downloaded directly from GMKF neuroblastoma dataset, and a detailed description was listed here (https://github.com/kids-first/kf-somatic-workflow). We filtered the SVs called by Manta that were supported by only spanning read pairs (PR), less than 3 PR (PR < 3), or less than 3 split-reads (SR < 3) because most of these SVs were not supported by CNV breakpoints (**Fig. 13**). After checking the mapped reads in the tumor WGS bam files, we found most of those PR < 3 or SR < 3 SVs were not observed and likely to be false positive. If the breakpoints of a Manta SV and a Meerkat SV fell within 50bp, they were considered the same SV. To avoid algorithm-specific biases induced by individual SV callers, we also called somatic SVs using Meerkat [46] according to the user manual. The final somatic SVs were the union of Meerkat SVs and filtered Manta SVs.

Predicting enhancer hijacking genes with HYENA

The analytic pipeline of HYENA has been extensively described in the first chapter of this dissertation. The input files for HYENA to detect putative oncogenes included the hg38 gene annotation file included in the HYENA package (https://github.com/yanglab-computationalgenomics/HYENA), SV bedpe files, formatted gene expression, CNV files

mapped to genes, and clinical information (gender and age at diagnosis), according to HYENA manual (https://github.com/yanglab-computationalgenomics/HYENA/blob/main/User_manual_0_5_4.pdf).

In summary, 0 to 5 principal components (PCs) were tested and for each PC level, 100 permutation tests were run to generate the empirical P-value. The model included gene copy number, sex and age. Finally, PC0 results were determined to be final results (**Table 1**). Another model without gene copy number was run for all samples and high-risk samples. PC0 results were taken as final results (**Table 2 and 3**).

Results

SV calling for the 189 neuroblastoma samples

In this chapter, we analyzed 189 neuroblastoma samples downloaded from GMKF including 97 low-risk, 44 intermediate-risk, and 48 high-risk cases (**Fig. 12**).

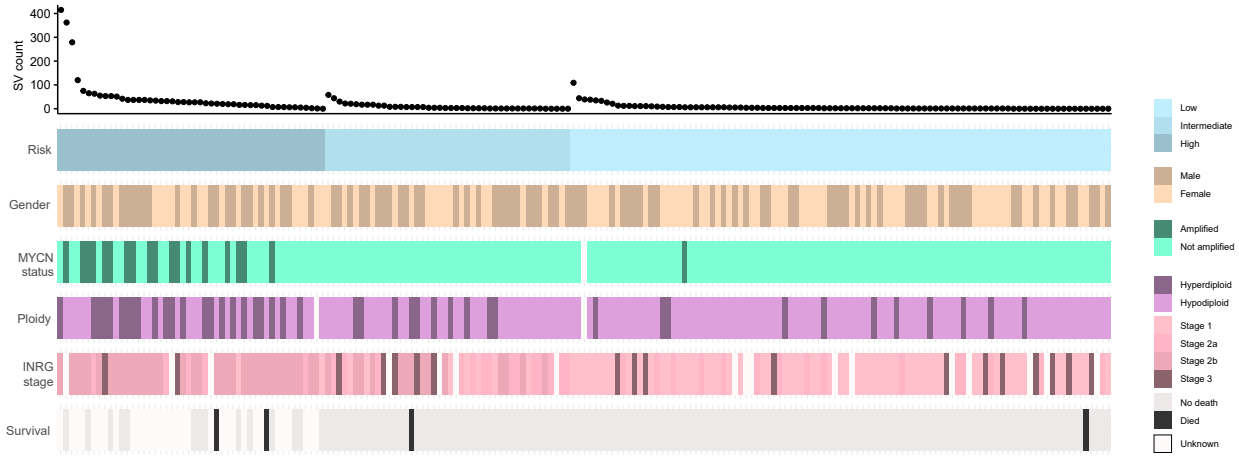

Figure 12. Landscape of 189 neuroblastoma cases.
Data tracks showed the SV count, risk level, gender, MYCN amplification status, ploidy status, INRG stage, and survival status for each individual case included in this study. All the information was from the clinical information of the samples except for the SV counts. SV calling process could be found in Methods.

To call somatic SVs as the input for HYENA analysis, we checked how the Manta SVs from GMKF were supported by CNV breakpoints and how the reads were mapped to reference genome in Integrative Genomics Viewer (IGV). We found most SVs with less than three split-reads (SR) or spanning read pairs (PR) were likely to be artefacts (**Fig. 13**, Methods), so they were filtered out and then taken the union with Meerkat SVs (Methods) to generate the SV

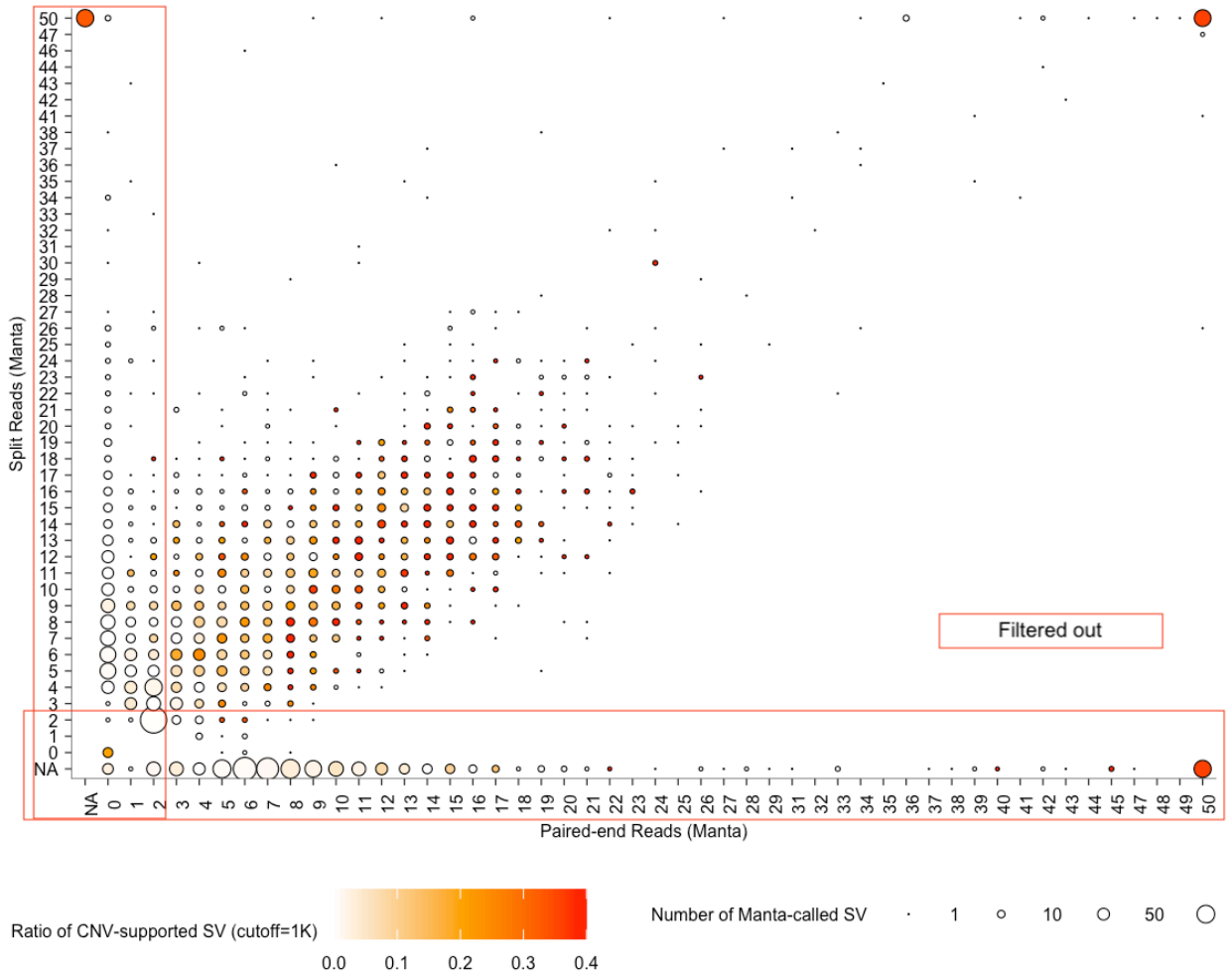

Figure 13. The ratio of CNV-supported Manta SVs and the SV counts supported by SR and PR count combinations.
The color scale of each dot shows the ratio of the SVs (at specific SR and PR combinations shown in x and y axis) were supported by CNVs. The size of each dot shows the number of Manta called SVs at the represented SR and PR counts. All SVs with SR or PR larger than 50 were included in the counts at 50. Red rectangles represent the Manta SVs that were filtered out.

counts shown in **Figure 12**. High-risk samples had more somatic SVs per sample compared to intermediate- and low-risk samples, with a maximum of 415 somatic SVs per sample.

Enhancer hijacking candidates detected in GMKF neuroblastomas

After SV calling and formatting the input data, we first applied HYENA pipeline to the 189 neuroblastoma samples. With the default parameters of PC0, sex, age, recurrency larger than 5% as well as copy numbers no larger than 10, HYENA output included five enhancer hijacking candidates, *GZF1*, *NBAS*, *TTC32*, *TRIP13* and *TERT* (**Table 1**). *TERT* has been reported to carry frequent SVs nearby and be activated by rearranged enhancers or super enhancers [111]. The detection of *TERT* suggested that HYENA was able to find novel oncogene candidates activated by SVs.

Table 1. HYENA default setting predicted oncogenes with 189 neuroblastoma samples

Gene ID	Chrom	Start	End	Gene Type	Gene Name	Ratio	Freq	P Emp FDR
ENSG00000125812	20	23362182	23373062	protein_coding	GZF1	3/51	5.6	0.052
ENSG00000151779	2	15166914	15561334	protein_coding	NBAS	9/77	10.5	0.052
ENSG00000183891	2	19896631	19901983	protein_coding	TTC32	4/68	5.6	0.052
ENSG00000071539	5	892884	919357	protein_coding	TRIP13	4/72	5.3	0.077
ENSG00000164362	5	1253147	1295068	protein_coding	TERT	5/72	6.5	0.077

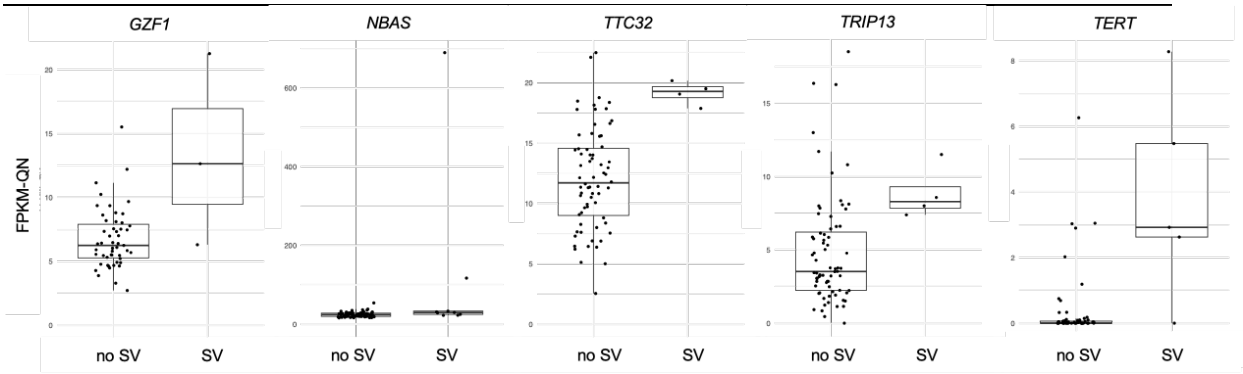

Figure 14. Gene expression levels of the five candidate genes detected in neuroblastoma. Gene names are listed on top of the plots. Y-axis represents gene expression level in FPKM quantile normalized values (FPKM-QN). X-axis shows the SV breakpoint status of the up- and down-stream 500kb of the gene TSSs.

High-risk neuroblastomas frequently carry CNVs of chromosomal arms, including gain of chromosome 1q, gain of chromosome 2p, gain of chromosome 17q (17q+), loss of chromosome 1p (1p-), loss of chromosome 3p, and loss of chromosome 11q (11q-). 17q+ happens in almost all high-risk neuroblastoma patients [98]. These CNVs that cause gains of chromosome arms can induce higher gene expression levels and more breakpoints. Because HYENA detects putative enhancer hijacking genes based on the association between SV breakpoint and gene expression level, there might be passenger events in the candidate gene list. Therefore, we examined the co-occurrence of known recurrent CNVs and the SV breakpoints near the candidate genes (**Fig. 15**). We would like to exclude potential passenger events and identify driver events that are not associated with known driving CNVs.

GZF1 encodes ZNF336, which may regulate the spatial and temporal expression of the *HOXA10* gene, which plays a role in morphogenesis [209]. It was found to be frequently deleted in esophageal cancer, but the underlying mechanisms remain unclear [210]. *GZF1* is on chromosome 20, and only one sample had the co-occurrence of *GZF1* SV and other known events, suggesting it might be an independent gene. However, only two out of five samples that had *GZF1* SVs were high-risk, suggesting this gene is not driving high-risk neuroblastomas (**Fig. 15**). TRIP13 (Thyroid Hormone Receptor Interacting Protein 13) plays a key role in regulating mitotic processes, including spindle assembly checkpoint and DNA repair pathways, which may account for chromosome instability. It is overexpressed and associated with poorer survival in multiple cancer types including lung, breast, prostate, head and neck as well as colorectal cancers, considered to be a potential target for treatment [211]. *TRIP13* SVs are co-occurring with TERT SVs (**Fig. 15**), and the gene is approximately 350kb upstream of *TERT*, so it is likely to be a passenger associated with *TERT*. *TTC32* was detected as an essential gene in a previous

CRISPR knockout screen using neuroblastoma cell line KP-N-YS [212], but its functions were not investigated either. Four out of five *TTC32* SVs co-occur with *MYCN* amplification (**Fig. 15**), indicating this gene is likely to be a passenger gene associated with *MYCN*.

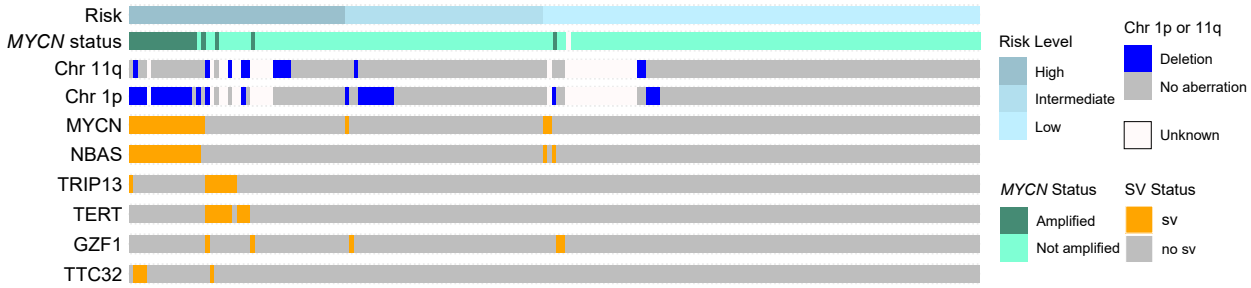

Figure 15. The co-occurrence of recurrent CNVs, *MYCN* status, and SV status of candidate genes detected with HYENA default parameters.
The plot shows the risk level, *MYCN* status, deletion of chromosome 1p or 11q, and SV status of *MYCN*, *NBAS*, *TRIP13*, *TERT*, *GZF1* and *TTC32* in all samples. Each row is a gene except for risk, *MYCN* status, CNVs on chromosome 11q or 1p. Each line is one sample.

As *MYCN* amplification is a marker for high-risk neuroblastoma, we examined the *MYCN* copy numbers and whether there was any SV breakpoint located in 500kb or 3mb up- or down-stream of its transcription starting site (TSS). Together with *MYCN* expression level shown by RNA-Seq FPKM, we saw that just as reported in other studies, high-risk neuroblastoma samples carry high copy numbers of *MYCN* [98], and the gene expression level was associated with copy number (**Fig. 16**) as expected. Notably, there were some samples with *MYCN* amplification (> 4 copies) but low expression level. Considering that samples carrying genes with more than 10 copies were excluded in the regression model (Methods of the HYENA Chapter) and gene copy number is positively correlated with gene expression, this could explain why HYENA did not detect *MYCN* as an enhancer hijacking gene (**Table 1**) although in previous studies it has been shown that *MYCN* can be activated by rearranged enhancers [213].

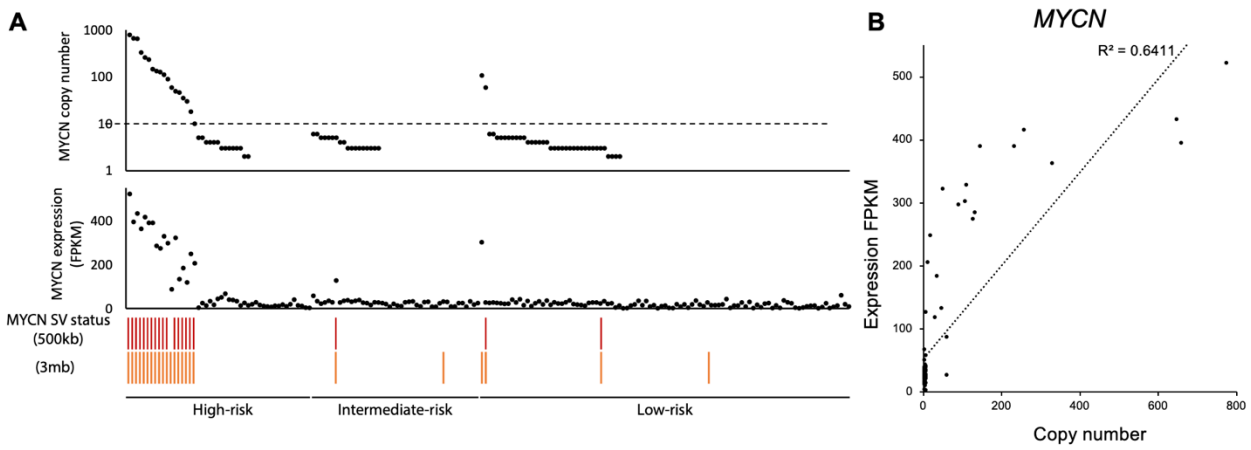

Figure 16. *MYCN* expression is positively correlated with copy number, and SVs near *MYCN* reflected CNVs.
A. Samples were sorted by risk levels and *MYCN* copy numbers. Tracks representing *MYCN* copy number, expression FPKM levels and somatic SV breakpoint status in individual samples. Copy numbers and gene expression data were downloaded from GMKF, with blank representing NAs in the copy number track. Red bars represent there is SV breakpoint mapped to the 500kb window upstream or downstream of *MYCN* TSS, while no red bar means there is not a breakpoint within this window. Orange bars represent there is SV breakpoint mapped to the 3mb window upstream or downstream of *MYCN* TSS. **B**. Scatter plot showing the correlation between *MYCN* gene expression FPKM and copy number. Each dot is one sample.

While examining the results, we found there were gaps of DNA segments without copy number data in the CNV files provided by GMKF datasets. When feeding into the analysis pipeline, these missing values of copy numbers would cause the sample to be excluded in the HYENA model. It might hinder the detection of enhancer hijacking genes by reducing sample size in the model. In addition, when we included CNV info into the analysis, HYENA would consider gene copy as a variant and exclude samples with larger than 10 copies for the gene under test. Therefore, we performed the analysis again with 3% recurrent rate and without putting CNV information into the model (**Methods**). There were 58 putative oncogenes detected, listed in **Table 2**. *MYCN* and two long non-coding RNAs (lncRNAs) next to *MYCN* showed up as top candidates, indicating that it was not detected in the analysis including CNV filter because its high copy numbers in high expression samples were excluded in the model. In addition to the

known genes *MYCN* and *TERT*, *CCND1* is also an enhancer hijacking gene reported in B cell malignancies [214]. Because the model did not include copy number in the regression model, or filter out the samples with gene copy larger than 10 (Methods), many genes in Table 2 might be upregulated in neuroblastoma due to copy gain instead of enhancer hijacking. Note that a substantial proportion of the candidate genes might be passengers of known driving CNVs instead of driver genes.

Table 2. Enhancer hijacking candidates in neuroblastoma with lower frequency requirement including high-copy genes

Gene ID	Chrom	Start	End	Gene Type	Gene Name	Ratio	Freq	P Emp FDR
ENSG00000134323	2	15940550	15947007	protein_coding	MYCN	20/169	10.6	0.000
ENSG00000233718	2	15918350	15942249	lncRNA	MYCNOS	20/169	10.6	0.000
ENSG00000079785	2	15591178	15634346	protein_coding	DDX1	19/170	10.1	0.000
ENSG00000151779	2	15166914	15561334	protein_coding	NBAS	18/171	9.5	0.000
ENSG00000223850	2	15920399	15936017	lncRNA	MYCNUT	20/169	10.6	0.000
ENSG00000226041	2	16202430	16204226	lncRNA	AC010745.1	20/169	10.6	0.000
ENSG00000228876	2	16224047	16333978	lncRNA	AC010745.2	20/169	10.6	0.000
ENSG00000236289	2	16013928	16087201	lncRNA	GACAT3	20/169	10.6	0.001
ENSG00000149716	11	69653076	69675416	protein_coding	LTO1	6/183	3.2	0.001
ENSG00000162344	11	69698238	69704022	protein_coding	FGF19	6/183	3.2	0.001
ENSG00000231031	2	15690782	15744339	lncRNA	LINC01804	19/170	10.1	0.001
ENSG00000169016	2	11444375	11466177	protein_coding	E2F6	6/183	3.2	0.001
ENSG00000118961	2	20684014	20823130	protein_coding	LDAH	6/183	3.2	0.001
ENSG00000164363	5	1225381	1246189	protein_coding	SLC6A18	8/181	4.2	0.001
ENSG00000174358	5	1201595	1225111	protein_coding	SLC6A19	9/180	4.8	0.001
ENSG00000108883	17	44849948	44899445	protein_coding	EFTUD2	7/182	3.7	0.001
ENSG00000196208	2	11482341	11642788	protein_coding	GREB1	7/182	3.7	0.001
ENSG00000236989	2	16085222	16105841	lncRNA	AC142119.1	20/169	10.6	0.001
ENSG00000237326	2	15801747	15810877	lncRNA	AC113608.1	20/169	10.6	0.003
ENSG00000234022	2	15564170	15573868	lncRNA	AC008278.2	18/171	9.5	0.004
ENSG00000161692	17	44708608	44752264	protein_coding	DBF4B	6/183	3.2	0.004
ENSG00000279663	2	16541690	16545695	TEC	RP11-542H15.1	20/169	10.6	0.005
ENSG00000142319	5	1392794	1445440	protein_coding	SLC6A3	12/177	6.3	0.006
ENSG00000214842	2	17510584	17518439	protein_coding	RAD51AP2	6/183	3.2	0.008
ENSG00000186185	17	44924709	44947773	protein_coding	KIF18B	7/182	3.7	0.011
ENSG00000071539	5	892884	919357	protein_coding	TRIP13	8/181	4.2	0.013

ENSG00000229224	2	29088649	29097586	lncRNA	AC105398.3	6/183	3.2	0.013
ENSG00000232444	2	16316324	16319566	lncRNA	AC010745.4	20/169	10.6	0.013
ENSG00000172992	17	45023340	45061109	protein_coding	DCAKD	7/182	3.7	0.013
ENSG00000073670	17	44758988	44781846	protein_coding	ADAM11	8/181	4.2	0.014
ENSG00000172927	11	69294151	69367726	protein_coding	MYEOV	8/181	4.2	0.014
ENSG00000182963	17	44798448	44830816	protein_coding	GJC1	6/183	3.2	0.014
ENSG00000108352	17	40177010	40195656	protein_coding	RAPGEFL1	6/183	3.2	0.015
ENSG00000164362	5	1253147	1295068	protein_coding	TERT	9/180	4.8	0.015
ENSG00000239899	2	11584773	11585047	misc_RNA	RN7SL674P	7/182	3.7	0.019
ENSG00000180336	17	44656404	44690308	protein_coding	MEIOC	7/182	3.7	0.034
ENSG00000185344	12	123712353	123761755	protein_coding	ATP6V0A2	6/183	3.2	0.034
ENSG00000214657	2	15869939	15870243	pseudogene	RPLP1P5	20/169	10.6	0.034
ENSG00000267334	17	44947912	44948939	lncRNA	KIF18B-DT	7/182	3.7	0.034
ENSG00000179270	2	29060976	29074523	protein_coding	PCARE	6/183	3.2	0.037
ENSG00000162341	11	69048932	69136316	protein_coding	TPCN2	6/183	3.2	0.037
ENSG00000240125	17	40439467	40439917	pseudogene	RPL23AP75	6/183	3.2	0.037
ENSG00000229087	2	15397435	15397782	pseudogene	RPS26P18	15/174	7.9	0.045
ENSG00000284713	11	69155478	69159752	protein_coding	SMIM38	7/182	3.7	0.057
ENSG00000243541	2	15950689	15950981	misc_RNA	RN7SL104P	20/169	10.6	0.063
ENSG00000247872	5	816346	817001	pseudogene	SPCS2P3	8/181	4.2	0.065
ENSG00000261070	11	69147228	69171564	lncRNA	RP11-554A11.8	7/182	3.7	0.070
ENSG00000132740	11	68903863	68940602	protein_coding	IGHMBP2	6/183	3.2	0.071
ENSG00000180329	17	44673069	44689779	protein_coding	CCDC43	6/183	3.2	0.071
ENSG00000233622	19	40808474	40812100	pseudogene	CYP2T1P	6/183	3.2	0.073
ENSG00000028310	5	850291	892801	protein_coding	BRD9	8/181	4.2	0.073
ENSG00000110092	11	69641156	69654474	protein_coding	CCND1	7/182	3.7	0.073
ENSG00000188818	5	795606	850986	protein_coding	ZDHHC11	8/181	4.2	0.073
ENSG00000256508	11	69012283	69018447	lncRNA	MRGPRF-AS1	6/183	3.2	0.073
ENSG00000146872	17	62458658	62615481	protein_coding	TLK2	6/183	3.2	0.076
ENSG00000033627	17	42458844	42522582	protein_coding	ATP6V0A1	6/183	3.2	0.093
ENSG00000111361	12	123620406	123633766	protein_coding	EIF2B1	6/183	3.2	0.094
ENSG00000008838	17	40019097	40061215	protein_coding	MED24	6/183	3.2	0.097

To investigate if these genes are close to known oncogenes in neuroblastoma, and to check if the SVs near candidate genes have any association with the SVs near known oncogenes, we plotted the SV status with known CNVs and *MYCN* status again to rule out the potential passenger genes. Indeed, most of the candidates, including both coding and none coding genes,

that are close to known oncogenes *MYCN* or *TERT* had SV co-occurrence, suggesting these genes were not likely to be cancer drivers, but rather upregulated when *MYCN* or *TERT* got over-expressed in these tumors (**Fig. 17**).

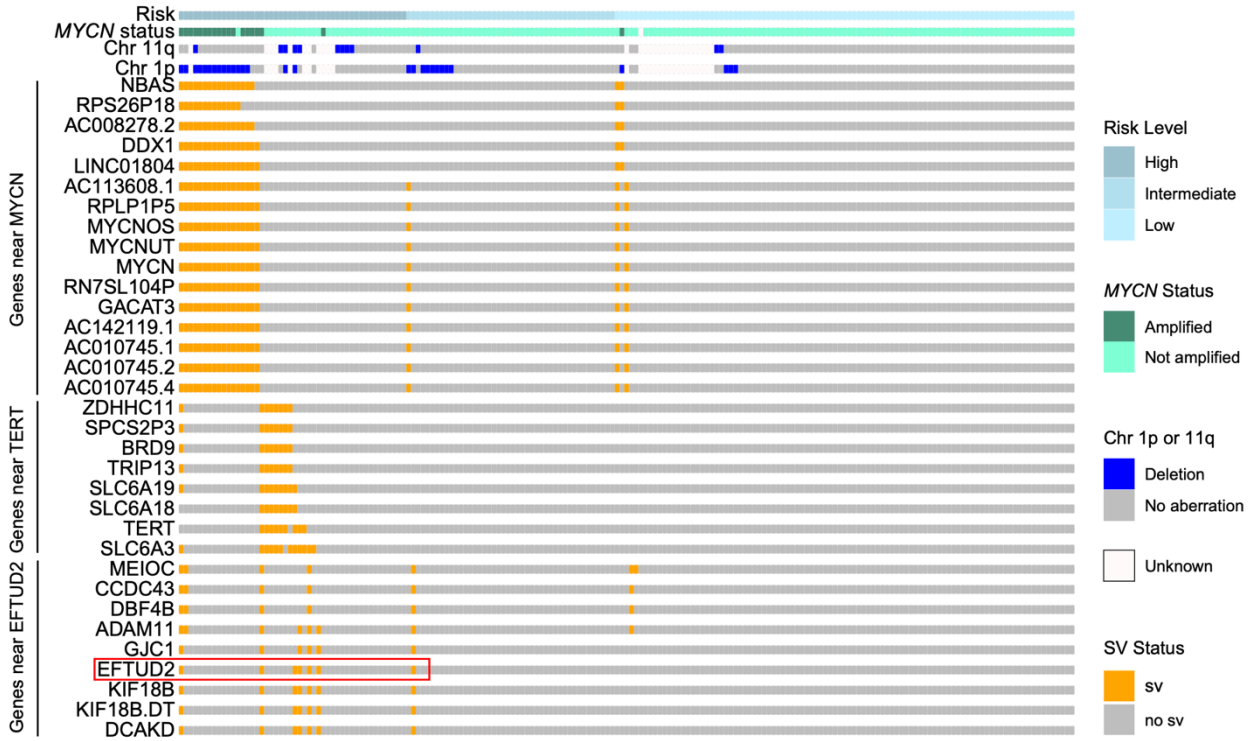

Figure 17. The co-occurrence of recurrent CNVs, *MYCN* status, and the SV status of selected candidate genes detected with HYENA without CNV information input and 3% recurrence rate cutoff.
The plot shows the risk level, *MYCN* status, deletion of chromosome 1p or 11q, and SV status of candidate genes in all samples. The genes are grouped by their locations relative to *MYCN*, *TERT* or *EFTUD2*. Red rectangle highlights the gene of interest, *EFTUD2*. Each row is a gene except for risk, *MYCN* status, CNVs on chromosome 11q or 1p. Each line is one sample.

Next, to explore the potential novel oncogenes in high-risk neuroblastomas, we performed HYENA analysis in the 48 high-risk samples. Since HYENA pipeline filters genes to be tested by a recurrent rate, and small sample size could limit its ability to detect important events, we used a cutoff of 3% frequency in this analysis (Methods) and did not include CNV information in this analysis as mentioned above. As listed in **Table 3**, there were 26 putative oncogenes detected by HYENA. We noticed that *ALK* showed up as an enhancer hijacking gene

here (**Table 3**), which is within expectation because *ALK* can be activated by point mutations or amplifications [215, 216]. When comparing the SV status of the candidate genes detected in high-risk neuroblastomas, we noticed that *MYCNOS*, *MYCNUT*, *DDXT*, *NBAS* and many other genes close to *MYCN* often had SVs which co-occurred with *MYCN* SVs (**Fig. 17**), indicating they are likely to be passengers amplified together with *MYCN* but not driver genes in high-risk neuroblastoma. *TERT* carried nearby SVs exclusively to *MYCN* SVs (**Fig. 17**). This is consistent with a previously published study that *TERT* can drive a subset of high-risk neuroblastomas [111]. *SLC6A18* and *SLC6A3* are both from the pharmacologically important family of transporter proteins, solute carriers family 6 (SLC6). SLC6 transporters, which include the serotonin, dopamine, norepinephrine, GABA, taurine, creatine, as well as amino acid transporters, are important to normal nervous system functions and associated with a number of human neurological disorders [217]. *SLC6A18*, *SLC6A3* and other genes close to *TERT* had co-occurred SVs, indicating they might be passengers (**Fig. 17**) for high-risk neuroblastomas. However, it was reported that SLC34A2, a member of SLC transporters, promoted neuroblastoma cell stemness via enhancement of Wnt/β-catenin signaling and thus were considered as an oncogene [218]. Therefore, the oncogenic functions of SLC6 family genes in neuroblastoma need further study, whether they are independent oncogenes, and how they are activated by enhancer hijacking remain to be investigated.

Table 3. Enhancer hijacking candidates in high-risk neuroblastoma including high-copy genes

Gene ID	Chrom	Start	End	Gene Type	Gene Name	Ratio	Freq	P Emp FDR
ENSG00000233718	2	15918350	15942249	lncRNA	MYCNOS	17/31	35.4	0.000
ENSG00000134323	2	15940550	15947007	protein_coding	MYCN	17/31	35.4	0.000
ENSG00000223850	2	15920399	15936017	lncRNA	MYCNUT	17/31	35.4	0.002
ENSG00000079785	2	15591178	15634346	protein_coding	DDX1	17/31	35.4	0.003

ENSG00000229224	2	29088649	29097586	lncRNA	AC105398.3	4/44	8.3	0.035
ENSG00000230737	2	29890371	29892354	lncRNA	AC106870.1	4/44	8.3	0.040
ENSG00000255774	11	69475567	69481545	lncRNA	LINC02747	5/43	10.4	0.040
ENSG00000164363	5	1225381	1246189	protein_coding	SLC6A18	8/40	16.7	0.046
ENSG00000118960	2	20560448	20651130	protein_coding	HS1BP3	4/44	8.3	0.065
ENSG00000151779	2	15166914	15561334	protein_coding	NBAS	16/32	33.3	0.065
ENSG00000171094	2	29192774	29921586	protein_coding	ALK	3/45	6.2	0.065
ENSG00000228876	2	16224047	16333978	lncRNA	AC010745.2	17/31	35.4	0.065
ENSG00000108883	17	44849948	44899445	protein_coding	EFTUD2	6/42	12.5	0.080
ENSG00000109118	17	28905250	28951771	protein_coding	PHF12	4/44	8.3	0.079
ENSG00000118369	11	78188812	78215232	protein_coding	USP35	3/45	6.2	0.080
ENSG00000118961	2	20684014	20823130	protein_coding	LDAH	4/44	8.3	0.080
ENSG00000142319	5	1392794	1445440	protein_coding	SLC6A3	12/36	25	0.079
ENSG00000158125	2	31334321	31414742	protein_coding	XDH	3/45	6.2	0.079
ENSG00000162344	11	69698238	69704022	protein_coding	FGF19	4/44	8.3	0.079
ENSG00000169016	2	11444375	11466177	protein_coding	E2F6	5/43	10.4	0.079
ENSG00000203643	2	11721619	11724222	lncRNA	AC012456.3	3/45	6.2	0.080
ENSG00000224177	2	11372612	11403175	lncRNA	LINC00570	5/43	10.4	0.079
ENSG00000251718	2	11561194	11561306	snRNA	RNU2-13P	5/43	10.4	0.080
ENSG00000278797	19	46533669	46534351	pseudogene	LLNLR-276E7.1	3/45	6.2	0.079
ENSG00000279757	16	72973374	72973832	TEC	CTD-2326C4.1	4/44	8.3	0.079
ENSG00000230203	22	26422071	26423193	pseudogene	CTB-1048E9.7	3/45	6.2	0.097

***EFTUD2* was a potential enhancer hijacking gene in high-risk neuroblastoma**

While looking for candidate genes as independent drivers, we noticed *EFTUD2* SVs were not associated with *MYCN* or *TERT* SVs (**Fig. 17**, **Table 2**). *EFTUD2* is also an enhancer hijacking candidate detected in high-risk samples (**Table 3**). Although it is located on chromosome 17q and 17q+ is very common in high-risk samples [98], we found that the largest copy number was 10 for this gene, and its expression level was not statistically correlated with copy number (**Fig. 18**). Importantly, all 6 samples with SVs near *EFTUD2* were high-risk. Therefore, we consider it as a potential oncogene that is activated independent of other known oncogenes to drive high-risk neuroblastoma.

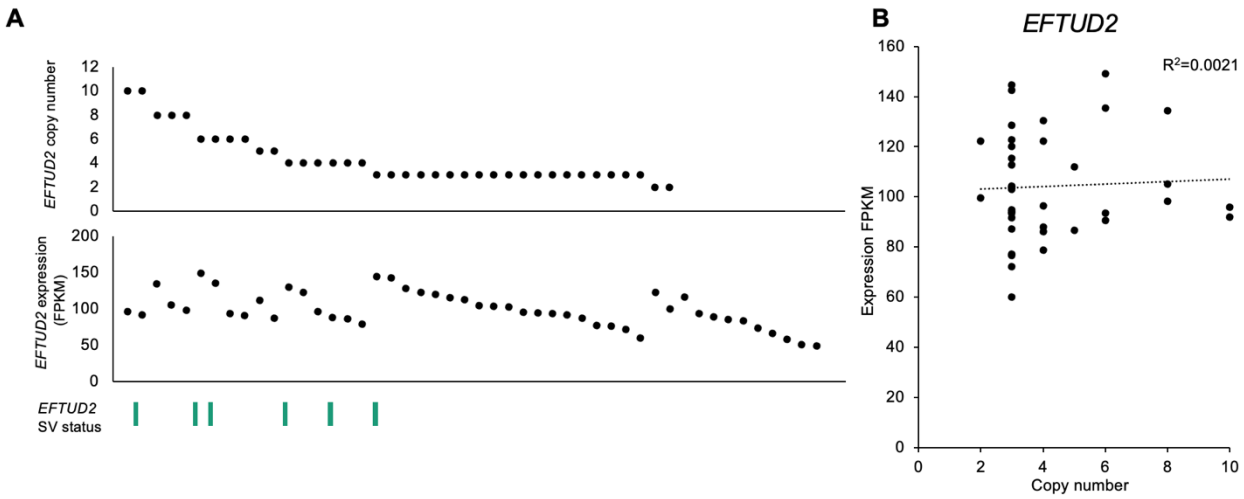

Figure 18. *EFTUD2* expression is not correlated with copy number in high-risk neuroblastoma.
A. Samples were sorted by *EFTUD2* copy number and gene expression. Tracks representing *EFTUD2* copy number, expression FPKM levels and somatic SV breakpoint status in individual samples. Copy numbers and gene expression data were downloaded from GMKF, with blank representing NAs in the copy number track. Green bars represent there is SV breakpoint mapped to the 500kb window upstream or downstream of *EFTUD2* TSS, while no bar means there is not a breakpoint within this window. **B**. Scatter plot showing the correlation between *EFTUD2* gene expression FPKM and copy number. Each dot is one sample. R-square shows there is no correlation between the two values.

EFTUD2 (Elongation factor Tu GTP binding domain containing 2) plays a pivotal role in splicing precursor mRNAs (pre-mRNAs) into mature mRNAs [219]. It is an oncogene associated with tumor progression and poor survival in multiple cancer types including liver, breast, and endometrial cancers [220-222]. Although there was no report about *EFTUD2* in neuroblastoma, its generally high expression level suggested this gene could play some role via reported pathways in neuroblastoma development (**Fig. 19A**).

In the HYENA results, 6 (12.5%) out of 48 high-risk tumors had some form of somatic SVs near *EFTUD2*. To investigate if this gene was activated by the rearrangement of enhancers, the 3D genome interaction prediction was deployed. A translocation between chromosome 11 and 17 rearranged the regulatory sequences on *ANO1* gene body to the *EFTUD2* region, and induced new chromatin interactions to activate *EFTUD2* and the genes close to it (**Fig. 19B**). In

summary, our data indicated that *EFTUD2* is a potential enhancer hijacking oncogene in high-risk neuroblastoma.

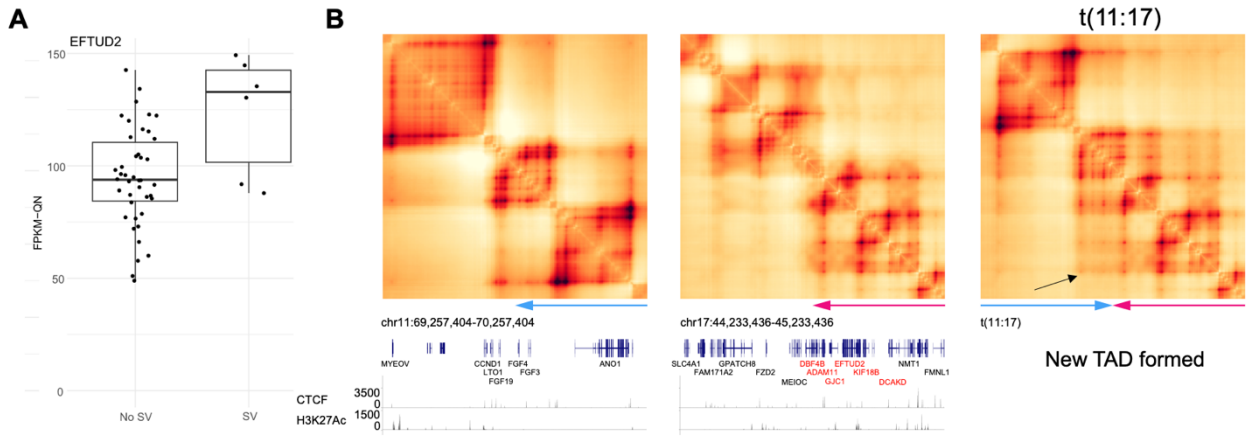

Figure 19. Expression levels of *MYCN*, *SLC6A18*, *EFTUD2*, and *SLC6A3* in high-risk samples.
A, Normalized expression of *EFTUD2* in samples with (n=6) and without (n=183) nearby SVs in neuroblastomas. The boxplot shows median values (thick black lines), upper and lower quartiles (boxes), and 1.5× interquartile range (whiskers). Individual tumors are shown as black dots. For each gene, tumors are grouped based on SV status (no SV or SV). Quantile normalized FPKM values are shown for each group. The boxplots show median values (thick black lines), upper and lower quartiles (boxes), and 1.5× interquartile range (whiskers). Individual tumors are shown as black dots. **B**, 3D genome structures predicted by deep-learning based algorithm Orca. Heatmaps show the predicted 3D chromatin interaction maps of the chromosome 11 SV partner (left panel), chromosome 17 SV partner (middle panel), and the translocated region in the translocation between chromosome 11 and 17 (t11:17) (right panel). H3K27Ac and CTCF ChIP-Seq data of neuroblastoma cancer cell line are shown below the interaction maps.

Discussion

In this study, we analyzed 189 neuroblastoma samples including low-, intermediate-, and high-risk cases, to call their somatic SVs with two algorithms, as well as integrate their gene expression profile and clinical information to predict putative oncogenes activated by enhancer hijacking using our analytical pipeline HYENA. Other than *TERT*, we identified four more candidate genes: *GZF1*, *NBAS*, *TTC32*, and *TRIP13*, to be putative oncogenes activated by distal enhancers in all risk levels of neuroblastoma (**Table 1**). When excluding gene copy information

in the analysis and using a recurrency cutoff of 3%, we detected 58 putative oncogenes in all risk levels, and 26 putative oncogenes specifically in high-risk tumors (**Table 2**, **Table 3**). Most of the candidate genes had SVs associated with *MYCN* or *TERT* SVs (**Fig. 17**). We identified that *EFTUD2*, an oncogene in other solid tumor types, is a potential oncogene in high-risk neuroblastoma activated by rearranged enhancers.

There are a few aspects to discuss here. First is the SV calling process. Published analytic algorithms that identify putative CRE rearrangement as well as our pipeline HYENA depend heavily on SV detection with WGS data to infer the CRE hijacking events [223]. Therefore, it is especially important to have high-quality SV calls for the purpose of infer gene deregulation. There are numerous published SV callers widely applied in cancer research now, including Meerkat [46], Manta [208], Delly [224], and novoBreak [225]. Manta is the tool used for the GMKF published somatic SV data. Although filtered and passed the quality control by the tool, the Manta SVs contained artefacts and it is reported this tool should not be used solely to get reliable SV calls [226]. In our analysis, we used both Manta and Meerkat. After manually checking bam file reads in IGV, we decided to filter out the Manta SVs with SR < 3 or PR < 3 to get reliable calls (**Fig. 12**), and this step significantly reduced the SV counts (2170 kept out of 8089 SVs provided by GMKF dataset). These Manta SVs were then taken union with Meerkat SVs (if the breakpoints of a Manta SV and a Meerkat SV fell within 50bp, they were considered as one SV) to generate the final SV input for HYENA running. When comparing the SV breakpoints near *MYCN*, we noticed there were two samples with *MYCN* amplifications, but without SV breakpoints around *MYCN*. This suggested that even we used two algorithms to call SVs, there was still the possibility that we missed some breakpoints. A future direction should be to apply more SV callers to have more sensitive and reliable SV results.

It has been discussed that HYENA's sensitivity can be limited by small sample sizes. In this study, we analyzed 189 neuroblastoma samples, but there were only 48 high risk samples, which is not a big size to detect enhancer hijacking genes specifically for high-risk group with a stringent recurrent rate cutoff. Although we performed the analysis without CNV input and with a loose cutoff, there could still be genes missed by HYENA. HYENA is designed to eliminate the effect of copy gain and identify the genes activated by SVs, so it would be ideal if including all information into the analysis instead of using a loose parameter setting. While high-risk neuroblastoma is especially in need of novel biomarkers to improve outcome prediction and develop new therapies, a future direction is to include more high-risk samples into analysis whenever there are normal-matched WGS data and tumor RNA-Seq data available.

Deploying the HYENA pipeline is just the first step of the identification of novel oncogenes activated by distal enhancers. To validate the functions, experimental approaches must be taken, and 3D genome interactions should be analyzed for the samples carrying interesting SV events. In the scale of this study, we did not perform any validation study, but a future direction should be to thoroughly investigate the functions of the oncogene candidates, especially *EFTUD2*, to understand how it involves in high-risk tumors and to confirm regulatory sequences are hijacked and activate gene transcription.

Further Discussion

In the first chapter, we presented a computational algorithm HYENA to detect candidate oncogenes activated by distal enhancers brought by somatic SVs. These SV breakpoint partners fell in potential regulatory sequences and caused shuffling of regulatory elements, leading to abnormal gene expression. The candidate genes we detected included protein-coding and non-coding genes. Our in vitro and in vivo experiments suggested that a lncRNA identified by HYENA, *TOB1-AS1*, was a potent oncogene in pancreatic cancers and promoted tumor metastasis. In the second chapter, we performed CRISPR activation screens to further explore the functions of the candidate genes detected by HYENA using the PCAWG dataset. With cell proliferation and migration screens, genes that can promote these phenotypes after transcription activation, to mimic the context of enhancer hijacking, were enriched and provided deeper insights for understanding the functions of HYENA candidates. We found *RCCD1*, a gene reported to drive breast tumor [186], can promote MCF-7, a breast cancer cell line proliferation; *POLR2F*, a subunit of RNA Pol II complex known to be overexpressed in multiple cancer types and involved in mechanisms of cisplatin resistance in gastric cancer [190-192, 194], can promote MCF-7 cell migration. In the third part of this dissertation, we further deployed HYENA and analyzed 189 neuroblastoma samples including low-, intermediate-, and high-risk cases, to predict putative oncogenes activated by enhancer hijacking. When using a loosened parameter setting compared to the PCAWG analysis by including high copy genes into the analysis and a recurrency cutoff of 3%, we detected 58 putative oncogenes in all risk levels, and 26 putative oncogenes specifically in high-risk neuroblastomas. In summary, HYENA is a robust tool to predict enhancer hijacking genes; our CRISPR screens added another layer of experimental validation; the application in neuroblastoma samples showed HYENA's capability to detect

putative oncogenes in not only adult cancers but also pediatric cancers. With further validations and functional studies, the candidate genes can be new biomarkers or therapeutic targets in the corresponded cancer type.

As a complement to the discussion sectors in previous chapters, here are some key points that I would like to further discuss on this dissertation.

Limitations of HYENA pipeline

The superior performance of HYENA compared to existing tools has been described in detail in its chapter, so here I focus more on the limitations. Although HYENA is a robust and sensitive tool to detect enhancer hijacking genes using cohort data, there are still limitations associated with how it was designed and how it can be applied.

First of all, the results rely on the datasets very much. The data size can limit the discovery when the gene SV frequency was not high enough for the gene to be tested in the model. Increasing data size or input multiple cohorts of uniformed data would work, but the inconsistency of how sequencing data was processed across different cohort studies makes it tedious and challenging to merge multiple cohorts together into one analysis. In addition, the quality of data matters substantially in HYENA analysis, especially the quality of SV calls and CNV calls (if including CNV in the regression), because the frequency of SVs decides whether a gene can be regressed, and the SV status around a gene is a coefficient in the regression model, and directly determines whether the gene is a candidate or not. SV false positives would lead to false positive prediction by HYENA, while SV false negatives would lead to HYENA missing candidates. If the CNV files include too many 'NA' values, the sample with NA copy number for a gene will not be included in the model, thus it will reduce the frequency of the gene SVs and

cause false negative due to missed information. Users should be very cautious when inputting existing data which are not generated by the users from raw data.

Second, the final FDRs were generated from permutation tests, which induce instability to the output. Depending on how many genes were included into the regression, different numbers (100 to 1,000) of permutation tests were applied in our analysis to calculate empirical FDRs and increase the reliability of HYENA results. However, when gene expression levels were shuffled in permutation tests, it induced uncertainty and results could be different for different users even from the same set of data. While other algorithms like PANGEA and CESAM do not include such tests, the instability here can be a drawback of HYENA. Increasing the number of permutation tests can potentially reduce the instability, but meanwhile it will increase the calculation workload and consume more time and resources.

Third, data interpretation, especially for significant non-coding genes detected by HYENA, should be approached with more validations. While including non-coding gene annotations in the analysis is an important advantage of HYENA, the fact that many non-coding oncogene candidates fell close to the cancer driving coding genes makes it hard to distinguish their functions from the passenger effects of the activated protein coding genes. For example, in **Table 3**, significant genes identified in high-risk neuroblastoma samples that were close to *MYCN* included *MYCNOS*, *MYCNUT*, *NBAS*, and *DDX1*. We could not conclude whether these genes play driver roles in tumorigenesis, but it is very likely that their activation was passenger effects of *MYCN* amplification. Another example would be the two non-coding candidate genes *RNU7-143P* and *SNORD65* close to *IGF2BP3* in thyroid cancer. We tested the effects of their overexpression or knockdown in thyroid cancer cell lines (data not shown), but the changes were either not significant or associated with *IGF2BP3*, which is an oncogene [141]. It cannot be

emphasized too much that all the genes predicted to be oncogene candidates with computational tools need to be investigated by manipulating gene expression levels and observing the effects in proper tumor models.

Unaddressed questions on lncRNA *TOB1-AS1*

TOB1-AS1 was identified as an enhancer hijacking oncogene by HYENA from PCAWG pancreatic cancer cohort, and our experiments demonstrated its functions in promoting cancer cell invasion and tumor metastasis. However, how *TOB1-AS1* drives these biological processes, and why its roles are dramatically different in pancreatic cancer [162] compared to that reported in other cancers [153, 154, 227] remain to be elucidated.

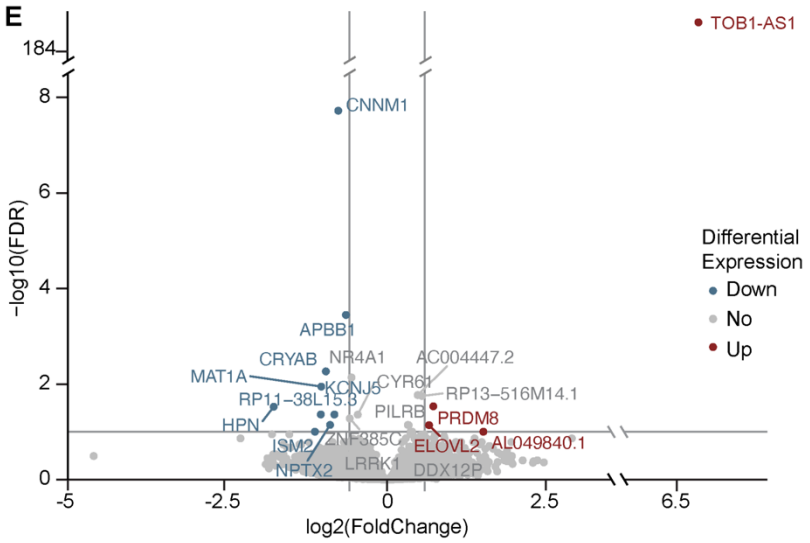

Figure 20. Differentially expressed genes caused by *TOB1-AS1* overexpression. Volcano plot showing the differentially expressed genes in *TOB1-AS1* overexpression PANC-1 tumors (n=6) compared to vector control tumors (n=6). Red and blue dots with gene labels represent significantly (FDR <0.1) upregulated and downregulated genes with fold-change larger than 1.5 and smaller than 1/1.5, respectively. Grey dots represent all other genes. Grey lines represent -log10(FDR) of 1 (horizontal), log2(FoldChange) of log2(1.5) (vertical, right) and log2(1/1.5) (vertical, left). This figure is adapted from the Supplementary Fig. S13E of Yu, *et al.*, *Nucleic Acids Research*, 2024.

To figure out the mechanisms of *TOB1-AS1* promoting tumor metastasis, we performed RNA-Seq on *TOB1-AS1* overexpression tumors and control tumors (**Fig. 17**). The results were

not very informative for us to reach a conclusion because there were no significantly enriched pathway, and the differentially expressed genes had diverse functions in cancers. For example, *HPN* encodes a type II transmembrane serine protease and promotes epithelial-mesenchymal transition and cell invasion in prostate cancer [228]. It seems to conflict with what we observed in pancreatic tumor models where HPN was downregulated after *TOB1-AS1* overexpression. Another significantly downregulated gene *CNNM1* is a cyclin and CBS domain divalent metal cation transport mediator and is predicted to be involved in ion transport [156], but its cancer driving functions are unclear. Therefore, the information we have for *TOB1-AS1* is very limited and further studies including proteomics or pull-down assays should be performed to understand how *TOB1-AS1* interacts with other genes to promote tumor metastasis.

TOB1-AS1 is detected from Australian pancreatic cancer cohort (PACA-AU) with 10% recurrent rate. However, in the analysis using other pancreatic cancer cohort like PACA-CA, which is a Canadian cohort, we did not see any gene overlapping with the candidate genes detected in PACA-AU (data not shown). Such results suggested that the output of HYENA can change dramatically across different populations, even for the same tumor type. This is likely due to the intrinsic genetic differences, age or sex compositions from different patient populations. It was unclear whether *TOB1-AS1* is activated by any mechanisms other than enhancer hijacking in other pancreatic cancer cohorts. A future direction should be to explore all the mutations (including SNV, indel and SV) that can lead to the expression change of *TOB1-AS1*, and to get insights from the genes associated with it to infer the possible oncogenic mechanisms of this gene.

Future directions for studying enhancer hijacking events

Here I discuss the future directions in the field of enhancer hijacking studies in two angles: computational approaches and validations.

As the importance of enhancer hijacking is emerging and more extensively studied, there have been a lot of tools detecting such events with NGS data. The computational tools take advantages of Hi-C, RNA-Seq, and/or WGS data, and can detect rearranged enhancers or super enhancers that activate oncogenes based on the data from either one sample or a group of samples. Tools including CESAM [93], PANGEA [95], NeoLoopFinder [125] and cis-X [96] have been introduced in the HYENA chapter. CESAM and PANGEA use linear regression and elastic net model (based on linear regression) to associate increased gene expression with nearby SVs. Cis-X and NeoLoopFinder can detect enhancer hijacking target genes based on individual samples utilizing heterozygous SNVs and Hi-C (or other chromatin interaction data), respectively. Since SVs can disrupt normal TAD structures and form new 3D genome interactions, NeoLoopFinder can be considered as the most appropriate model to detect enhancer hijacking events, validating the neo-interactions between gene promoters and distal enhancers. However, Hi-C, or other chromatin conformation data are limited for patient samples, and most studies were done within cell lines. This factor restricted the application of tools like NeoLoopFinder. Overcoming this limitation, tools like HYENA take more commonly available data, RNA-Seq and WGS, to build the models and predict enhancer hijacking, and the trade-off becomes the lack of validation that can confirm the SVs indeed change 3D genome structures and induce enhancer-promoter interactions. It would be a great breakthrough if computational approaches that can predict 3D genome interactions based on solely DNA sequences [128, 130] can be incorporated with the tools that can detect enhancer hijacking based on widely available RNA-Seq and WGS data. Therefore, as far as I am concerned, given an acceptable quality of

input files (such as SV calling), an important future direction would be to develop tools that implement WGS and RNA-Seq data from individual samples to predict the association between SV breakpoints and nearby genes, and at the same time to infer the disrupted genome interactions using the SV profiles and DNA sequences of SV partners. With inferred events from individual samples within a cohort, this approach can identify both individual events and recurring events associated with genes of interest.

For the purpose of confirming the robust performance of developed algorithms and improving the clinical practices, validation is a necessity. It involves confirming the promoter-enhancer interaction of a gene induced by SVs, and demonstrating the oncogenic functions of the gene. To confirm an oncogene is activated by distal enhancers, the SV partner should be annotated as enhancers or carry active enhancer markers like H3K27ac, and the gene should be inactivated without the SV or without an active enhancer located at the SV partner region. Large consortium studies like ENCODE [229] have annotated a large number of regulatory sequences, and ChIP-seq and CTCF binding profiles are available for a large amount of cell lines. The limiting factor is searching for a cell line carrying a SV that activates the gene of interest. With a cell line, we can use CRISPR/Cas9 to disrupt the regulatory sequences, to achieve the goal of validation for the hijacked enhancers. An induced deletion to model the removal of a TAD boundary can also be done, using a normal cell line, to investigate the consequences of a deletion observed in patient data. While this part of validation seems to be straightforward, revealing the oncogenic functions of a gene and underlying mechanisms usually take more efforts. Gene overexpression and KD can directly show how the gene expression levels affect phenotypes in vitro and in vivo, but only clarified pathways and elucidated oncogenic mechanisms of cancer driver genes can truly bring bench-side discoveries to translational applications.

In summary, future directions in enhancer hijacking studies should emphasize the importance of accurate SV input for reliable local assemblies, recognizing that the quality of these inputs is crucial for uncovering the full scope of chromatin interactions in cancer. Advanced tools like HYENA, which utilizes rank-based regression rather than traditional linear regression-based methods, offer sensitive and reliable detection of enhancer hijacking genes with WGS and RNA-Seq data, overcoming challenges posed by outliers or limited data availability. Moreover, the application of HYENA on neuroblastoma samples further supported that identifying enhancer hijacking events is helpful for discovering potential oncogenes and high-risk mutations. In the future, understanding the processes and underlying mechanisms could lead to the identification of novel biomarkers for cancer diagnosis and prognosis, as well as new therapeutic targets.

Appendix

Supplementary Figures

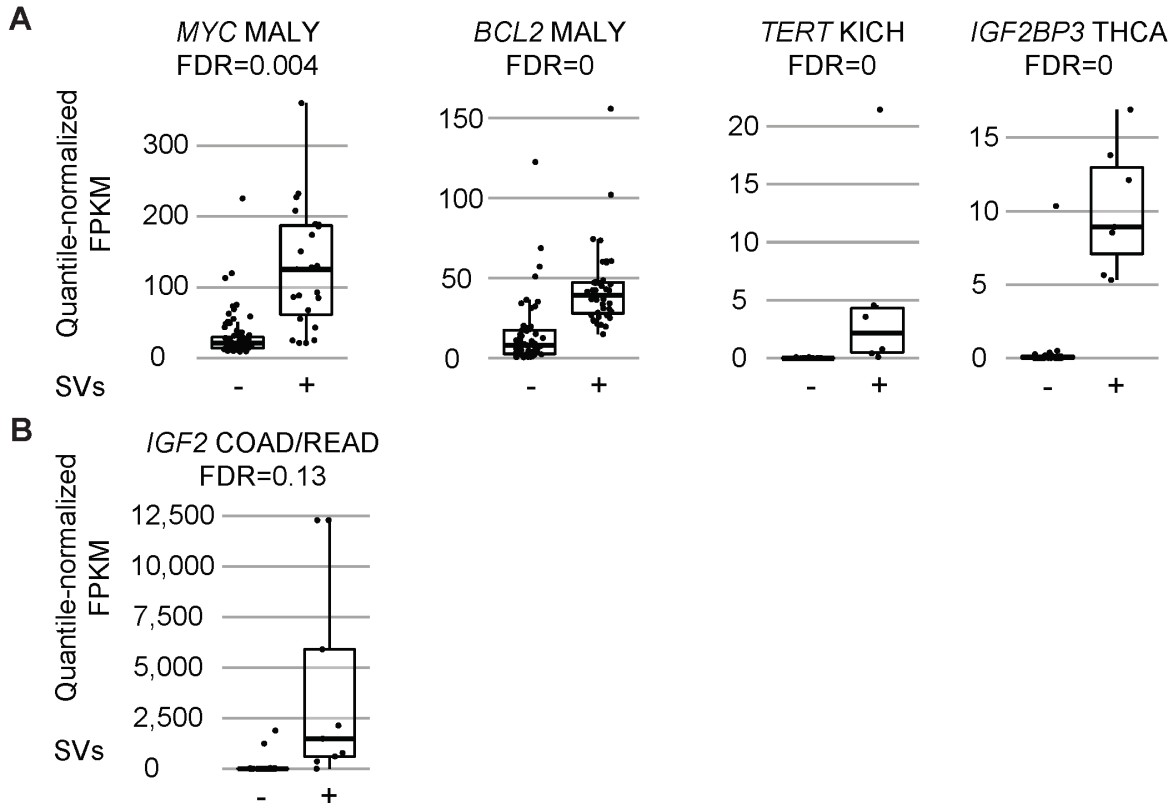

Fig. S1. Expression levels for five known enhancer hijacking target oncogenes. For each gene, tumors are grouped based on SV status (- or +). Quantile normalized FPKM values are shown for each group. The boxplots show median values (thick black lines), upper and lower quartiles (boxes), and 1.5× interquartile range (whiskers). Individual tumors are shown as black dots. **A**, Genes detected by HYENA. **B**, Gene not detected by HYENA.

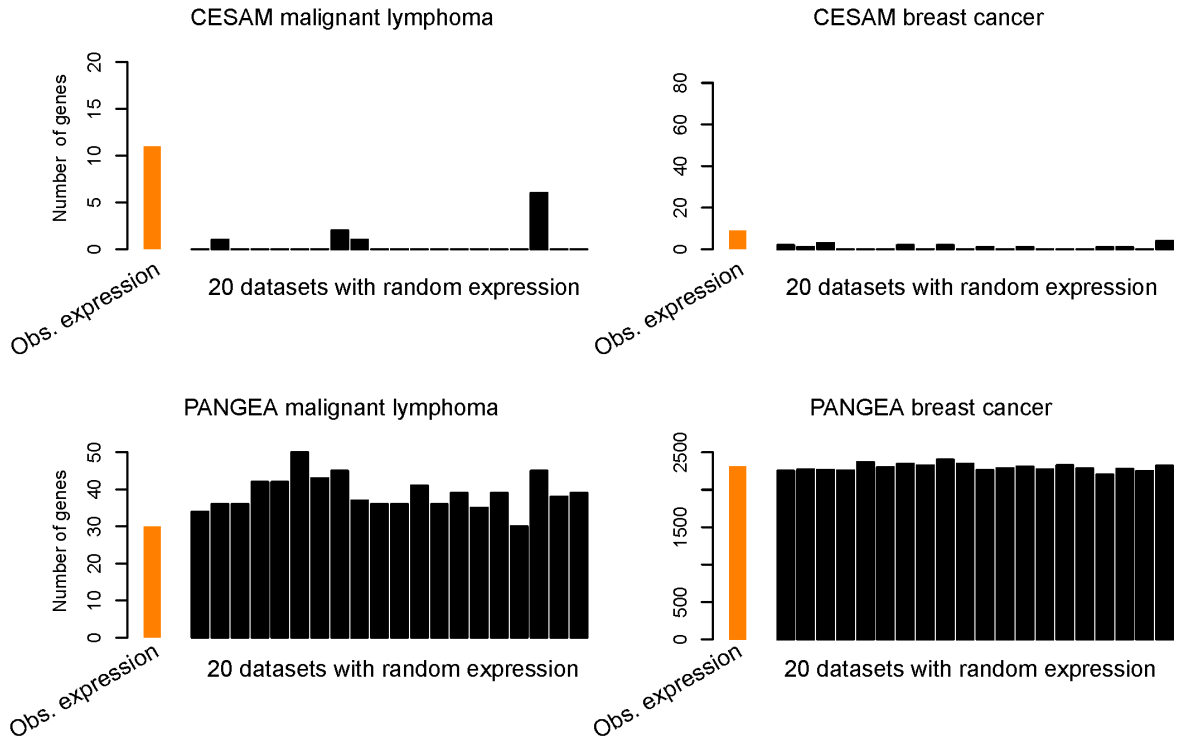

Figure S2. Numbers of genes detected by CESAM and PANGEA in two PCAWG tumor types using observed gene expression and randomized expression. Genes detected when expression was randomized were false positives.

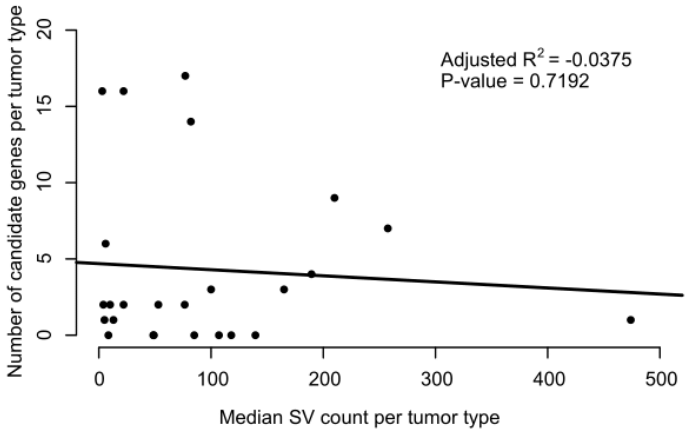

Figure S3. Number of candidate enhancer-hijacking genes detected by HYENA is not associated with genome instability. Scatter plot of median SV count and number of candidate gene detected by HYENA in each tumor type. One dot represents one tumor type. The line represents the linear regression with its statistics labeled at the upper-right corner.

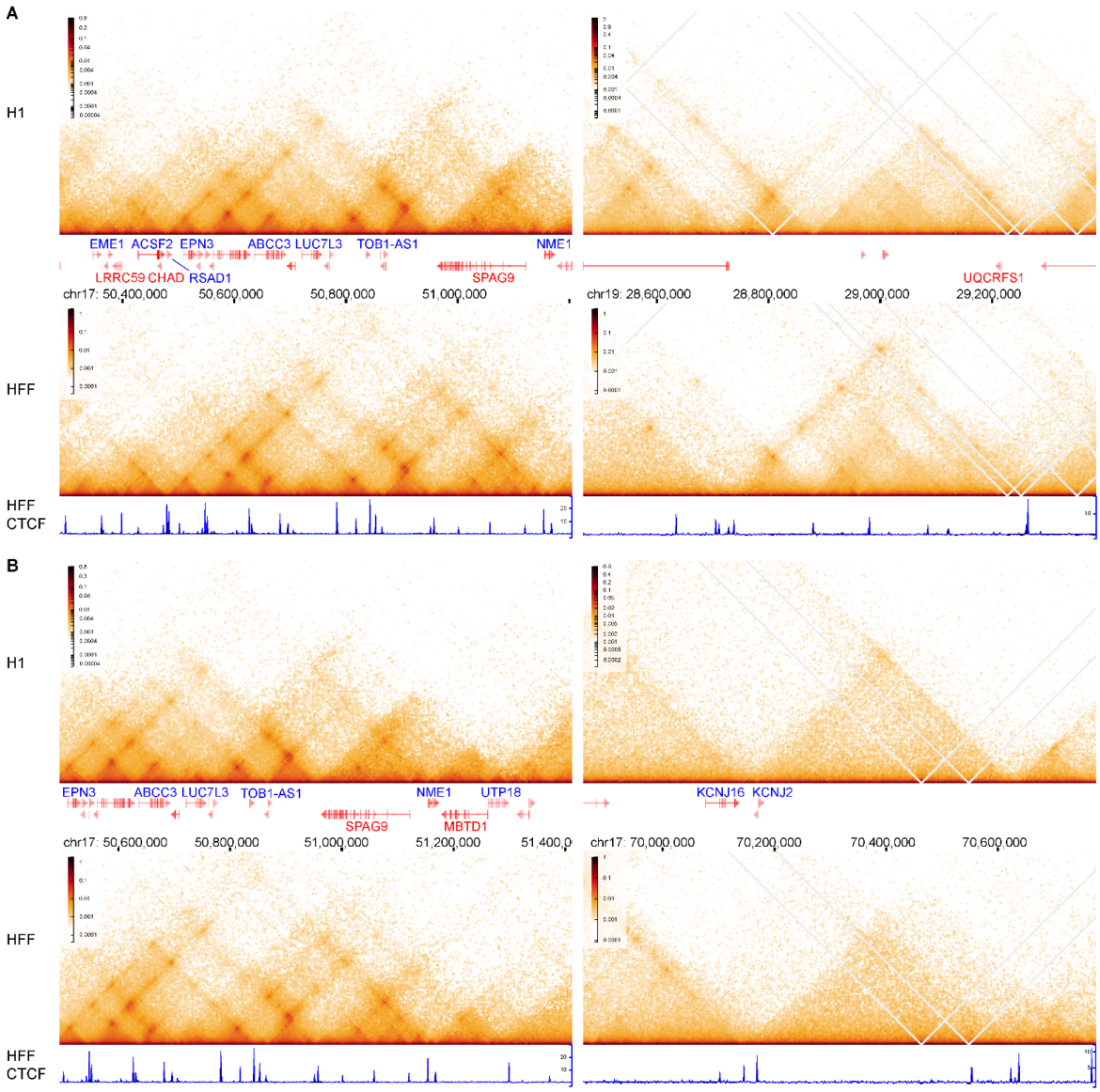

Figure S4. Hi-C maps of *TOB1-AS1*, *UQCRFS1*, and *KCNJ2* loci from H1 and HFF cell lines. A, *TOB1-AS1* (left panels) and *UQCRFS1* (right panels) loci. **B**, *TOB1-AS1* (left panels) and *KCNJ2* (right panels) loci. CTCF ChIP-seq of the HFF cell line is shown at the bottom. These experiment-based Hi-C maps are very similar to predicted Hi-C maps for the same loci in **Fig. 5D** and **5E** left and middle panels.

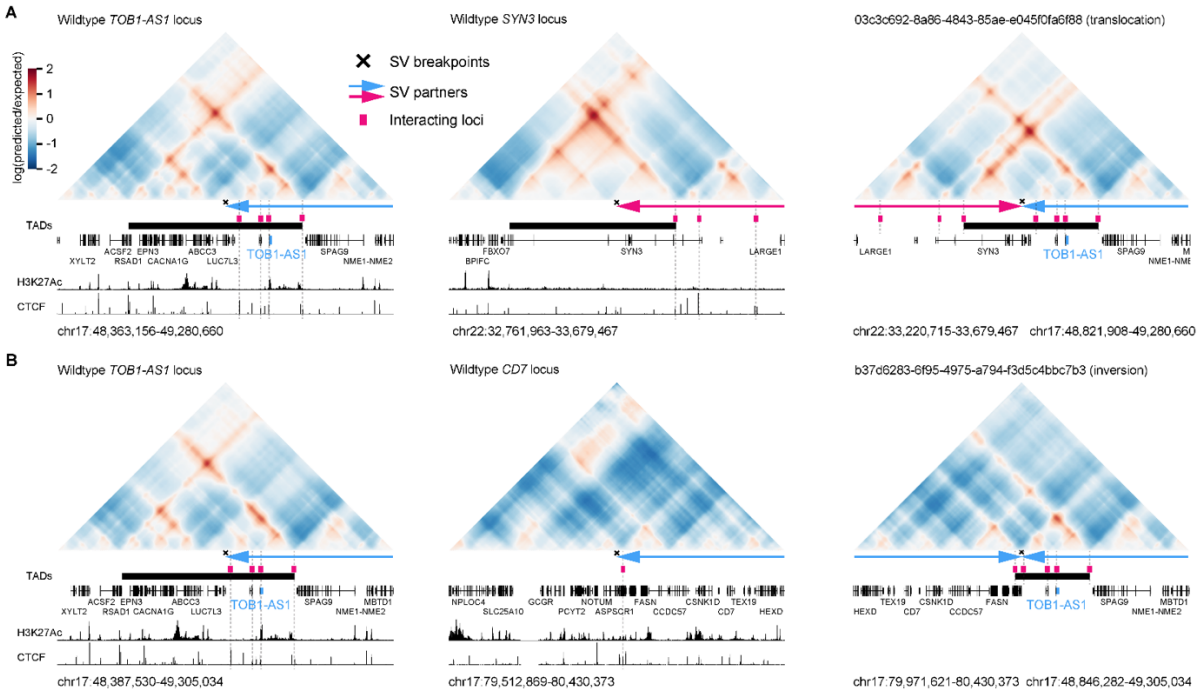

Figure S5. Predicted 3D chromatin interaction maps for two pancreatic cancers with SVs near *TOB1-AS1*. **A**, Predicted maps for regions without translocations (left and middle panels) and with translocation in tumor 03c3c692-8a86-4843-85ae-e045f0fa6f88 (right panel). **B**, Predicted maps for regions without inversion (left and middle panels) and with inversion in tumor b37d6283-6f95-4975-a794-f3d5c4bbc7b3 (right panel).

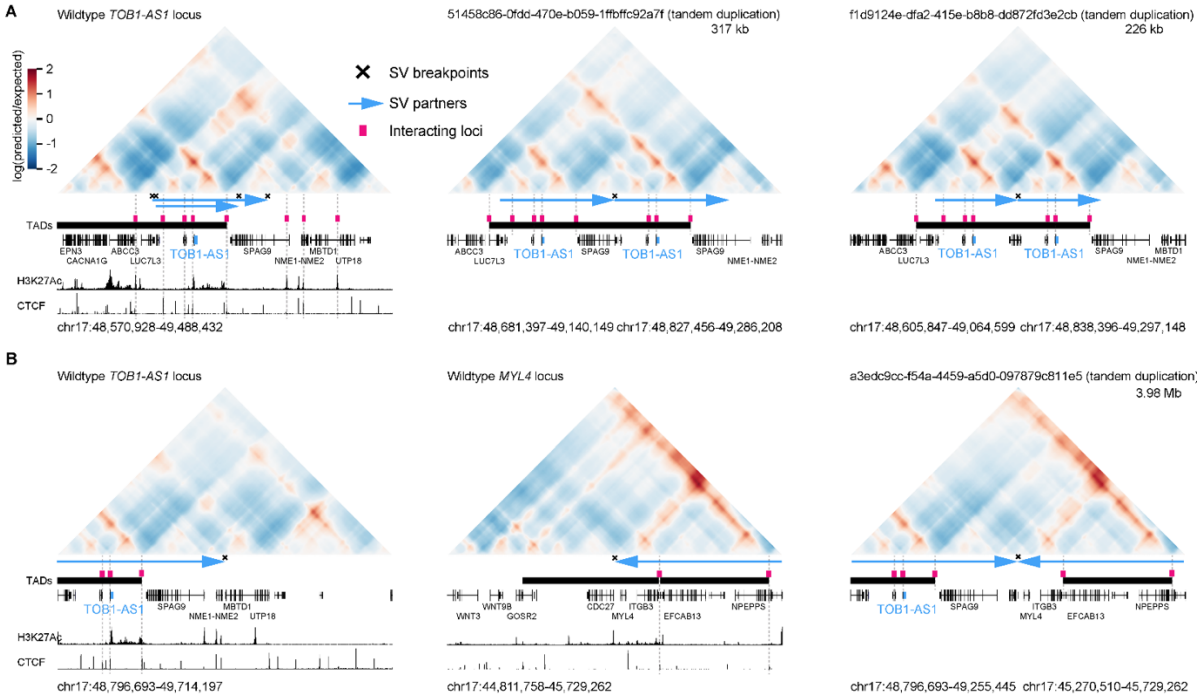

Figure S6. Predicted 3D chromatin interaction maps for three pancreatic cancers with SVs near *TOB1-AS1*. **A**, Predicted maps for regions without tandem duplication (left panel) and with

tandem duplications in two tumors 51458c86-0fdd-470e-b059-1ffbffc92a7f (middle panel) and f1d9124e-dfa2-415e-b8b8-dd872fd3e2cb (right panel). **B**, Predicted maps for regions without tandem duplication (left and middle panels) and with tandem duplication in tumor a3edc9cc-f54a-4459-a5d0-097879c811e5 (right panel).

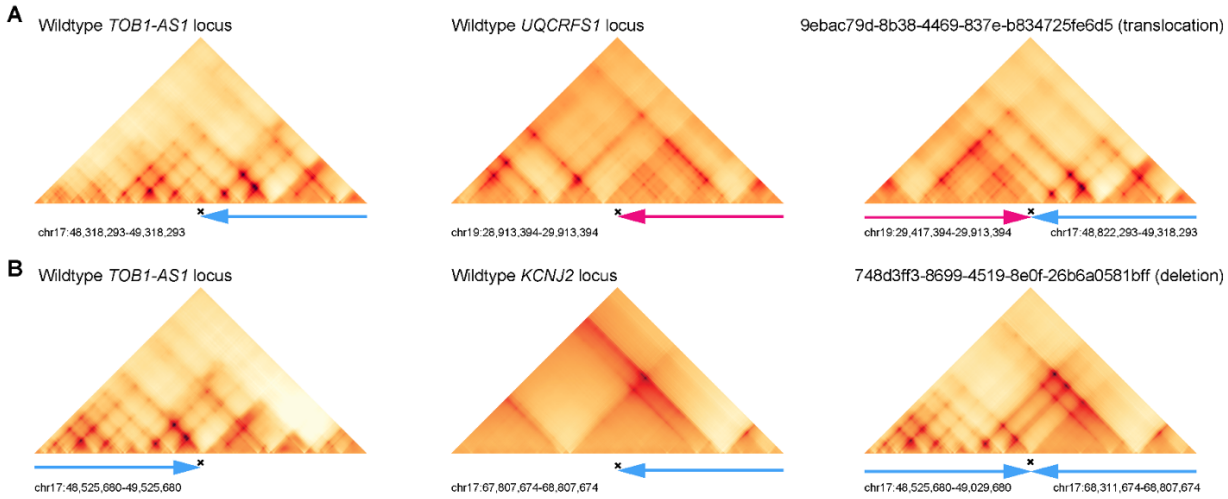

Figure S7. 3D genome structures predicted by deep-learning based algorithm Orca. A, Predicted 3D chromatin interaction maps of the *TOB1-AS1* (left panel), *UQCRFS1* (middle panel), and the translocated region in tumor 9ebac79d-8b38-4469-837e-b834725fe6d5 (right panel). **B**, Predicted 3D chromatin 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). The 6 regions in this figure are the same regions shown in **Fig. 5D** and **5E**.

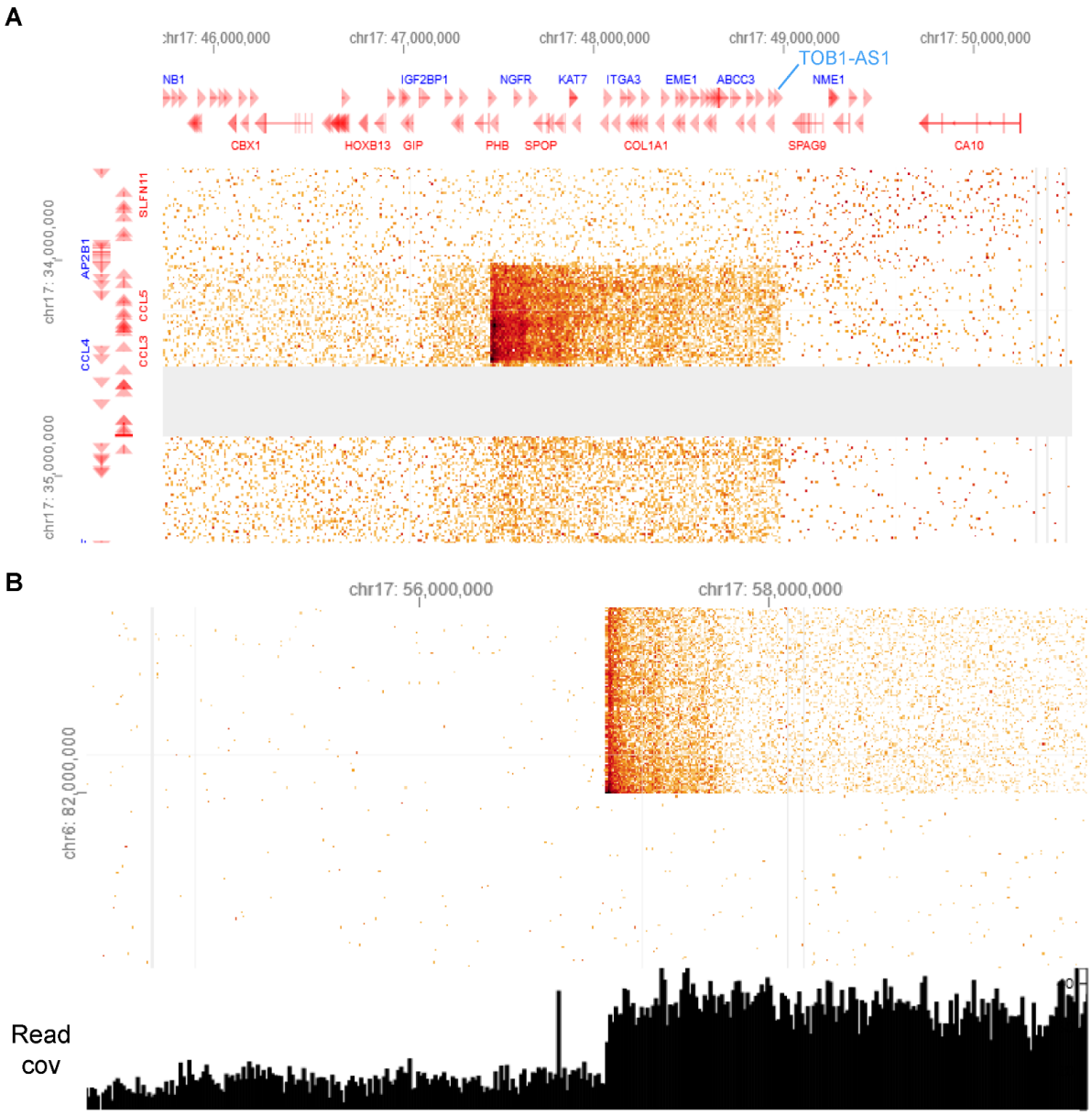

Figure S8. SVs in Panc 10.05 detected by Hi-C. A, HiGlass view showing a deletion of chr17:34,460,000-47,450,000. **B**, HiGlass view showing a translocation between chromosomes 6 and 17. Read coverage is shown below the Hi-C contact map. The chromosome 17 translocation breakpoint is 8 Mb downstream of the CNV breakpoint shown in **Fig. 6C** left most panel.

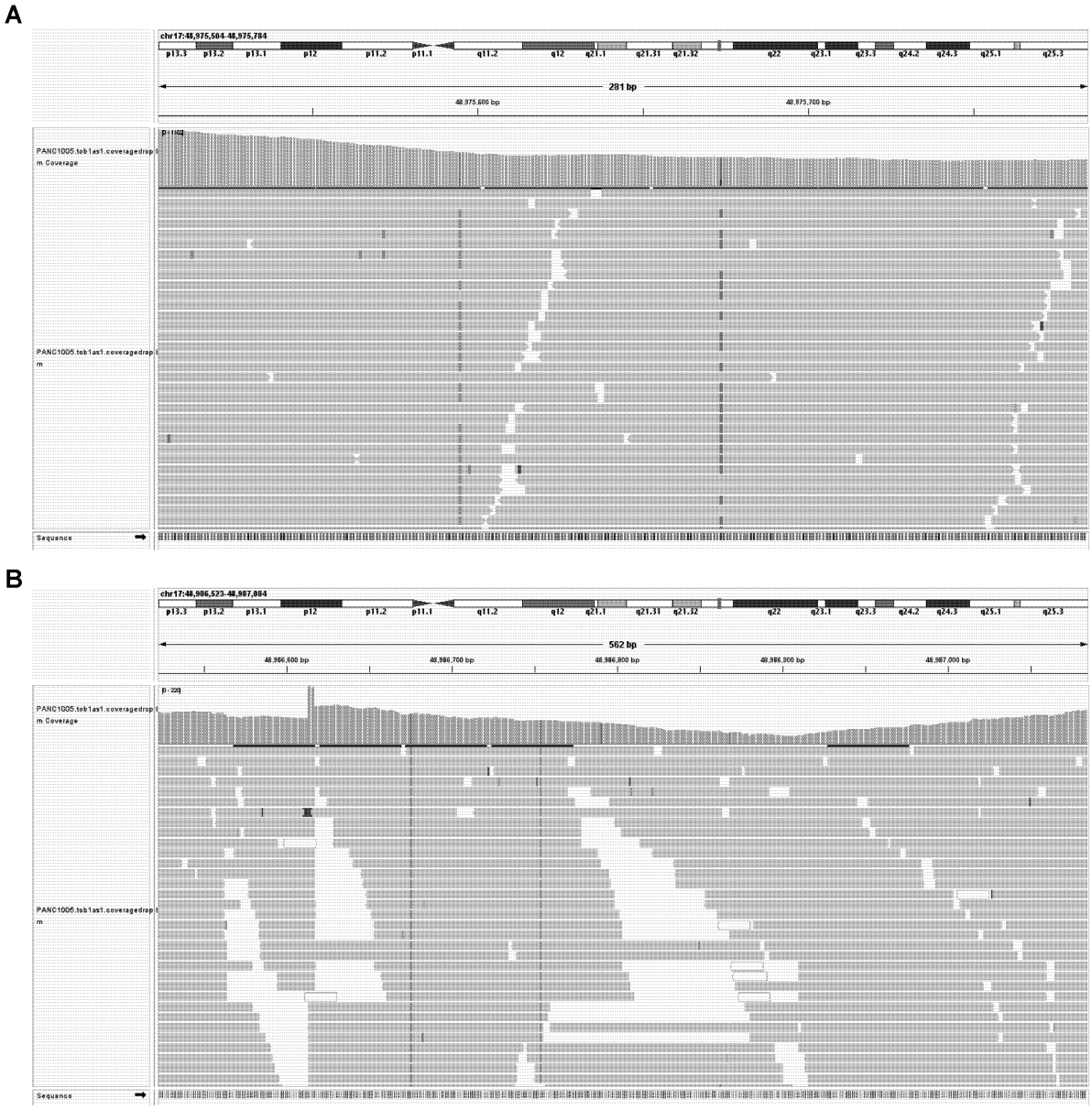

Figure S9. SNPs in Panc 10.05 near CNV and foldback inversion breakpoint. A and **B**, IGV screenshots showing reads mapped to five-copy and one-copy regions in Panc 10.05 in **Fig. 6C** left most panel. Horizontal grey bars are Hi-C sequencing reads. Colored lines are mismatches of reads compared to the reference genome. Grey vertical bars are read depth. Colored vertical bars represent SNPs. The two SNPs in **A** are heterozygous SNPs, whereas the four in **B** are homozygous.

Figure S11. HiGlass views showing the shared SV near *TOB1-AS1* in PATU-8988S (top) and PATU-8988T (bottom). The SV is about 50 kb upstream of *TOB1-AS1* and points away from *TOB1-AS1*. The locations of *TOB1-AS1* are shown in the x-axis at the top.

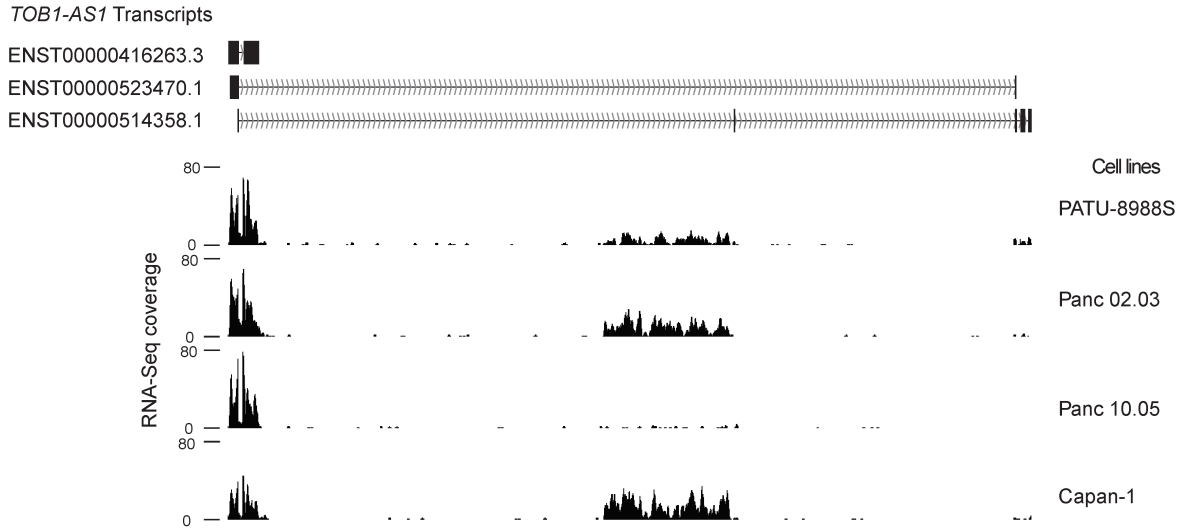

Figure S12. RNA-Seq coverage of *TOB1-AS1* isoforms. RNA-Seq coverage of three *TOB1-AS1* isoforms from four pancreatic cancer cell lines with high *TOB1-AS1* expression (PATU-8988S, Panc 02.03, Panc 10.05, and Capan-1). The major isoform is ENST00000416263.3.

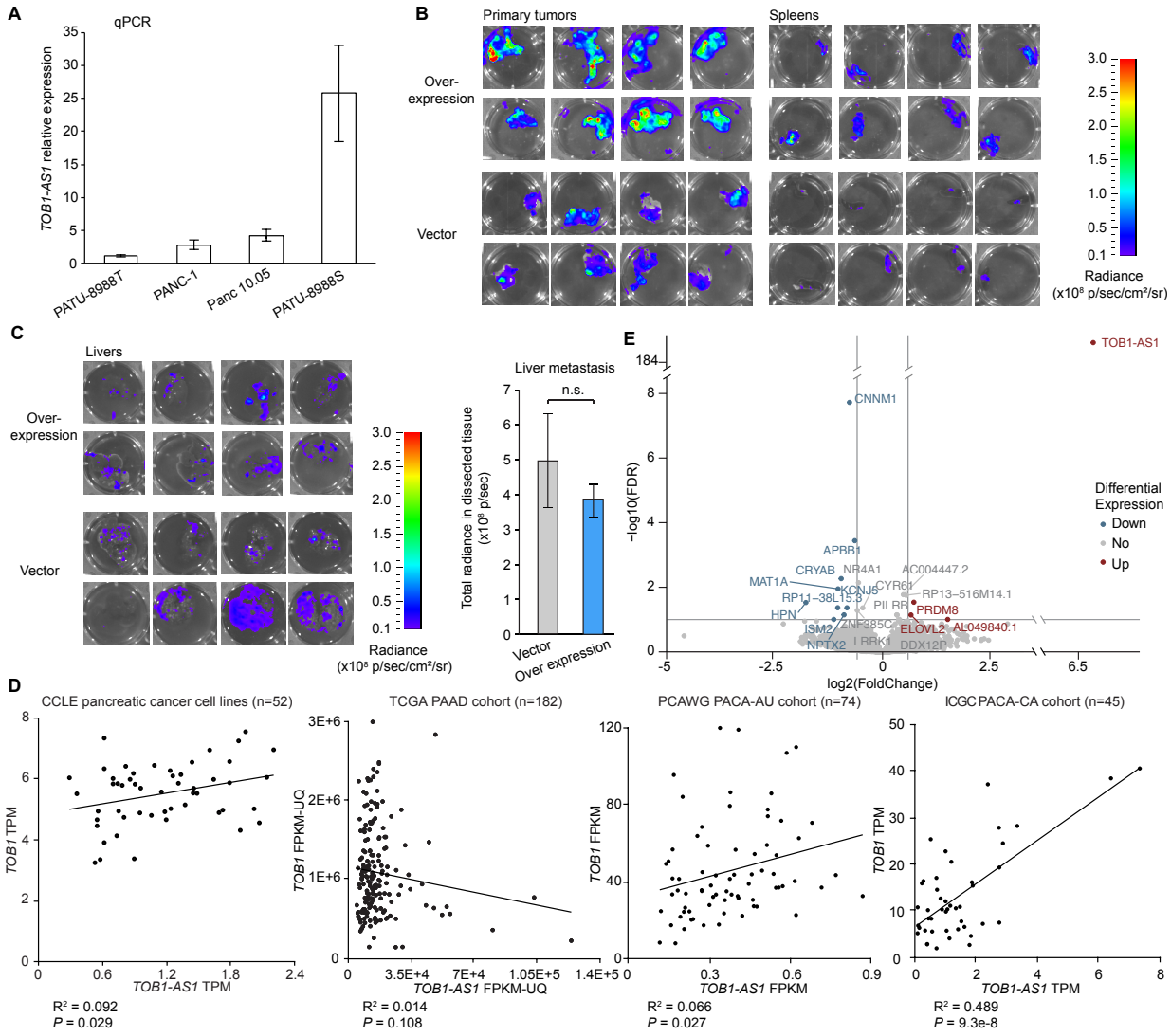

Figure S13. *TOB1-AS1* overexpression. A, *TOB1-AS1* relative expression levels in PATU-8988T, PANC-1, Panc 10.05, and PATU-8988S cell lines based on quantitative RT-PCR. The relative expression of the other three cell lines was calculated relative to PATU-8988T. Error bars indicate standard error of the mean. **B**, Ex vivo IVIS images showing primary tumors and spleen metastatic tumors from mice orthotopically injected with PANC-1. **C**, Ex vivo IVIS images and radiance quantification (p/sec) of whole wells showing liver metastatic tumors in mice orthotopically injected with PANC-1. Two-sided student t test was used. Error bars indicate the standard error of the mean. **D**, Scatter plots showing the correlations between *TOB1* and *TOB1-AS1* RNA expression in CCLE pancreatic cancer cell lines, TCGA PAAD, PCAWG PACA-AU, and ICGC PACA-CA cohorts. Sample sizes, gene expression normalization methods, squared-Rs and *P* values are labeled. In CCLE cell lines and the PCAWG PACA-AU cohort, the two genes have very weak positive associations with marginal *P* values of 0.029 and 0.027. In the ICGC PACA-CA cohort, the two genes have a strong positive correlation. However, the correlation is mainly driven by two outliers. On the contrary, in the TCGA PAAD cohort, the two genes are not significantly correlated. Therefore, *TOB1-AS1* and *TOB1* do not have consistent associations in patient samples and cell lines. **E**, Volcano plot showing the

differentially expressed genes in *TOB1-AS1* overexpression PANC-1 tumors (n=6) compared to vector control tumors (n=6). Red and blue dots with gene labels represent significantly (FDR <0.1) upregulated and downregulated genes with fold-change larger than 1.5 and smaller than 1/1.5, respectively. Grey dots represent all other genes. Grey lines represent -log10(FDR) of 1 (horizontal), log2(FoldChange) of log2(1.5) (vertical, right) and log2(1/1.5) (vertical, left).

Supplementary Tables

The supplementary tables for the published part of this dissertation can be downloaded at NAR Online (https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkae646#supplementary-data).

CRISPRa oligo library

Gene ID	Oligo Sequence
ENSG00000238098	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGACAGAGCAGTGCGCAAGCACGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000238098	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCGCGCCCCCTGCTGGACATGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000238098	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCCCTGCTGGACATAGGCGCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000236871	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCATACAGAGTGCTGCACGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000236871	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGTTCCCTCCCAGATACGAGGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000236871	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTTGTTCCCTCCCAGATACGGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000257818	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGGATTTAGAGGCCATTTCCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000257818	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCTTCCTCTTCGCCCCCAGGGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000257818	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGATGTGAGAAATAGACTTCCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000206775	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGATCAAGGGCTGGGTCAAGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000206775	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAATGTTGGGTGACAGAGAAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000206775	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGAATGTTGGGTGACAGAGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000253974	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGTTACCATAGAGATTGCACGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000253974	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAACTAATAGAAAGTGAAGCAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000253974	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGGCGCCTGTGCAATCTCTAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000203999	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGCGGGCCTGCAGCTCAAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000203999	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTCGTCACAGCTGAGCCGGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000203999	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCTCAAGAGGCAGGAGATGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000270852	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCAGATAGGGTAGTCCTGTTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000270852	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCCAACAGGACTACCCTATCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000270852	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCTGGCAATATGTTCCCAACGTTTTAGAGCTAGGCCAACATGAGG ATCACC

ENSG00000240364	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGGTGATTTACTGACATGCTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000240364	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGAGTTCGAGTCTGCAGAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000240364	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAGAGGGAGGCCTTGTCTCTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000207805	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGCAGGAGCCCCATCACGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000207805	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGCAGGTGCCATCAGCCGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000207805	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAAAGGGCAGGTGCCATCAGCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000237456	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGAGGGTGGGCTAGATCTAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000237456	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGTCTGAATACCTGGAGTCAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000237456	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCTAGAGATGTTGTGTTGCCCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000261472	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAAAGGGGACCCGTTGTTCCTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000261472	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCTCTGATCACCAGTATGTCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000261472	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCGGACATACTGGTGATCAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000261014	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGGCATTGTGACATAATATTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000261014	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTAGGCTACTATTGACCTTCCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000261014	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCCGGGGATACAGCCAAAGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000187791	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAAATGACGAGGTGTAAAATCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000187791	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTGATGCTATCCCTGCCTTGGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000187791	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAAAGCCCTCCTGTTCCACAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000206921	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCTAATCTGGCAAGAATCTAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000206921	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTAGAGGAGAAGGTGGATTGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000206921	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGATTGTGGGATGAATAAGCAAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000220204	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAAACTGCCTCAAAGAGGTTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000220204	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGCATTTAGCAAATAAGTGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000220204	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTGAGAATCAAACACAATATGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000240216	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCACAACACACTGGATCGTTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000240216	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGAAATGTTGCACAACACACGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000240216	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGATCAAGGGACAGTGGATGATGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000265912	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGAGTATATGTAAAGACCAAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000265912	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGGTCTTTACATATACTCAGGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000265912	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAAATGCTGAAAAGGATGAGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000267766	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGGGCGAGTTCACTCAGGCAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000267766	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGAGAAGGGCGAGTTCACTCGTTTTAGAGCTAGGCCAACATGAG GATCACC

ENSG00000267766	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAATACAAGCAAGAATGTCCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000172965	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCATGTGGACGTCCTTGCCCCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000172965	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGTCACGTCTGAGCTGCCAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000172965	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGACACCAGTGATATGGGATCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000284130	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCATCGACTGGCGTCTGCCAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000284130	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGAGTGTGGCGTCCATCGACGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000284130	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCGTGGCAGACGCCAGTCGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000197475	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCTCCGGTGCAGCCCACTGTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000197475	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGGTGCAGCCCACTGTGGGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000197475	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAGCTCCGTGGATCCCGAGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000199755	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAATAATCAAGGGCCAGGCAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000199755	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAAAAAAGAATAATTCAGTCTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000199755	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTTAAAGAATAATCAAGGGCCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000212347	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTTGGTAAATGAAGAGAGTTCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000212347	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGTTGCTGAGGGCTGTAGATGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000212347	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGCGTATCAAGTTTCAGTTTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000229980	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGGGCCAATGAACCGACGATGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000229980	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAAGTCCCGCCTGCCCCTGTCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000229980	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGCAGGCGGGACTTGGCCAAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000266913	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGGCTCTTGAGGCTTCAACCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000266913	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGACTCATTATATTCTACGCCTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000266913	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCAAGAGCCGGGACTAGGGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000260835	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGATGACGTTGGCCATCAGCACGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000260835	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTATGACGTTGGCCATCAGCAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000260835	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCTGCTGTGTGACATGCTCAGGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000264587	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGTGTGTGTCAGTATACAAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000264587	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTCTAGAAACTCAAATTCAAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000264587	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCACACTCCAAAAAATATATTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000254321	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCCCCTCCTGCAGTCCCAAGGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000254321	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGGGACAAAGGGGAACTCCCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000254321	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGTCCCCTCCTGCAGTCCCAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000239699	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTTAGAACTCAACACAACATGTTTTAGAGCTAGGCCAACATGAGG ATCACC

ENSG00000239699	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGAGTTCTAACACAATGGATGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000239699	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTAGAAAAAAAGGCACATTCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000258077	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAAAGTGTGAAAGGACTCCAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000258077	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTTCACACTTTCAAACCTGCTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000258077	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCAAACCTGCTAGGCAGGCCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000212264	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCTACCCACCTTAGCTCTGGTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000212264	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGCCCTACCCACCTTAGCTCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000212264	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGTTTCTCCCACCAGAGCTAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000232627	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAAGAGTAATAAAAGTTAAGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000232627	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCACACACATACAAATGATACGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000232627	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTATATAAATATTTATCACTTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000224265	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGACCACTAGACCTCCGTGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000224265	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCTGTTTGTCAGGAGGAGTTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000224265	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGACTCCTCCTGACAAACAGACGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000236654	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTGTAGAGAGGAGAGTGAACGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000236654	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGTAGAGAGGAGAGTGAACAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000236654	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGATGCTAGAAAAACAGCTTGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000232818	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCGGATGACGCCAGTGCAGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000232818	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGATGACGCCAGTGCAGAGGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000232818	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGACGCCAGTGCAGAGGGCGGGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000252590	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCCGAGTATAAGACAATCAAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000252590	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGACAGCCTCCCAATCCTCGGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000252590	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGAGGCTGTCAAAGCTCATGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000249628	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCTTTAGGCAGGAAGCCAGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000249628	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGTGCCCAGGCGGGACAGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000249628	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCAGCCCCATGGGAGTGCCCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000229630	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTATTGACCGTGATTCTAATCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000229630	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCGCGTCTGAACCCATAAGTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000229630	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGCTCACCCGATTAGAATCAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000234800	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTAGAGGATCATGCCTGTAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000234800	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGAAGTCCTTAATGACCTGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000234800	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCAGCTCCATCAGGTCATTAGTTTTAGAGCTAGGCCAACATGAGG ATCACC

ENSG00000218676	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGCAAGTCTAATCCCACAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000218676	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCTTGTGAGAGGCTTCTCACGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000218676	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGGGCAAGTCTAATCCCACAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000227616	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTAGGAAATGGCTAGCAATCAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000227616	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAAATGGCTAGCAATCAGGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000227616	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTAGGAAATGGCTAGCAATCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000259676	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGATATTCTGCTCCTACCGGTAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000259676	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGATGTTTGGCCATACCGGTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000259676	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCAACGATGTTTGGCCATACGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000254364	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTAGCAGGTGCACCCTTGTACGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000254364	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTAATAGGTTAGTGCCTCAGCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000254364	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGAGGCACTAACCTATTAAAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000258384	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGGGCGGGAGGCCTTTCCGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000258384	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAGCGGGCGCCGACAGCCACGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000258384	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGACGGGGAGAAAAGGTTTGAAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000273312	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGAGCCTCAGGCCACGCGGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000273312	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCCGGACTGACGAGGAGGGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000273312	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGACGCCGGACTGACGAGGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000227045	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGACTGTTGGACATCTATACGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000227045	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGATGATGATGTAGATATGGATGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000227045	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTAATATTTGGGGCACTGTTTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000246228	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTCCCAGAGATGGTGAGCTCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000246228	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTTGGCTGAAGCTCAATTGCAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000246228	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTTGCAAGACAGTCCCAGAGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000230613	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCTGGAGGGAATCCAAACGCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000230613	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCTGGAGGGAATCCAAACGCAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000230613	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGCATTCCGGGCCAGGGCGCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000256944	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTAAACCAGCGCCCCGAGTTGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000256944	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCAAACCCGCGCCTCAACTCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000256944	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGATGAGCTCGGAGACTAGCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000232874	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTTGTCCCCTAGGTATCAGAGGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000232874	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGACAAACCACTCTGATACCTAGTTTTAGAGCTAGGCCAACATGAGG ATCACC

ENSG00000232874	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAAACCACTCTGATACCTAGGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000248676	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGGGCTTCTGATTCCATCAGGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000248676	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGACTGAACGACCACAGAGAAAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000248676	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCTTCTGATTCCATCAGAGGGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000258837	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGGAGAAAATACAGTAGTGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000258837	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCTGTGAATCTGCTTAATGTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000258837	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAATCAATTGCTACATAGACCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000263279	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGATTACTTAAGCTCAGCCTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000263279	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGACACAGCCTGTGTTCTACTAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000263279	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGAGATTAAAGGCATGAAGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000204581	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGATCTTAACAACATCAGTGGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000204581	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAGACATGACCACCATGTGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000204581	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGATGGGGAATATAGAGGGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000240498	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGGCGCCCGCGCTGAGGGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000240498	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCATCTTCCCACCCTCAGCGCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000240498	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGGGTGGGAAGATGGTGGTGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000272181	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAATATCCAATCGGTGACCCAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000272181	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAAGACATATGACAAAACTGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000272181	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGATAGAATACAATATCCAATGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000269019	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGAAACCGGAGCCCCTGAGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000269019	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCTGAAACCGGAGCCCCTGAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000269019	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCCGCCCCACCCCCACTCAGGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000114779	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGTCTCACATGCCCCAAACGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000114779	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGACTGCTCCATGTCCATGTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000114779	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGATGTGAGACCGAGAGGCTTTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000167107	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCACCCTCGAAGAATAGCCCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000167107	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCGAATCCGCCGCACCCCATGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000167107	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGGGTGGCCTCTGCCTATGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000183773	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCTCCTCCCAGCTCAAGCTGGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000183773	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGCCAGTCCACTGGTCCCCAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000183773	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCTCAAGCTGTGGCCAGGAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000125449	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTTAGCCTGTGCTGGAACCAGGTTTTAGAGCTAGGCCAACATGAG GATCACC

ENSG00000125449	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGCGGCGAGCGGCTTCGCTGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000125449	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTGCTGGAACCAGCGGCGAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000239388	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTTAATGGAATGGTCACTACAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000239388	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGATTAATGGAATGGTCACTACGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000239388	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTAATGGAATGGTCACTACAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000171791	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTTACGCACAGGAAACCGGTCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000171791	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTTACGCACAGGAAACCGGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000171791	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGTGCAGAGAATGAAGTAAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000099385	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGCGTCCAGGCGCTCCAAGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000099385	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGCGTCCAGGCGCTCCAAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000099385	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCCCCGCCCCGGCCTGTCCCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000197299	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTTGGAGATACGCGTCCCTCCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000197299	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCCGCCCCAGCAGCCTGAGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000197299	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCGGGGTCGCCGTACAGCGCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000165714	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGACCCGCTCCTGCGACTTAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000165714	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGATCGCCCTAAGTCGCAGGAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000165714	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTAGGGCGATGCCACCTTAAAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000128346	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTCGTAGTGTCGCTGTTTGCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000128346	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAGCGGGGTCGCTGAGGCGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000128346	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCGCAGGCGCAAGATAAGCTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000107159	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGATGGAGCCAAAGTCTCACGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000107159	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCCATACCAAAGCTAGGATGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000107159	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGTATGGGGGAGAGGGCACAGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000105173	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGTCCCGCGCGGCCGCTGAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000105173	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGCGGCTCCCAGCCCCTCAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000105173	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCCGTCCCTCGCCCCGCCCCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000120217	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCTGACCTTCGGTGAAATCGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000120217	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCAGTTTAGGTATCTAGTGTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000120217	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTCCCGCCCACCTCTGCCCAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000125726	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAGTCTGAAGATCCTAAAGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000125726	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGGGACTTGAGCAATTGGCGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000125726	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCAGACTGGCAGCGGTTGGAGTTTTAGAGCTAGGCCAACATGAG GATCACC

ENSG00000117399	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTGATAGCTGAGACTTTCCCGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000117399	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCGGAGAGGCCAATGGGCTAGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000117399	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGCTAGGGCAACGGTTGCGAGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000168802	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTGGGGCCGCTGGTGAGTTGGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000168802	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGAGAGGCGAGCACCGGGAGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000168802	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGCGGAGAGAGGCGAGCACGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000138433	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAGCTACCCGAAGTGACAGGGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000138433	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCCCAGCTACCCGAAGTGACGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000138433	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTACCCGAAGTGACAGGTGGAGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000092853	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGATAGGAGATTGGGCGGCCAGGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000092853	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGCCTGGGTAATAGGAGATTGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000092853	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCTGGGTAATAGGAGATTGGGGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000107175	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGTTCCCTCCAACTGAGAAGGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000107175	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGTGCAGCGCCACGTCCCAGGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000107175	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAGGGGCTATGCAAATGTAGGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000099942	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCCCACCCTGATCGTCGCGAGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000099942	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGAGCACCGTCGCGACGATCGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000099942	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCGCTGTGTGACGTAACGGGGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000174177	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCCGCGCTGACGTCATCGTAGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000174177	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGCCGCGCTGACGTCATCGTGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000174177	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGACGTCAGCGCGGCGCAGTAGGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000205279	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGACAGAGCACTGGTAGATTTAGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000205279	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTAACATTCAAAAGGAGTACGGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000205279	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTTAACATTCAAAAGGAGTACGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000115866	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGGTGGCTGGCTGTAGACCTGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000115866	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGAAACTACCGCGAATTCCCGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000115866	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGGGAATTCCCGGGAATTCGGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000150990	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGACAGCCCAGGACTTAAACTCGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000150990	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGACCCCAAGCGAGAACAGCCCGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000150990	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGACTATTTAGAGAAGTAGGGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000186047	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGGTCTGCAGAGCCACCATTGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000186047	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGGCTGATGGAGGCCACTAAGTTTTAGAGCTAGGCCAACATGAGGATCACC

ENSG00000186047	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGGAGGCCACTAAGGGCGCAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000129295	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGGCAACGGGGACTCTACGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000129295	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCTGCTGAGCCCCTTCACTGGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000129295	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAGAGGAGACCGCAGTGAAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000177692	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAAATGATGTAAAGACCGAGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000177692	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCTCTCGGGAGGGACTTAGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000177692	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAACTCCCCACGCCGCCAACTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000107223	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGGCGCCGGCGACGTAGGGAGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000107223	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGACGTAGGGAAGGCGACGTCGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000107223	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGCCGGGCGCCGGCGACGTAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000196411	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGCAGCTCTCCCAACGCGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000196411	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAAGCGGAGAGGGGCACCGAGGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000196411	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGACGCTACTGAATAATTCATGGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000117868	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGATACTGACGTCATCGCGCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000117868	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTAGTCTCCGACCCGGCCCAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000117868	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTATCCACCCCGCCCGCTCCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000139083	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGCGCGGGCGGAGGAAACCCGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000139083	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGAAATAAAAGCTGCGCGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000139083	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTGGGGCCGCGGCTGCGAGCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000164002	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGACAGACAACGGCGCAAACAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000164002	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAGACAACGGCGCAAACAGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000164002	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCTCTGGGAAGGCGGTCCGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000123737	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCAGACTCAAAGCGTGATTGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000123737	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGACGGCCTCCCAAAGCCGCAAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000123737	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCGCAGACTCAAAGCGTGATGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000173727	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGCCCCAGGGCTTCCAGGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000173727	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGACCGACTAGAGCACAGTGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000173727	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGCCCTGGGGCTCTCCGAGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000167244	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCCGCAACCCGAGCCAAGAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000167244	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCGCAACCCGAGCCAAGAGCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000167244	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGCAACCCGAGCCAAGAGCGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000136231	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAGAAGGGTGACTGGCAGGAGTTTTAGAGCTAGGCCAACATGAG GATCACC

ENSG00000136231	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCCGCTGGCCGAAGCCAAAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000136231	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGGGTGACTGGCAGGAGGGAGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000143061	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCCGGGGCACCGGAGGACCTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000143061	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGAGGCCCGGCCCCGCCCCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000143061	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGCCGGGCTGGAGCTGGAGGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000161405	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCCGCTGTAACCCCGCGCACGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000161405	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGCAGGAGCCGGCGACCTGCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000161405	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCCGGTGCGCGGGGTTACAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000113430	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGACCTCACCTACCTGCGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000113430	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAGGCGGAGCACCTGGCGAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000113430	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGTTGTCGCCGGCTCCCCGCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000079999	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGCGCCCGTCGGCGAGGAGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000079999	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGATGCCCCACTCCTCGCCGAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000079999	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCGGCGAGGAGTGGGGCATCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000142687	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGCCGGCCGCGAGACCAAGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000142687	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAATGCCGGCCGCGAGACCAAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000142687	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCCCCCGGGAAGGGCAATGCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000124702	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTAGTCCATACGTTTCCTGAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000124702	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGACGAGAGGGCTGAGGAGTGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000124702	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGACGAGAGGGCTGAGGAGGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000187905	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGCTGGACGGCACATTCGGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000187905	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAATAGCGACGCGCTCTCCCGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000187905	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCGGGAGAGCGCGTCGCTATGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000146006	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGACCACACAGAACTGTATATGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000146006	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCATGCTTTGTAACAGCATCGGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000146006	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGTATATAGGCTGACGTCACGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000176204	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGACAGATTGCACATTAAAGACGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000176204	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTATAACTGAAAGGAACTCCTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000176204	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGATACACTGAGTGCCCTCCGCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000173212	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGATTTATATCAGGCTGTGCATGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000173212	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAAAGCTGGAGGTAATGACGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000173212	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCTATCAGAGGGTCATTGGAGGTTTTAGAGCTAGGCCAACATGAG GATCACC

ENSG00000185022	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGTGACGTCACCGCATGACTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000185022	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGTTGTAAGGCGAGCTTCCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000185022	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCCCGGAATTACTCACTCACGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000116353	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGCCTGGCTAGACTGCGTGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000116353	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCCCTTCAAAGGGACCAAGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000116353	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCTAGCCAGGCGCCACCACTTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000136146	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGACACCGGCGCCATCTGTTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000136146	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCCGGCCCAACAGATGGCGCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000136146	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGGAGGCGCACTGCGGACACGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000100139	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCGAGCGGGCAGCGCGGAGTGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000100139	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGCAGCGCGGAGTAGGCGGGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000100139	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGCCGGGCCGGGCCACGCGGGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000146410	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGACCCGTAACCAGCCTCATTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000146410	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGATTGGTCATTGCATGATGTCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000146410	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGGGCAGGGCAGGTCGCGTTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000087053	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTTCACCAGACCCCCTCACCCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000087053	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCACCAGACCCCCTCACCCTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000087053	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGGCGGCTCCCAGGGTGAGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000144959	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGTGACTGCCACACTTTGCAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000144959	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCGCGGACATGCCCCCTCTAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000144959	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTTGCAAGGACACCGTAGAGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000196498	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGCGGAGGCTGGCTCGAGAGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000196498	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGCGGAGGCTGGCTCGAGAGGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000196498	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCACAGGCAGGCGGACTTGTGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000110717	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCACTGTAGGACGCTGCCATTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000110717	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTTGGCCTCCCCAAACCAACAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000110717	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGCAAGGCTGCTGAAAGAGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000164190	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCAGGAATGACTCCCTCCGCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000164190	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCTGCAGCTGCACTTCCGGTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000164190	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAAGTGGAGTGGGGAAGAGGGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000135838	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGATGTTGTTCAGGTGACACCCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000135838	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGCGCAGAACTACAACCCCCGTTTTAGAGCTAGGCCAACATGAG GATCACC

ENSG00000135838	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGCCACGCCAGGGAGCCTGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000143257	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGACACGGGGAGGGACTCCAGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000143257	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGCATGACAAAAGTGCTGGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000143257	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTCACAAGGTTCCCACCCACGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000203757	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGTGATTCACTTCAGTGCAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000203757	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTTAATTAACCCCATGAGAGAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000203757	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAATAACAGGAAAGGGAACTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000197702	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGGGAGCGCAGCTCCTTCCAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000197702	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAGGGCAGAGGGCGGCGCGAGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000197702	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCGCCGCCCTCTGCCCTGGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000126249	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCCCTTACTAGGATCCGATGGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000126249	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTTAAAGCCACGCCTCCGCATGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000126249	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAGAGCCACAGAGGCGCCGCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000129292	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTTGACGGACCGCTAGCGGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000129292	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCCTACGGCGGCCCGCCAATGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000129292	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGGAAAGAACGCGAGCTCTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000021300	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTACAAGTCCCAGTACGCCTCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000021300	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGTGAAGAGATGCTGACCCGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000021300	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGATGCTGACCCGAGGCGTACTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000100142	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCAGAAATACTGCGCGATCTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000100142	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCTTTACAGCCGCATCCGCACGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000100142	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAAGCAGTGGTCACCGCCTCCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000172531	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCCGCGCCTCACGTCCAGCGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000172531	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGAAGGAGAGCCAGGCCGGAGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000172531	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCCCCGCCTCCAGGCCTCCCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000172179	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTATGGGGGTAATCTCAATGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000172179	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAATGACGGAAATAGATGACCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000172179	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCATATTCAGGAAGACATACGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000185238	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGAGAGCCCGGCGTGTCTGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000185238	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCCACCTCACGTGACCCGCAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000185238	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGCGCCGTGCGGGTCACGTGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000126067	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTGACCTGCACAGCCTGCTCGTTTTAGAGCTAGGCCAACATGAGG ATCACC

ENSG00000126067	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGTCGCGAGAGGTTGCAAAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000126067	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCTTGTCTCTGGGATCGTACGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000149177	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAAGCGATGAATATTCAGAGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000149177	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTTCATCGCTTCCTCCCGCCCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000149177	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGCCCCCGCCCCCTCCGAGCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000166965	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCACCCCAGCACGTTCGAGGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000166965	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCCCAAGCCGCCCTTCAACCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000166965	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCCAGCTGAGCCTGGTTGAAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000139547	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAAGATAGTATCTTCCTACTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000139547	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAAGATACTATCTTCCTACTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000139547	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCAGGACTTGAGGTGTGCAAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000115255	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGGGCAGGGCGGACAAAGGGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000115255	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGTGGCGGAGGGCTATGCGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000115255	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCAGGGCGGACAAAGGAGGAGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000165731	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGCGCAGCCAGAGCAAGCACGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000165731	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGGGGCTCCAGTGCTTGCTCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000165731	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCCCGCACCCCACCCGCCTCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000133874	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAAGCCGGGAGGATCTTCGCAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000133874	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGCCTGCAAATAATAGCCGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000133874	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGCAGGAGAAAGGCTCCGATGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000170633	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTGACCGGCCTCCCGAAAGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000170633	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGATAGGAAGCCGACTTTCGGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000170633	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCGGTAAAAGCGCAGAGCGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000175634	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGATTCCCGAGAGGCCTTGCGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000175634	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAAGGGATCTTACTCCCCCTCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000175634	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCTAGGACTACAATTCCCGAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000176783	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAACCTTTCGAAACAGAGGGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000176783	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGAAGGCGAGGCGGTGAAGAGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000176783	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGAAGGCGAGGCGGTGAAGGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000169976	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTAGCGTAACTCTCGCTCATGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000169976	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCGCTCATAGGGCTCAGAGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000169976	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGATAGGGCTCAGAGGCGGAAGGTTTTAGAGCTAGGCCAACATGAG GATCACC

ENSG00000185437	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCAGAGCTGGGGCTGTAACAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000185437	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAACAGGGTAACCGGCTGAGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000185437	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAACAGCGCGGAGCAGGTAAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000167114	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGCCCGCCCTGCTCGCGCAAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000167114	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGACAGGCGGGCTTCACCCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000167114	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGAGTCCGCGAGGAGACAGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000040487	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGAGGAGCGCGGAGGCAGTTGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000040487	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGGAGCGCGGAGGCAGTTGGGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000040487	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGGAGCGCGGAGGCAGTTGGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000130821	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGGGCGGAGTGTGACGAGGAGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000130821	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTAGTGACATCACCCCGGAGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000130821	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTGTGACGAGGAGGGCGGGAGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000183963	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCCCACTCTCGTGCCCCATTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000183963	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCCCAATGGGGCACGAGAGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000183963	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAATGGGGCACGAGAGTGGGGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000087087	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGATCTCGCAAGTCTCGCGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000087087	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCGGGGAGAGCAGGGCGTGAGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000087087	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGAGCAGGGCGTGATGGGAAGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000141380	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCGCAGAAGCGAGACATCCCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000141380	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGAAGGAGGCACTCTGGCCAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000141380	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAGGCCTGAAGGAGGCACTCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000136840	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGCCCTCCCCAGGTCGCGCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000136840	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGATTCACGTCCCGCCCCGTCCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000136840	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGACGGGGCGGGACGTGAAGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000117632	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAAGTGTAGTCCTGTCCCGGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000117632	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCGGGACCCCGAAGCACCTCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000117632	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCCGCATTGGCCGAGAGCGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000243244	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGAGAGAAGGGAGGCTCAGCGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000243244	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGAGAAGGGAGGCTCAGCAGGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000243244	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGAGAAGGGAGGCTCAGCAGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000103266	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTTAAGGGTGGGCGTTCGCGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000103266	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTGCGCCCAACCAGCCTGGAGTTTTAGAGCTAGGCCAACATGAG GATCACC

ENSG00000103266	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGAGCGCGGGGACAGGGAACGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000162227	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCCCCAATACAGCCAGTCGGGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000162227	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGAGGGCCGGAAACTTTCCAGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000162227	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCAAGAGCCAGGAAGCCTCCGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000158710	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGGGACAGTAGACCAGAGCAGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000158710	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGACAGTAGACCAGAGCAAGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000158710	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGAGGACTGCTTGAGACAGAGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000164362	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGCGGAGCTGGAAGGTGAAGGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000164362	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGGCGGAGCTGGAAGGTGAGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000164362	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGGCGGAGCTGGAAGGTGAAGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000198133	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTGCTCCTAGGCGCCCGTGGGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000198133	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCCGGGCTGCGGGAGGCAGAGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000198133	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCTCTGCCTCCCGCAGCCCGGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000137747	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAACTGGGATGGCCTCGATGAGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000137747	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGGGGCAGAGGAACAGGTCGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000137747	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGACTGGGATGGCCTCGATGAGGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000105576	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTAGTTCCAGGGCTTATCGGAGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000105576	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTAGTTCCAGGGCTTATCGGGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000105576	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCCCACCCTCCGATAAGCCCGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000170777	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAAGTAGACACGCTAGCTCGGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000170777	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTAGAATCAGTGCTCAGCTGGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000170777	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTGGGGAGCCGGGCTGGTCAGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000164548	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCTCCCGCGAGGCTTTGTGTGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000164548	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAGCCGACACAAAGCCTCGCGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000164548	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTCGGCTGGGCCTGGAGCGAGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000127191	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGCCAACCAGCCAGCCCTCGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000127191	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCTCTCGCTACAGCTTCCTGAGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000127191	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGCGGGGCGTATCTGGCCAAGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000118271	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTCTAGAGAGATTAGAGCATGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000118271	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGGGATAAGCAGCCTAGCTCGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000118271	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGTCAATAATCAGAATCAGCGTTTTAGAGCTAGGCCAACATGAGGATCACC
ENSG00000160803	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGGCTGCGGCTGGTCCGGAGGTTTTAGAGCTAGGCCAACATGAGGATCACC

ENSG00000160803	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGCCGCAGCCACAGCGCCCCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000160803	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAGCCGCAGCCACAGCGCCCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000130717	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGGGAGTTGTAGTCCACCGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000130717	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGTCCCTGCCAGCCAGCTTTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000130717	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGGGGCCGCGCATGCGTGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000176125	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGGCTGGAGGGCAAGAAACGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000176125	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGGCAGGCTAGCTGGCACGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000176125	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAAGAAACAGGCGAGCTCCGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000140553	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCTAAGGAGGGAGCGCGCCCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000140553	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAAGCGCCCCGCCCCTGCCCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000140553	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGAGGGGCAGGGCTAAGGAGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000137288	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCACCGCTACTCTCCGTGTAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000137288	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTGGGCATGCGCGACTTTGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000137288	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGCACCGCTACTCTCCGTGTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000183066	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGAGGAGGAGGAGCGAGACCGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000183066	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGCGAGACCGGGTCACGTGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000183066	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGGGGAGCGGGGCGGAGTCAGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000133316	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGACTAGGGCTGGCTTGATCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000133316	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGGCAGACAGTTCACACTTCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000133316	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGGTGACGCACACGCTGCGCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000111186	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGGTGGGCAGGAGCGAGCCGGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000111186	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCCTGGGAGGAGCCCTCGGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000111186	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGCCGCGACACCCTCCCGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000188033	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGGCAGCAGGTCTAGTAGCTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000188033	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCAGATCTCGCTTCCGCCGGTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000188033	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGCACCAGTTCCGTCCCACGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000105732	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCGTTGCTGAGGGTAGCTGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000105732	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGCCCGAGCGGGCAATGATGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000105732	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCGAACGGCAGGGCCCGAGCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000117010	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCGACGGATGGCCTACACCCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000117010	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTCCGACGGATGGCCTACACCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000117010	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCCGGGTGTAGGCCATCCGTGTTTTAGAGCTAGGCCAACATGAG GATCACC

ENSG00000173875	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCACCAGTTCCGTCCCACCGGGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000173875	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCGTAGCCGTACCCTCCTTTCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000173875	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTAGGCAGATCTCGCTTCCGCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000284034	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCAGACTGTGAAGCTGAGTGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000284034	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGAGTGGGGAAACAAGGTGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000284034	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGTGAAGCTGAGTGGGGAAACAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000176075	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAAGTGGTTCCTCCCATACTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000176075	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAAGGATGGATTGAAGAGACAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000176075	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGCCCTGCAGATGCCCCAGTATGTTTTAGAGCTAGGCCAACATGAGG ATCACC
ENSG00000141736	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCGCTAGGAGGGACGCACCCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000141736	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCCCAGGCCTGCGCGAAGAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000141736	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGTCCGGGATAAATTCCCTGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000136997	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCTGCTTTGGCAGCAAATTGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000136997	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCCTAGCCCAGCTCTGGAACGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000136997	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGCCGCGAGCAGCACAGCTCGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000121879	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCGGAAGCGAAATTGAGGCGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000121879	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAAGCAGATGCGCAAAGAAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000121879	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGGCGGAAAAGCAAGACGCGTTTTAGAGCTAGGCCAACATGA GGATCACC
ENSG00000110092	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCAAACGCCGGGAGCAGCGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000110092	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGCCCAAAAGCCATCCCTGGTTTTAGAGCTAGGCCAACATGAG GATCACC
ENSG00000110092	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTCAAAGCCCGGCAGAGAATGTTTTAGAGCTAGGCCAACATGAG GATCACC
non-targeting_00000	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTGTCGTGATGCGTAGACGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
non-targeting_00001	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTCATCAAGGAGCATTCCGTGTTTTAGAGCTAGGCCAACATGAG GATCACC
non-targeting_00002	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGACCCTGACATGTATGTAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
non-targeting_00003	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCACTTAGCAGTTTGCAATGGTTTTAGAGCTAGGCCAACATGAGG ATCACC
non-targeting_00004	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGATAAATCGAAGTGTGACAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
non-targeting_00005	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGTGTAAGGGGCGTAATAAAGTTTTAGAGCTAGGCCAACATGAG GATCACC
non-targeting_00006	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTGGGAATTCTCGCATTCGTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
non-targeting_00007	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGCTCATAGATACGTCTTAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
non-targeting_00008	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGGATACAATCTTGGTCCGGTTTTAGAGCTAGGCCAACATGAG GATCACC
non-targeting_00009	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCGGACCTTGACCGCGGGAAGTTTTAGAGCTAGGCCAACATGAG GATCACC
non-targeting_00010	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGTATATAAAACGAGATTGGTTTTAGAGCTAGGCCAACATGAG GATCACC

non-targeting_00011	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAACACCAAATATGTCGGTGGTTTTAGAGCTAGGCCAACATGAG GATCACC
non-targeting_00012	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAATCCGACCCAGACTGAGAGTTTTAGAGCTAGGCCAACATGAG GATCACC
non-targeting_00013	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAATCGCCCGTAGAGCCTCCGTTTTAGAGCTAGGCCAACATGAG GATCACC
non-targeting_00014	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTGGTAGTGAGAAGTACTAGGTTTTAGAGCTAGGCCAACATGAG GATCACC
non-targeting_00015	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGATCAAGCCTAGGGGGCAGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
non-targeting_00016	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGACCATTGACCAAGCTGAGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
non-targeting_00017	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGAGCGGCACGGATGAGATGTTTTAGAGCTAGGCCAACATGA GGATCACC
non-targeting_00018	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGTCGTCCTGTCGTTGACTAGTTTTAGAGCTAGGCCAACATGAGG ATCACC
non-targeting_00019	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGCCCAGGGTACAAGTTGGGTTTTAGAGCTAGGCCAACATGAG GATCACC
non-targeting_00020	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGTCATGACGACTCTAAATCGTTTTAGAGCTAGGCCAACATGAGG ATCACC
non-targeting_00021	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGAGATATACTAGTTGGAAGTTTTAGAGCTAGGCCAACATGAG GATCACC
non-targeting_00022	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGACCATGTAGATATATTTACGTTTTAGAGCTAGGCCAACATGAGG ATCACC
non-targeting_00023	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTCGCTACCTTGACGTTCGGGTTTTAGAGCTAGGCCAACATGAGG ATCACC
non-targeting_00024	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTCCCTGACTACCTGTGCGTGTTTTAGAGCTAGGCCAACATGAGG ATCACC
non-targeting_00025	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTGGCTAGTCTATAATAAATGTTTTAGAGCTAGGCCAACATGAGG ATCACC
non-targeting_00026	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGCATCAGCGGACGTAGCACGTTTTAGAGCTAGGCCAACATGAG GATCACC
non-targeting_00027	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGAACCGTTGCTTGTGTCGGTTTTAGAGCTAGGCCAACATGAG GATCACC
non-targeting_00028	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGTGTTTTGACAGGAATCACGTTTTAGAGCTAGGCCAACATGAG GATCACC
non-targeting_00029	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGCCCCGCCGCTTCGGATATGGTTTTAGAGCTAGGCCAACATGAG GATCACC
non-targeting_00030	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAGATCAGGGGTGGTCCGTGGTTTTAGAGCTAGGCCAACATGAG GATCACC

MCF-7 proliferation screen significantly enriched genes with MAGeCK (D7 vs D0)

id	num	pos|score	pos|p-value	pos|fdr	pos|rank	pos|goodsgrna	pos|lfc
ENSG00000229980	3	0.000	4.49E-05	0.009	1	2	0.402
ENSG00000138433	3	0.003	0.0145	0.746	2	1	0.009
ENSG00000238098	3	0.004	0.0145	0.746	3	2	0.276
ENSG00000117399	3	0.005	0.0145	0.746	4	3	0.248
ENSG00000133874	3	0.006	0.0270	1	5	3	0.244
ENSG00000176125	3	0.009	0.0470	1	6	2	0.326

MCF-7 proliferation screen significantly enriched genes with MAGeCK (D14 vs D0)

id	num	pos|score	pos|p-value	pos|fdr	pos|rank	pos|goodsgrna	pos|lfc
ENSG00000116353	3	0.0027	4.98E-06	0.000171	1	1	0.1593
ENSG00000240364	3	0.0081	4.98E-06	0.000171	2	1	0.4705
ENSG00000166965	3	0.0135	4.98E-06	0.000171	3	1	0.4228

ENSG00000232627	3	0.0188	4.98E-06	0.000171	4	1	0.6219
ENSG00000252590	3	0.0241	4.98E-06	0.000171	5	1	0.2914
ENSG00000257818	3	0.0294	4.98E-06	0.000171	6	1	0.0621

MCF-7 migration screen significantly enriched genes with MAGeCK

id	num	pos|score	pos|p-value	pos|fdr	pos|rank	pos|goodsgrna	pos|lfc
ENSG00000100142	3	0.0027003	0.0028458	0.350964	1	1	2.0852

PATU-8988T migration significantly enriched genes with MAGeCK

id	num	pos|score	pos|p-value	pos|fdr	pos|rank	pos|goodsgrna	pos|lfc
ENSG00000258077	3	0.0025175	0.0067832	0.698665	1	3	0.52719
ENSG00000232818	3	0.0027003	0.0067832	0.698665	2	1	-0.024502

Sample information of the GMKF neuroblastomas included in the analysis

Case ID	Risk	Age	Gender	MYCN	Ploidy	INRG Stage	Survival	Stime	SV Count
PT_YMDFCE4V	High	1559	Female	Not amplified	Hyperdiploid	Stage 2b	Unknown	NA	415
PT_3WF5J3PZ	High	458	Male	Amplified	Hypodiploid	Unknown	No death	3920	362
PT_69AGBVQ5	High	1112	Male	Not amplified	Hypodiploid	Stage 2b	Unknown	NA	279
PT_ASJZTDRF	High	1614	Female	Not amplified	Hypodiploid	Stage 2b	Unknown	NA	120
PT_K5709E5B	High	583	Male	Amplified	Hypodiploid	Stage 2b	Unknown	NA	75
PT_D508JGWE	High	1278	Female	Amplified	Hypodiploid	Stage 2b	No death	3598	65
PT_1YAJEAMJ	High	418	Male	Amplified	Hyperdiploid	Stage 2a	Unknown	NA	63
PT_2RZN4HR2	High	929	Female	Not amplified	Hyperdiploid	Stage 2b	Unknown	NA	55
PT_GQBEY0JD	High	71	Male	Amplified	Hyperdiploid	Stage 3	Unknown	NA	53
PT_Q50YZ2T5	High	837	Male	Amplified	Hyperdiploid	Stage 2b	No death	2787	53
PT_2Y7Q85BM	High	1120	Female	Not amplified	Hypodiploid	Stage 2b	Unknown	NA	51
PT_AGYJR7PZ	High	1438	Male	Not amplified	Hyperdiploid	Stage 2b	No death	2483	42
PT_ASH4P45D	High	270	Male	Amplified	Hyperdiploid	Stage 2b	No death	3554	37
PT_DP679T4D	High	979	Male	Amplified	Hyperdiploid	Stage 2b	Unknown	NA	37
PT_4FTZAAC4	High	1821	Male	Not amplified	Hyperdiploid	Stage 2b	Unknown	NA	37
PT_RSPKGFXS	High	1061	Male	Not amplified	Hypodiploid	Stage 2b	Unknown	NA	37
PT_GV2XJJTP	High	305	Male	Amplified	Hypodiploid	Stage 2b	Unknown	NA	35
PT_B39849MF	High	1154	Female	Amplified	Hyperdiploid	Stage 2b	Unknown	NA	34
PT_2DX56CE0	High	1448	Female	Not amplified	Hypodiploid	Stage 2b	Unknown	NA	32

PT_69EVASRX	High	1895	Female	Not amplified	Hyperdiploid	Stage 2a	Unknown	NA	32
PT_26E4RFYV	High	414	Female	Amplified	Hyperdiploid	Unknown	Unknown	NA	31
PT_6R3RJ6MY	High	189	Male	Amplified	Hypodiploid	Stage 3	Unknown	NA	28
PT_TTHE7B08	High	1010	Female	Not amplified	Hyperdiploid	Stage 2b	Unknown	NA	28
PT_5E269C8Z	High	944	Female	Amplified	Hypodiploid	Stage 2a	Unknown	NA	27
PT_B9X3H54Y	High	1616	Male	Not amplified	Hypodiploid	Stage 2b	No death	3493	27
PT_P9QJMTF8	High	1685	Female	Not amplified	Hypodiploid	Stage 2b	No death	2570	27
PT_3YW2V4JK	High	567	Female	Amplified	Hyperdiploid	Stage 2a	No death	3609	23
PT_RG7MMHFF	High	474	Male	Not amplified	Hyperdiploid	Unknown	Unknown	NA	22
PT_4WVGKQRX	High	1525	Male	Not amplified	Hypodiploid	Stage 2b	Died	1772	21
PT_A4VM4H5N	High	964	Female	Not amplified	Hyperdiploid	Stage 2b	Unknown	NA	20
PT_64B8K70Y	High	958	Male	Amplified	Hypodiploid	Stage 2b	Unknown	NA	19
PT_F0QD1YWQ	High	871	Male	Not amplified	Hyperdiploid	Stage 2b	Unknown	NA	19
PT_02SNWVRF	High	1126	Female	Amplified	Hypodiploid	Stage 1	No death	3753	16
PT_1EQHANKW	High	133	Male	Amplified	Hyperdiploid	Stage 2b	Unknown	NA	16
PT_3VNMNFT6	High	1724	Male	Not amplified	Hypodiploid	Stage 2b	No death	727	15
PT_J3X9NQ5F	High	1179	Female	Not amplified	Hyperdiploid	Stage 2b	Unknown	NA	15
PT_V1HR5C5P	High	1261	Male	Not amplified	Hyperdiploid	Stage 2b	Unknown	NA	13
PT_SDPQ63J1	High	1148	Female	Not amplified	Hypodiploid	Stage 2b	Died	397	12
PT_QH23VVKW	High	2079	Male	Amplified	Hyperdiploid	Stage 2b	No death	33	7
PT_1NDSW1JX	High	1280	Female	Not amplified	Hypodiploid	Stage 2b	Unknown	NA	7
PT_XPGEBQKA	High	1041	Male	Not amplified	Hyperdiploid	Stage 2b	Unknown	NA	7
PT_EXZSSRGH	High	1081	Male	Not amplified	Hypodiploid	Stage 2b	Unknown	NA	6
PT_YHWENHB0	High	549	Female	Not amplified	Hypodiploid	Stage 2b	No death	2911	6
PT_53M7K3JE	High	583	Female	Not amplified	Hyperdiploid	Stage 2b	No death	3706	5
PT_QF2A2F08	High	554	Female	Not amplified	Hypodiploid	Stage 2a	Unknown	NA	4
PT_GSWXPFPQ	High	1713	Male	Not amplified	Hypodiploid	Stage 2b	Unknown	NA	2
PT_B0YZ0H85	High	597	Female	Not amplified	Unknown	Stage 2b	Unknown	NA	1
PT_KWRFGRER	High	939	Female	Not amplified	Hypodiploid	Stage 1	No death	4181	0
PT_4W8PD8TR	Intermediate	211	Female	Not amplified	Hypodiploid	Stage 2b	No death	3121	58
PT_581CW7RN	Intermediate	378	Male	Not amplified	Hypodiploid	Stage 2a	No death	2445	44
PT_JYRSHSWJ	Intermediate	3	Male	Not amplified	Hypodiploid	Stage 3	No death	3259	30
PT_A77B7F2F	Intermediate	355	Female	Not amplified	Hypodiploid	Stage 2a	No death	3277	22
PT_HZ4VWQP5	Intermediate	503	Male	Not amplified	Hypodiploid	Stage 1	No death	3055	21
PT_CV0FE3Z3	Intermediate	181	Female	Not amplified	Hyperdiploid	Stage 2b	No death	3029	19
PT_2QB9MP9J	Intermediate	194	Male	Not amplified	Hyperdiploid	Stage 2b	No death	2784	17
PT_9A9Q2YB3	Intermediate	248	Male	Not amplified	Hypodiploid	Stage 1	No death	3164	17
PT_ATQMV6B3	Intermediate	249	Female	Not amplified	Hypodiploid	Stage 2b	No death	3175	17
PT_7E6A5N3P	Intermediate	470	Male	Not amplified	Hypodiploid	Stage 2a	No death	3360	13
PT_KBVX8B37	Intermediate	30	Male	Not amplified	Hypodiploid	Stage 3	No death	1345	13

PT_2G290D0G	Intermediate	67	Male	Not amplified	Hypodiploid	Unknown	No death	1167	8
PT_9GRB7EF0	Intermediate	164	Female	Not amplified	Hyperdiploid	Stage 3	No death	3859	8
PT_H2Q0BW73	Intermediate	9	Male	Not amplified	Hypodiploid	Stage 2b	No death	3193	8
PT_8BYCCC0V	Intermediate	307	Male	Not amplified	Hypodiploid	Stage 2b	No death	3638	7
PT_8DFBAQVQ	Intermediate	200	Female	Not amplified	Hyperdiploid	Stage 2b	Died	882	7
PT_9X3MV3GW	Intermediate	8	Male	Not amplified	Hyperdiploid	Stage 3	No death	3636	7
PT_KXWQXAR4	Intermediate	127	Male	Not amplified	Hypodiploid	Stage 2b	No death	3918	7
PT_22BQQFYM	Intermediate	201	Female	Not amplified	Hypodiploid	Stage 2b	No death	4072	4
PT_B9CP3H35	Intermediate	19	Female	Not amplified	Hyperdiploid	Stage 3	No death	2885	4
PT_DCBZYQAX	Intermediate	242	Female	Not amplified	Hypodiploid	Unknown	No death	2831	4
PT_21PJ8R0Z	Intermediate	85	Female	Not amplified	Hypodiploid	Stage 2b	No death	3454	3
PT_2HCWZNTR	Intermediate	1348	Female	Not amplified	Hyperdiploid	Stage 1	No death	254	3
PT_CCC65GCE	Intermediate	427	Male	Not amplified	Hypodiploid	Unknown	No death	2451	3
PT_XNBJNRXJ	Intermediate	401	Female	Not amplified	Hypodiploid	Stage 2a	No death	3150	3
PT_2FB9C15K	Intermediate	814	Male	Not amplified	Hyperdiploid	Stage 1	No death	3439	2
PT_C3YC0C9Q	Intermediate	518	Female	Not amplified	Hypodiploid	Stage 2a	No death	3780	2
PT_HZQ6TWR9	Intermediate	243	Male	Not amplified	Hypodiploid	Stage 1	No death	3893	2
PT_ZS5D8MVF	Intermediate	494	Female	Not amplified	Hypodiploid	Stage 2a	No death	3721	2
PT_70BK6DFW	Intermediate	885	Female	Not amplified	Hyperdiploid	Stage 2a	No death	3237	1
PT_86NG4W76	Intermediate	41	Male	Not amplified	Hyperdiploid	Stage 1	No death	3761	1
PT_9KB3ESTZ	Intermediate	212	Male	Not amplified	Hypodiploid	Stage 2b	No death	3386	1
PT_9RJY3GWC	Intermediate	228	Male	Not amplified	Hypodiploid	Stage 2b	No death	3779	1
PT_HC1QFR28	Intermediate	193	Male	Not amplified	Hypodiploid	Stage 2a	No death	3362	1
PT_KRHMQFFP	Intermediate	201	Male	Not amplified	Hypodiploid	Stage 2a	No death	3375	1
PT_M6QAJFS8	Intermediate	329	Female	Not amplified	Hypodiploid	Stage 2b	No death	2529	1
PT_PFRE83H3	Intermediate	176	Male	Not amplified	Hypodiploid	Stage 1	No death	679	1
PT_XDPN4357	Intermediate	141	Male	Not amplified	Hypodiploid	Stage 2a	No death	3647	1
PT_ZK8Z4WAK	Intermediate	283	Male	Not amplified	Hypodiploid	Stage 2a	No death	1841	1
PT_2JZNQGTR	Intermediate	140	Female	Not amplified	Hypodiploid	Stage 2b	No death	3714	0
PT_DS5XN67S	Intermediate	47	Male	Not amplified	Hypodiploid	Stage 2a	No death	3773	0
PT_FW0K9SXK	Intermediate	31	Male	Not amplified	Hypodiploid	Unknown	No death	2662	0
PT_W6AVZF18	Intermediate	45	Female	Not amplified	Hypodiploid	Stage 2a	No death	3499	0
PT_X8N7GE8X	Intermediate	68	Male	Not amplified	Hypodiploid	Stage 2a	No death	3344	0
PT_P2M0Q2KS	Low	676	Male	Not amplified	Hypodiploid	Stage 1	No death	3575	109
PT_PDYCQB6P	Low	17	Male	Not amplified	Hypodiploid	Stage 1	No death	3032	44
PT_K579G3KQ	Low	142	Male	Unknown	Unknown	Stage 1	No death	2766	39
PT_XPTE7785	Low	385	Female	Not amplified	Hypodiploid	Stage 1	No death	3220	38
PT_RJPEMEQV	Low	691	Female	Not amplified	Hyperdiploid	Stage 1	No death	3231	35
PT_81RSHW1D	Low	421	Female	Not amplified	Hypodiploid	Stage 1	No death	2533	33
PT_K0BJPWY9	Low	299	Female	Not amplified	Hypodiploid	Stage 1	No death	1380	26

PT_YJ8KZG27	Low	148	Male	Not amplified	Hypodiploid	Stage 1	No death	2807	21
PT_R94DDN50	Low	82	Female	Not amplified	Hypodiploid	Stage 3	No death	2903	13
PT_56ZM694R	Low	203	Male	Not amplified	Hypodiploid	Stage 2a	No death	1293	12
PT_ECTDZ6QS	Low	1618	Male	Not amplified	Hypodiploid	Stage 1	No death	876	12
PT_D5BYDHZ9	Low	142	Male	Not amplified	Hypodiploid	Stage 3	No death	2693	11
PT_G3Q35987	Low	777	Male	Not amplified	Hypodiploid	Stage 1	No death	2012	11
PT_M4ETZ912	Low	139	Female	Not amplified	Hypodiploid	Stage 3	No death	2842	11
PT_ZW22K0YF	Low	699	Male	Not amplified	Hypodiploid	Stage 1	No death	3656	10
PT_6M0TPG4X	Low	394	Male	Not amplified	Hypodiploid	Stage 2a	No death	783	9
PT_8HFHWZH9	Low	968	Female	Not amplified	Hyperdiploid	Stage 1	No death	3868	8
PT_5CPS8GNT	Low	21	Female	Not amplified	Hyperdiploid	Stage 1	No death	3309	7
PT_HB9JT4G5	Low	205	Female	Not amplified	Hypodiploid	Stage 2a	No death	2178	7
PT_P7V330C5	Low	338	Female	Not amplified	Hypodiploid	Stage 1	No death	2230	7
PT_E3R0MRXN	Low	231	Female	Amplified	Hypodiploid	Stage 1	No death	3633	6
PT_5MA1YQ49	Low	504	Female	Not amplified	Hypodiploid	Stage 1	No death	2958	6
PT_6WE8JADD	Low	1210	Male	Not amplified	Hypodiploid	Stage 1	No death	2384	6
PT_8RQQWAQR	Low	168	Female	Not amplified	Hypodiploid	Stage 2a	No death	2440	6
PT_FZ3XEWEK	Low	744	Female	Not amplified	Hypodiploid	Stage 1	No death	2499	6
PT_KH0H9EZS	Low	359	Male	Not amplified	Hypodiploid	Stage 1	No death	3720	6
PT_S4EJKTME	Low	708	Female	Not amplified	Hypodiploid	Stage 1	No death	2872	6
PT_WG51EA8V	Low	381	Male	Not amplified	Hypodiploid	Stage 2a	No death	1977	6
PT_6TM0T48Z	Low	140	Female	Not amplified	Hypodiploid	Stage 1	No death	2516	5
PT_EKP4F49T	Low	1326	Male	Not amplified	Hypodiploid	Unknown	No death	288	5
PT_NZ3F3J67	Low	1012	Female	Not amplified	Hypodiploid	Stage 1	No death	3571	5
PT_HYJB8Y4N	Low	58	Male	Not amplified	Hypodiploid	Unknown	No death	2578	4
PT_M8RHAK5K	Low	501	Female	Not amplified	Hypodiploid	Unknown	No death	2668	4
PT_VA8GM98Z	Low	409	Female	Not amplified	Hypodiploid	Stage 2a	No death	3102	4
PT_0XAWD5CE	Low	1567	Male	Not amplified	Hypodiploid	Stage 1	No death	1628	3
PT_1396H6SD	Low	396	Male	Not amplified	Hypodiploid	Stage 1	No death	2261	3
PT_5FCYBT0S	Low	165	Female	Not amplified	Hypodiploid	Stage 3	No death	2879	3
PT_D4SZQV48	Low	8	Female	Not amplified	Hypodiploid	Stage 2a	No death	178	3
PT_H3GBG09Q	Low	859	Female	Not amplified	Hyperdiploid	Stage 1	No death	2528	3
PT_HQ23GQ23	Low	13	Male	Not amplified	Hypodiploid	Stage 1	No death	3878	3
PT_PV869ZYE	Low	1871	Male	Not amplified	Hypodiploid	Stage 1	No death	3475	3
PT_QW5Q0G84	Low	860	Female	Not amplified	Hypodiploid	Stage 1	No death	3487	3
PT_WWRAC6EH	Low	505	Female	Not amplified	Hypodiploid	Stage 2a	No death	3300	3
PT_XNDPC9TT	Low	835	Female	Not amplified	Hypodiploid	Stage 1	No death	3570	3
PT_YYGH8EMR	Low	196	Female	Not amplified	Hypodiploid	Stage 1	No death	2269	3
PT_11XN6CG5	Low	310	Female	Not amplified	Hyperdiploid	Stage 2a	No death	2929	2
PT_1X6CJ589	Low	284	Male	Not amplified	Hypodiploid	Stage 1	No death	2363	2

PT_66Y5KGME	Low	21	Male	Not amplified	Hypodiploid	Unknown	No death	2821	2
PT_89D6BFGP	Low	2	Male	Not amplified	Hypodiploid	Stage 1	No death	2374	2
PT_AQS8CCAB	Low	519	Male	Not amplified	Hypodiploid	Stage 1	No death	2321	2
PT_BZCXTAH9	Low	559	Female	Not amplified	Hypodiploid	Unknown	No death	2056	2
PT_D9XF79J4	Low	439	Male	Not amplified	Hypodiploid	Stage 1	No death	3774	2
PT_HA7TBZ1V	Low	191	Female	Not amplified	Hypodiploid	Stage 1	No death	3015	2
PT_JBQT2QPG	Low	1029	Male	Not amplified	Hypodiploid	Stage 1	No death	3393	2
PT_NYMKWAZT	Low	1373	Female	Not amplified	Hyperdiploid	Stage 1	No death	1882	2
PT_V3BXBVVV	Low	66	Male	Not amplified	Hypodiploid	Stage 1	No death	652	2
PT_WWQGABFP	Low	468	Female	Not amplified	Hypodiploid	Stage 1	No death	141	2
PT_YPK89ADE	Low	416	Female	Not amplified	Hypodiploid	Stage 1	No death	784	2
PT_1MWZEHCT	Low	1464	Female	Not amplified	Hyperdiploid	Stage 1	No death	538	1
PT_2YBKT6RW	Low	185	Female	Not amplified	Hypodiploid	Stage 2a	No death	4028	1
PT_49FZV0HC	Low	291	Male	Not amplified	Hypodiploid	Stage 1	No death	3935	1
PT_4Y3P2N1P	Low	111	Male	Not amplified	Hypodiploid	Stage 1	No death	2325	1
PT_5W51TAZS	Low	11	Male	Not amplified	Hypodiploid	Stage 1	No death	3712	1
PT_6HZH56MX	Low	44	Male	Not amplified	Hypodiploid	Stage 1	No death	2622	1
PT_7XV9SBKQ	Low	1160	Female	Not amplified	Hyperdiploid	Stage 1	No death	2909	1
PT_APMAKP20	Low	1076	Female	Not amplified	Hypodiploid	Stage 1	No death	3632	1
PT_BZZY1BM4	Low	424	Male	Not amplified	Hypodiploid	Stage 1	No death	3966	1
PT_C6429DZZ	Low	131	Female	Not amplified	Hypodiploid	Stage 3	No death	3701	1
PT_E6CZS2KF	Low	301	Male	Not amplified	Hypodiploid	Unknown	No death	1892	1
PT_E7PFZT6E	Low	422	Male	Not amplified	Hypodiploid	Stage 2a	No death	3335	1
PT_ESKA5P5B	Low	1618	Male	Not amplified	Hyperdiploid	Stage 2a	No death	3730	1
PT_JD8FVX6G	Low	578	Male	Not amplified	Hypodiploid	Unknown	No death	3311	1
PT_MG3HP8D9	Low	532	Female	Not amplified	Hypodiploid	Stage 1	No death	3503	1
PT_QCMS0C3W	Low	1197	Female	Not amplified	Hypodiploid	Stage 1	No death	1231	1
PT_QZFYXPJK	Low	79	Female	Not amplified	Hypodiploid	Stage 3	No death	3781	1
PT_SBS3N6ZT	Low	1570	Female	Not amplified	Hyperdiploid	Stage 2a	No death	2678	1
PT_WH6RANZQ	Low	1772	Female	Not amplified	Hypodiploid	Stage 1	No death	92	1
PT_Z4S0193A	Low	37	Female	Not amplified	Hypodiploid	Stage 3	No death	2585	1
PT_ZT2NW6WA	Low	92	Female	Not amplified	Hypodiploid	Stage 1	No death	2557	1
PT_10KTTTPD	Low	5	Male	Not amplified	Hypodiploid	Stage 1	No death	2611	0
PT_1X9YQF9W	Low	81	Male	Not amplified	Hypodiploid	Stage 1	No death	3601	0
PT_4A1B95TK	Low	1295	Female	Not amplified	Hyperdiploid	Stage 1	No death	3135	0
PT_58J0PB4V	Low	291	Female	Not amplified	Hypodiploid	Unknown	No death	2557	0
PT_7BAFX5PZ	Low	52	Male	Not amplified	Hypodiploid	Stage 3	No death	3832	0
PT_92RR9C8D	Low	174	Female	Not amplified	Hypodiploid	Stage 2a	No death	2664	0
PT_9DD8F0VD	Low	2011	Female	Not amplified	Hypodiploid	Unknown	No death	579	0
PT_9K8VF0Z0	Low	277	Male	Not amplified	Hypodiploid	Stage 3	No death	1877	0

PT_F2AFSP66	Low	55	Female	Not amplified	Hypodiploid	Stage 1	No death	2541	0
PT_GGJ9E0VV	Low	1	Female	Not amplified	Hypodiploid	Stage 1	No death	2506	0
PT_K3QMVST1	Low	177	Male	Not amplified	Hypodiploid	Stage 3	No death	3351	0
PT_MK375DCF	Low	410	Male	Not amplified	Hypodiploid	Stage 1	No death	1956	0
PT_R07QYFJ0	Low	429	Female	Not amplified	Hypodiploid	Stage 1	No death	1637	0
PT_RS3TBZV5	Low	116	Female	Not amplified	Hypodiploid	Stage 1	Died	10	0
PT_RVTVP55V	Low	55	Male	Not amplified	Hypodiploid	Stage 3	No death	2280	0
PT_SV8ETF29	Low	381	Male	Not amplified	Hypodiploid	Unknown	No death	240	0
PT_VVVS471N	Low	992	Female	Not amplified	Hypodiploid	Stage 2a	No death	3536	0
PT_XKZYFJZV	Low	470	Male	Not amplified	Hypodiploid	Stage 1	No death	2421	0

Reference

1. Torre, L.A., et al., *Global Cancer Incidence and Mortality Rates and Trends--An Update.* Cancer Epidemiol Biomarkers Prev, 2016. **25**(1): p. 16-27.
2. Hanahan, D., *Hallmarks of Cancer: New Dimensions.* Cancer Discov, 2022. **12**(1): p. 31-46.
3. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* Cell, 2000. **100**(1): p. 57-70.
4. Duesberg, P.H. and P.K. Vogt, *Differences between the ribonucleic acids of transforming and nontransforming avian tumor viruses.* Proc Natl Acad Sci U S A, 1970. **67**(4): p. 1673-80.
5. Stehelin, D., et al., *DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA.* Nature, 1976. **260**(5547): p. 170-3.
6. Bister, K. and P.H. Duesberg, *Structure and specific sequences of avian erythroblastosis virus RNA: evidence for multiple classes of transforming genes among avian tumor viruses.* Proc Natl Acad Sci U S A, 1979. **76**(10): p. 5023-7.
7. Duesberg, P.H., K. Bister, and P.K. Vogt, *The RNA of avian acute leukemia virus MC29.* Proc Natl Acad Sci U S A, 1977. **74**(10): p. 4320-4.
8. Slamon, D.J., et al., *Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer.* Science, 1989. **244**(4905): p. 707-12.
9. Pon, J.R. and M.A. Marra, *Driver and passenger mutations in cancer.* Annu Rev Pathol, 2015. **10**: p. 25-50.
10. Shlien, A. and D. Malkin, *Copy number variations and cancer.* Genome Med, 2009. **1**(6): p. 62.

11. Ostroverkhova, D., T.M. Przytycka, and A.R. Panchenko, *Cancer driver mutations: predictions and reality.* Trends in Molecular Medicine, 2023. **29**(7): p. 554-566.

12. Huang, L., et al., *KRAS mutation: from undruggable to druggable in cancer.* Signal Transduction and Targeted Therapy, 2021. **6**(1).

13. Skoulidis, F., et al., *Sotorasib for Lung Cancers with KRAS p.G12C Mutation.* New England Journal of Medicine, 2021. **384**(25): p. 2371-2381.

14. Kamio, T., et al., *Immunohistochemical expression of epidermal growth factor receptors in human adrenocortical carcinoma.* Hum Pathol, 1990. **21**(3): p. 277-82.

15. Brand, T.M., et al., *The nuclear epidermal growth factor receptor signaling network and its role in cancer.* Discov Med, 2011. **12**(66): p. 419-32.

16. Abourehab, M.A.S., et al., *Globally Approved EGFR Inhibitors: Insights into Their Syntheses, Target Kinases, Biological Activities, Receptor Interactions, and Metabolism.* Molecules, 2021. **26**(21).

17. Akher, F.B., A. Farrokhzadeh, and M.E.S. Soliman, *Covalent vs. Non-Covalent Inhibition: Tackling Drug Resistance in EGFR - A Thorough Dynamic Perspective.* Chem Biodivers, 2019. **16**(3): p. e1800518.

18. Hanahan, D. and A. Weinberg, Robert, *Hallmarks of Cancer: The Next Generation.* Cell, 2011. **144**(5): p. 646-674.

19. Negrini, S., V.G. Gorgoulis, and T.D. Halazonetis, *Genomic instability — an evolving hallmark of cancer.* Nature Reviews Molecular Cell Biology, 2010. **11**(3): p. 220-228.

20. Fishel, R., et al., *The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer.* Cell, 1993. **75**(5): p. 1027-38.

21. Al-Tassan, N., et al., *Inherited variants of MYH associated with somatic G:C-->T:A mutations in colorectal tumors.* Nat Genet, 2002. **30**(2): p. 227-32.

22. Kennedy, R.D. and A.D. D'Andrea, *DNA repair pathways in clinical practice: lessons from pediatric cancer susceptibility syndromes.* J Clin Oncol, 2006. **24**(23): p. 3799-808.

23. Ripperger, T., et al., *Breast cancer susceptibility: current knowledge and implications for genetic counselling.* Eur J Hum Genet, 2009. **17**(6): p. 722-31.

24. Rajagopalan, H. and C. Lengauer, *Aneuploidy and cancer.* Nature, 2004. **432**(7015): p. 338-41.

25. Lee, J.K., et al., *Mechanisms and Consequences of Cancer Genome Instability: Lessons from Genome Sequencing Studies.* Annu Rev Pathol, 2016. **11**: p. 283-312.

26. Yang, L., *A Practical Guide for Structural Variation Detection in the Human Genome.* Curr Protoc Hum Genet, 2020. **107**(1): p. e103.

27. Jeggo, P.A., L.H. Pearl, and A.M. Carr, *DNA repair, genome stability and cancer: a historical perspective.* Nat Rev Cancer, 2016. **16**(1): p. 35-42.

28. Boveri, T., *Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris.* J Cell Sci, 2008. **121 Suppl 1**: p. 1-84.

29. Boveri, T., *Zur frage der entstehung maligner tumoren*. 1914, Jena,: G. Fischer. 2 p.l., 64 p.

30. Burdette, W.J., *The significance of mutation in relation to the origin of tumors: a review.* Cancer Res, 1955. **15**(4): p. 201-26.

31. Falini, B. and D.Y. Mason, *Proteins encoded by genes involved in chromosomal alterations in lymphoma and leukemia: clinical value of their detection by immunocytochemistry.* Blood, 2002. **99**(2): p. 409-26.

32. Tomescu, O. and F.G. Barr, *Chromosomal translocations in sarcomas: prospects for therapy.* Trends Mol Med, 2001. **7**(12): p. 554-9.

33. Nowell, P.C. and C.M. Croce, *Chromosomal approaches to the molecular basis of neoplasia.* Symp Fundam Cancer Res, 1986. **39**: p. 17-29.

34. Rowley, J.D., *Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining.* Nature, 1973. **243**(5405): p. 290-3.

35. Kantarjian, H., et al., *Hematologic and cytogenetic responses to imatinib mesylate in chronic myelogenous leukemia.* N Engl J Med, 2002. **346**(9): p. 645-52.

36. Soda, M., et al., *Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer.* Nature, 2007. **448**(7153): p. 561-6.

37. Kwak, E.L., et al., *Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer.* N Engl J Med, 2010. **363**(18): p. 1693-703.

38. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

39. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.

40. Sjoblom, T., et al., *The consensus coding sequences of human breast and colorectal cancers.* Science, 2006. **314**(5797): p. 268-74.

41. Wood, L.D., et al., *The genomic landscapes of human breast and colorectal cancers.* Science, 2007. **318**(5853): p. 1108-13.

42. Li, Y., et al., *Patterns of somatic structural variation in human cancer genomes.* Nature, 2020. **578**(7793): p. 112-121.

43. Trost, B., L.O. Loureiro, and S.W. Scherer, *Discovery of genomic variation across a generation.* Hum Mol Genet, 2021. **30**(R2): p. R174-R186.

44. Weinhold, N., et al., *Genome-wide analysis of noncoding regulatory mutations in cancer.* Nat Genet, 2014. **46**(11): p. 1160-5.

45. Stephens, P.J., et al., *Complex landscapes of somatic rearrangement in human breast cancer genomes.* Nature, 2009. **462**(7276): p. 1005-10.

46. Yang, L., et al., *Diverse mechanisms of somatic structural variations in human cancer genomes.* Cell, 2013. **153**(4): p. 919-29.

47. Huang, M. and W.A. Weiss, *Neuroblastoma and MYCN.* Cold Spring Harb Perspect Med, 2013. **3**(10): p. a014415.

48. Slamon, D.J., et al., *Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene.* Science, 1987. **235**(4785): p. 177-82.

49. Cairns, P., et al., *Rates of p16 (MTS1) mutations in primary tumors with 9p loss.* Science, 1994. **265**(5170): p. 415-7.

50. Li, J., et al., *PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer.* Science, 1997. **275**(5308): p. 1943-7.

51. Salesse, S. and C.M. Verfaillie, *BCR/ABL: from molecular mechanisms of leukemia induction to treatment of chronic myelogenous leukemia.* Oncogene, 2002. **21**(56): p. 8547-59.

52. Kohno, T., et al., *RET fusion gene: translation to personalized lung cancer therapy.* Cancer Sci, 2013. **104**(11): p. 1396-400.

53. Kohno, T., et al., *Beyond ALK-RET, ROS1 and other oncogene fusions in lung cancer.* Transl Lung Cancer Res, 2015. **4**(2): p. 156-64.

54. Carroll, S.B., *Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution.* Cell, 2008. **134**(1): p. 25-36.

55. Ong, C.T. and V.G. Corces, *Enhancer function: new insights into the regulation of tissue-specific gene expression.* Nat Rev Genet, 2011. **12**(4): p. 283-93.

56. Wittkopp, P.J. and G. Kalay, *Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence.* Nat Rev Genet, 2011. **13**(1): p. 59-69.

57. Banerji, J., S. Rusconi, and W. Schaffner, *Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences.* Cell, 1981. **27**(2 Pt 1): p. 299-308.

58. Malik, S. and R.G. Roeder, *The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation.* Nat Rev Genet, 2010. **11**(11): p. 761-72.

59. Clapier, C.R. and B.R. Cairns, *The biology of chromatin remodeling complexes.* Annu Rev Biochem, 2009. **78**: p. 273-304.

60. Buenrostro, J.D., et al., *ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide.* Curr Protoc Mol Biol, 2015. **109**: p. 21 29 1-21 29 9.

61. Robertson, G., et al., *Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.* Nat Methods, 2007. **4**(8): p. 651-7.

62. Skene, P.J. and S. Henikoff, *An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites.* Elife, 2017. **6**.

63. Ruthenburg, A.J., et al., *Multivalent engagement of chromatin modifications by linked binding modules.* Nat Rev Mol Cell Biol, 2007. **8**(12): p. 983-94.

64. Bedford, D.C., et al., *Target gene context influences the transcriptional requirement for the KAT3 family of CBP and p300 histone acetyltransferases.* Epigenetics, 2010. **5**(1): p. 9-15.

65. Consortium, E.P., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.

66. Heintzman, N.D., et al., *Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.* Nat Genet, 2007. **39**(3): p. 311-8.

67. Heintzman, N.D., et al., *Histone modifications at human enhancers reflect global cell-type-specific gene expression.* Nature, 2009. **459**(7243): p. 108-12.

68. Pombo, A. and N. Dillon, *Three-dimensional genome architecture: players and mechanisms.* Nat Rev Mol Cell Biol, 2015. **16**(4): p. 245-57.

69. Schoenfelder, S. and P. Fraser, *Long-range enhancer-promoter contacts in gene expression control.* Nat Rev Genet, 2019. **20**(8): p. 437-455.

70. Cullen, K.E., M.P. Kladde, and M.A. Seyfred, *Interaction between transcription regulatory regions of prolactin chromatin.* Science, 1993. **261**(5118): p. 203-6.

71. Dekker, J., et al., *Capturing chromosome conformation.* Science, 2002. **295**(5558): p. 1306-11.

72. van de Werken, H.J., et al., *Robust 4C-seq data analysis to screen for regulatory DNA interactions.* Nat Methods, 2012. **9**(10): p. 969-72.

73. Kempfer, R. and A. Pombo, *Methods for mapping 3D chromosome architecture.* Nat Rev Genet, 2020. **21**(4): p. 207-226.

74. Hughes, J.R., et al., *Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment.* Nat Genet, 2014. **46**(2): p. 205-12.

75. Nagano, T., et al., *Single-cell Hi-C reveals cell-to-cell variability in chromosome structure.* Nature, 2013. **502**(7469): p. 59-64.

76. Slobodyanyuk, E., C. Cattoglio, and T.S. Hsieh, *Mapping Mammalian 3D Genomes by Micro-C.* Methods Mol Biol, 2022. **2532**: p. 51-71.

77. Fullwood, M.J., et al., *Chromatin interaction analysis using paired-end tag sequencing.* Curr Protoc Mol Biol, 2010. **Chapter 21**: p. Unit 21 15 1-25.

78. Rowley, M.J. and V.G. Corces, *Organizational principles of 3D genome architecture.* Nat Rev Genet, 2018. **19**(12): p. 789-800.

79. Deng, W., et al., *Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor.* Cell, 2012. **149**(6): p. 1233-44.

80. Morgan, S.L., et al., *Manipulation of nuclear architecture through CRISPR-mediated chromosomal looping.* Nat Commun, 2017. **8**: p. 15993.

81. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions.* Nature, 2012. **485**(7398): p. 376-80.

82. Rao, S.S., et al., *A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.* Cell, 2014. **159**(7): p. 1665-80.

83. Phillips-Cremins, J.E., et al., *Architectural protein subclasses shape 3D organization of genomes during lineage commitment.* Cell, 2013. **153**(6): p. 1281-95.

84. Fudenberg, G., et al., *Formation of Chromosomal Domains by Loop Extrusion.* Cell Rep, 2016. **15**(9): p. 2038-49.

85. Wan, L.B., et al., *Maternal depletion of CTCF reveals multiple functions during oocyte and preimplantation embryo development.* Development, 2008. **135**(16): p. 2729-38.

86. Guo, Y.A., et al., *Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers.* Nat Commun, 2018. **9**(1): p. 1520.

87. Narendra, V., et al., *CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation.* Science, 2015. **347**(6225): p. 1017-21.

88. Wang, J. and L.M. Boxer, *Regulatory elements in the immunoglobulin heavy chain gene 3'-enhancers induce c-myc deregulation and lymphomagenesis in murine B cells.* J Biol Chem, 2005. **280**(13): p. 12766-73.

89. Northcott, P.A., et al., *Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma.* Nature, 2014. **511**(7510): p. 428-34.

90. Abraham, B.J., et al., *Small genomic insertions form enhancers that misregulate oncogenes.* Nat Commun, 2017. **8**: p. 14385.

91. Helmsauer, K., et al., *Enhancer hijacking determines extrachromosomal circular MYCN amplicon architecture in neuroblastoma.* Nat Commun, 2020. **11**(1): p. 5823.

92. Haller, F., et al., *Enhancer hijacking activates oncogenic transcription factor NR4A3 in acinic cell carcinomas of the salivary glands.* Nat Commun, 2019. **10**(1): p. 368.

93. Weischenfeldt, J., et al., *Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking.* Nat Genet, 2017. **49**(1): p. 65-74.

94. 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, 2020. **69**(6): p. 1039-1052.

95. He, B., et al., *Diverse noncoding mutations contribute to deregulation of cis-regulatory landscape in pediatric cancers.* Sci Adv, 2020. **6**(30): p. eaba3064.

96. Liu, Y., et al., *Discovery of regulatory noncoding variants in individual cancer genomes by using cis-X.* Nat Genet, 2020. **52**(8): p. 811-818.

97. Wang, X., et al., *Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes.* Nature Methods, 2021. **18**(6): p. 661-668.

98. Pugh, T.J., et al., *The genetic landscape of high-risk neuroblastoma.* Nat Genet, 2013. **45**(3): p. 279-84.

99. Irwin, M.S., et al., *Revised Neuroblastoma Risk Classification System: A Report From the Children's Oncology Group.* J Clin Oncol, 2021. **39**(29): p. 3229-3241.

100. Meany, H.J., *Non-High-Risk Neuroblastoma: Classification and Achievements in Therapy.* Children (Basel), 2019. **6**(1).

101. Smith, V. and J. Foster, *High-Risk Neuroblastoma Treatment Review.* Children (Basel), 2018. **5**(9).

102. Mosse, Y.P., et al., *Identification of ALK as a major familial neuroblastoma predisposition gene.* Nature, 2008. **455**(7215): p. 930-5.

103. Maris, J.M., et al., *Chromosome 6p22 locus associated with clinically aggressive neuroblastoma.* N Engl J Med, 2008. **358**(24): p. 2585-93.

104. Capasso, M., et al., *Common variations in BARD1 influence susceptibility to high-risk neuroblastoma.* Nat Genet, 2009. **41**(6): p. 718-23.

105. Janoueix-Lerosey, I., et al., *Somatic and germline activating mutations of the ALK kinase receptor in neuroblastoma.* Nature, 2008. **455**(7215): p. 967-70.

106. Kohl, N.E., et al., *Transposition and amplification of oncogene-related sequences in human neuroblastomas.* Cell, 1983. **35**(2 Pt 1): p. 359-67.

107. Henriksen, J.R., et al., *Conditional expression of retrovirally delivered anti-MYCN shRNA as an in vitro model system to study neuronal differentiation in MYCN-amplified neuroblastoma.* BMC Dev Biol, 2011. **11**: p. 1.

108. Dzieran, J., et al., *MYCN-amplified neuroblastoma maintains an aggressive and undifferentiated phenotype by deregulation of estrogen and NGF signaling.* Proc Natl Acad Sci U S A, 2018. **115**(6): p. E1229-E1238.

109. Pinto, N., et al., *Predictors of differential response to induction therapy in high-risk neuroblastoma: A report from the Children's Oncology Group (COG).* Eur J Cancer, 2019. **112**: p. 66-79.

110. Zimmerman, M.W., et al., *MYC Drives a Subset of High-Risk Pediatric Neuroblastomas and Is Activated through Mechanisms Including Enhancer Hijacking and Focal Enhancer Amplification.* Cancer Discov, 2018. **8**(3): p. 320-335.

111. Valentijn, L.J., et al., *TERT rearrangements are frequent in neuroblastoma and identify aggressive tumors.* Nat Genet, 2015. **47**(12): p. 1411-4.

112. Krijger, P.H. and W. de Laat, *Regulation of disease-associated gene expression in the 3D genome.* Nat Rev Mol Cell Biol, 2016. **17**(12): p. 771-782.

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

114. Lupianez, D.G., et al., *Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions.* Cell, 2015. **161**(5): p. 1012-1025.

115. Zhang, W., et al., *A global transcriptional network connecting noncoding mutations to changes in tumor gene expression.* Nat Genet, 2018. **50**(4): p. 613-620.

116. Weischenfeldt, J., et al., *Phenotypic impact of genomic structural variation: insights from and for human disease.* Nat Rev Genet, 2013. **14**(2): p. 125-38.

117. Davis, C.F., et al., *The somatic genomic landscape of chromophobe renal cell carcinoma.* Cancer Cell, 2014. **26**(3): p. 319-330.

118. 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, 1985. **41**(3): p. 899-906.

119. Gostissa, M., et al., *Long-range oncogenic activation of Igh-c-myc translocations by the Igh 3' regulatory region.* Nature, 2009. **462**(7274): p. 803-7.

120. Hnisz, D., et al., *Activation of proto-oncogenes by disruption of chromosome neighborhoods.* Science, 2016. **351**(6280): p. 1454-1458.

121. Groschel, S., et al., *A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia.* Cell, 2014. **157**(2): p. 369-381.

122. Northcott, P.A., et al., *The whole-genome landscape of medulloblastoma subtypes.* Nature, 2017. **547**(7663): p. 311-317.

123. Kopp, F. and J.T. Mendell, *Functional Classification and Experimental Dissection of Long Noncoding RNAs.* Cell, 2018. **172**(3): p. 393-407.

124. Liu, S.J., et al., *Long noncoding RNAs in cancer metastasis.* Nat Rev Cancer, 2021. **21**(7): p. 446-460.

125. Wang, X., et al., *Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes.* Nat Methods, 2021. **18**(6): p. 661-668.

126. Consortium, I.T.P.-C.A.o.W.G., *Pan-cancer analysis of whole genomes.* Nature, 2020. **578**(7793): p. 82-93.

127. Consortium, G.T., et al., *Genetic effects on gene expression across human tissues.* Nature, 2017. **550**(7675): p. 204-213.

128. Fudenberg, G., D.R. Kelley, and K.S. Pollard, *Predicting 3D genome folding from DNA sequence with Akita.* Nat Methods, 2020. **17**(11): p. 1111-1117.

129. Krietenstein, N., et al., *Ultrastructural Details of Mammalian Chromosome Architecture.* Mol Cell, 2020. **78**(3): p. 554-565 e7.

130. Zhou, J., *Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale.* Nat Genet, 2022. **54**(5): p. 725-734.

131. Wang, X., Y. Luan, and F. Yue, *EagleC: A deep-learning framework for detecting a full range of structural variations from bulk and single-cell contact maps.* Sci Adv, 2022. **8**(24): p. eabn9215.

132. Grandi, F.C., et al., *Chromatin accessibility profiling by ATAC-seq.* Nat Protoc, 2022. **17**(6): p. 1518-1552.

133. Beroukhim, R., et al., *The landscape of somatic copy-number alteration across human cancers.* Nature, 2010. **463**(7283): p. 899-905.

134. Uhlen, M., et al., *A pathology atlas of the human cancer transcriptome.* Science, 2017. **357**(6352).

135. Zhang, X., et al., *Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers.* Nat Genet, 2016. **48**(2): p. 176-82.

136. Takeda, D.Y., et al., *A Somatically Acquired Enhancer of the Androgen Receptor Is a Noncoding Driver in Advanced Prostate Cancer.* Cell, 2018. **174**(2): p. 422-432 e13.

137. Turner, K.M., et al., *Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity.* Nature, 2017. **543**(7643): p. 122-125.

138. Wu, S., et al., *Circular ecDNA promotes accessible chromatin and high oncogene expression.* Nature, 2019. **575**(7784): p. 699-703.

139. Morton, A.R., et al., *Functional Enhancers Shape Extrachromosomal Oncogene Amplifications.* Cell, 2019. **179**(6): p. 1330-1341 e13.

140. Li, Y., et al., *Exaggerated false positives by popular differential expression methods when analyzing human population samples.* Genome Biol, 2022. **23**(1): p. 79.

141. Yun, J.W., et al., *Dysregulation of cancer genes by recurrent intergenic fusions.* Genome Biol, 2020. **21**(1): p. 166.

142. Richter, J., et al., *Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing.* Nat Genet, 2012. **44**(12): p. 1316-20.

143. Neuveut, C., Y. Wei, and M.A. Buendia, *Mechanisms of HBV-related hepatocarcinogenesis.* J Hepatol, 2010. **52**(4): p. 594-604.

144. Zapatka, M., et al., *The landscape of viral associations in human cancers.* Nat Genet, 2020. **52**(3): p. 320-330.

145. Sondka, Z., et al., *The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers.* Nat Rev Cancer, 2018. **18**(11): p. 696-705.

146. Wang, T., et al., *OncoVar: an integrated database and analysis platform for oncogenic driver variants in cancers.* Nucleic Acids Res, 2021. **49**(D1): p. D1289-D1301.

147. Franke, M., et al., *Formation of new chromatin domains determines pathogenicity of genomic duplications.* Nature, 2016. **538**(7624): p. 265-269.

148. 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, 2020. **107**(5): p. 802-814.

149. Melo, U.S., et al., *Complete lung agenesis caused by complex genomic rearrangements with neo-TAD formation at the SHH locus.* Hum Genet, 2021. **140**(10): p. 1459-1469.

150. Li, Y., et al., *Constitutional and somatic rearrangement of chromosome 21 in acute lymphoblastic leukaemia.* Nature, 2014. **508**(7494): p. 98-102.

151. Maciejowski, J., et al., *Chromothripsis and Kataegis Induced by Telomere Crisis.* Cell, 2015. **163**(7): p. 1641-54.

152. Dixon, J.R., et al., *Integrative detection and analysis of structural variation in cancer genomes.* Nat Genet, 2018. **50**(10): p. 1388-1398.

153. Shangguan, W.J., et al., *TOB1-AS1 suppresses non-small cell lung cancer cell migration and invasion through a ceRNA network.* Exp Ther Med, 2019. **18**(6): p. 4249-4258.

154. 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, 2018. **8**(8): p. 1483-1498.

155. Elsasser, H.P., et al., *Establishment and characterisation of two cell lines with different grade of differentiation derived from one primary human pancreatic adenocarcinoma.* Virchows Arch B Cell Pathol Incl Mol Pathol, 1992. **61**(5): p. 295-306.

156. Wang, C.Y., et al., *Molecular cloning and characterization of a novel gene family of four ancient conserved domain proteins (ACDP).* Gene, 2003. **306**: p. 37-44.

157. Li, J.H., J. Shou, and Q. Wu, *DNA fragment editing of genomes by CRISPR/Cas9.* Yi Chuan, 2015. **37**(10): p. 992-1002.

158. Wu, Q. and J. Shou, *Toward precise CRISPR DNA fragment editing and predictable 3D genome engineering.* J Mol Cell Biol, 2021. **12**(11): p. 828-856.

159. Claringbould, A. and J.B. Zaugg, *Enhancers in disease: molecular basis and emerging treatment strategies.* Trends Mol Med, 2021. **27**(11): p. 1060-1073.

160. Puissant, A., et al., *Targeting MYCN in neuroblastoma by BET bromodomain inhibition.* Cancer Discov, 2013. **3**(3): p. 308-23.

161. Frangoul, H., et al., *CRISPR-Cas9 Gene Editing for Sickle Cell Disease and beta-Thalassemia.* N Engl J Med, 2021. **384**(3): p. 252-260.

162. Yu, A., et al., *HYENA detects oncogenes activated by distal enhancers in cancer.* Nucleic Acids Res, 2024.

163. Anastasiadou, E., L.S. Jacob, and F.J. Slack, *Non-coding RNA networks in cancer.* Nat Rev Cancer, 2018. **18**(1): p. 5-18.

164. Bester, A.C., et al., *An Integrated Genome-wide CRISPRa Approach to Functionalize lncRNAs in Drug Resistance.* Cell, 2018. **173**(3): p. 649-664 e20.

165. Briggs, J.A., et al., *Mechanisms of Long Non-coding RNAs in Mammalian Nervous System Development, Plasticity, Disease, and Evolution.* Neuron, 2015. **88**(5): p. 861-877.

166. Vancura, A., et al., *Cancer LncRNA Census 2 (CLC2): an enhanced resource reveals clinical features of cancer lncRNAs.* NAR Cancer, 2021. **3**(2): p. zcab013.

167. Leucci, E., et al., *Melanoma addiction to the long non-coding RNA SAMMSON.* Nature, 2016. **531**(7595): p. 518-22.

168. Liu, S.J., et al., *CRISPRi-based radiation modifier screen identifies long non-coding RNA therapeutic targets in glioma.* Genome Biol, 2020. **21**(1): p. 83.

169. Gutschner, T., et al., *The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells.* Cancer Res, 2013. **73**(3): p. 1180-9.

170. Wang, H., et al., *LINC00680 Promotes the Progression of Non-Small Cell Lung Cancer and Functions as a Sponge of miR-410-3p to Enhance HMGB1 Expression.* Onco Targets Ther, 2020. **13**: p. 8183-8196.

171. Sun, C.C., et al., *Long Intergenic Noncoding RNA 00511 Acts as an Oncogene in Non-small-cell Lung Cancer by Binding to EZH2 and Suppressing p57.* Mol Ther Nucleic Acids, 2016. **5**(11): p. e385.

172. Wright, A.V., J.K. Nunez, and J.A. Doudna, *Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering.* Cell, 2016. **164**(1-2): p. 29-44.

173. Garneau, J.E., et al., *The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA.* Nature, 2010. **468**(7320): p. 67-71.

174. Rouet, P., F. Smih, and M. Jasin, *Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease.* Mol Cell Biol, 1994. **14**(12): p. 8096-106.

175. Cong, L., et al., *Multiplex genome engineering using CRISPR/Cas systems.* Science, 2013. **339**(6121): p. 819-23.

176. Chavez, A., et al., *Highly efficient Cas9-mediated transcriptional programming.* Nat Methods, 2015. **12**(4): p. 326-8.

177. Maeder, M.L., et al., *CRISPR RNA-guided activation of endogenous human genes.* Nat Methods, 2013. **10**(10): p. 977-9.

178. Gilbert, L.A., et al., *CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes.* Cell, 2013. **154**(2): p. 442-51.

179. Choi, A., et al., *iCSDB: an integrated database of CRISPR screens.* Nucleic Acids Res, 2021. **49**(D1): p. D956-D961.

180. Wu, D., et al., *Dual genome-wide coding and lncRNA screens in neural induction of induced pluripotent stem cells.* Cell Genom, 2022. **2**(11).

181. Esposito, R., et al., *Multi-hallmark long noncoding RNA maps reveal non-small cell lung cancer vulnerabilities.* Cell Genom, 2022. **2**(9): p. 100171.

182. Joung, J., et al., *Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening.* Nat Protoc, 2017. **12**(4): p. 828-863.

183. Li, W., et al., *MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens.* Genome Biol, 2014. **15**(12): p. 554.

184. Marcon, E., et al., *Human-chromatin-related protein interactions identify a demethylase complex required for chromosome segregation.* Cell Rep, 2014. **8**(1): p. 297-310.

185. Wu, L., et al., *A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer.* Nat Genet, 2018. **50**(7): p. 968-978.

186. Peng, Y., et al., *RCCD1 promotes breast carcinogenesis through regulating hypoxia-associated mitochondrial homeostasis.* Oncogene, 2023. **42**(50): p. 3684-3697.

187. Wu, J., et al., *RCCD1 depletion attenuates TGF-beta-induced EMT and cell migration by stabilizing cytoskeletal microtubules in NSCLC cells.* Cancer Lett, 2017. **400**: p. 18-29.

188. Cheng, Z., et al., *LINC01419 promotes cell proliferation and metastasis in lung adenocarcinoma via sponging miR-519b-3p to up-regulate RCCD1.* Biochem Biophys Res Commun, 2019. **520**(1): p. 107-114.

189. Osborne, C.K., K. Hobbs, and J.M. Trent, *Biological differences among MCF-7 human breast cancer cell lines from different laboratories.* Breast Cancer Res Treat, 1987. **9**(2): p. 111-21.

190. Antonacopoulou, A.G., et al., *POLR2F, ATP6V0A1 and PRNP expression in colorectal cancer: new molecules with prognostic significance?* Anticancer Res, 2008. **28**(2B): p. 1221-7.

191. Zhou, D., et al., *Combining multi-dimensional data to identify a key signature (gene and miRNA) of cisplatin-resistant gastric cancer.* J Cell Biochem, 2018. **119**(8): p. 6997-7008.

192. Naorem, L.D., M. Muthaiyan, and A. Venkatesan, *Integrated network analysis and machine learning approach for the identification of key genes of triple-negative breast cancer.* J Cell Biochem, 2019. **120**(4): p. 6154-6167.

193. Wang, X., et al., *Biochemical recurrence related metabolic novel signature associates with immunity and ADT treatment responses in prostate cancer.* Cancer Med, 2023. **12**(1): p. 862-878.

194. Yang, Y., et al., *Primary glioblastoma transcriptome data analysis for screening survival-related genes.* J Cell Biochem, 2020. **121**(2): p. 1901-1910.

195. Comsa, S., A.M. Cimpean, and M. Raica, *The Story of MCF-7 Breast Cancer Cell Line: 40 years of Experience in Research.* Anticancer Res, 2015. **35**(6): p. 3147-54.

196. Soule, H.D., et al., *Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10.* Cancer Res, 1990. **50**(18): p. 6075-86.

197. Muthuswamy, S.K., et al., *ErbB2, but not ErbB1, reinitiates proliferation and induces luminal repopulation in epithelial acini.* Nat Cell Biol, 2001. **3**(9): p. 785-92.

198. Seton-Rogers, S.E., et al., *Cooperation of the ErbB2 receptor and transforming growth factor beta in induction of migration and invasion in mammary epithelial cells.* Proc Natl Acad Sci U S A, 2004. **101**(5): p. 1257-62.

199. Zhang, H., et al., *Comprehensive analysis of oncogenic effects of PIK3CA mutations in human mammary epithelial cells.* Breast Cancer Res Treat, 2008. **112**(2): p. 217-27.

200. Dixit, A., et al., *Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens.* Cell, 2016. **167**(7): p. 1853-1866 e17.

201. Maris, J.M., et al., *Neuroblastoma.* Lancet, 2007. **369**(9579): p. 2106-20.

202. Rodriguez-Fos, E., et al., *Mutational topography reflects clinical neuroblastoma heterogeneity.* Cell Genom, 2023. **3**(10): p. 100402.

203. Song, Y., et al., *Identification of genomic alterations in oesophageal squamous cell cancer.* Nature, 2014. **509**(7498): p. 91-5.

204. Kandoth, C., et al., *Mutational landscape and significance across 12 major cancer types.* Nature, 2013. **502**(7471): p. 333-339.

205. Patel, R.R., et al., *Tumor mutational burden and driver mutations: Characterizing the genomic landscape of pediatric brain tumors.* Pediatr Blood Cancer, 2020. **67**(7): p. e28338.

206. Houghton, P.J., et al., *The pediatric preclinical testing program: description of models and early testing results.* Pediatr Blood Cancer, 2007. **49**(7): p. 928-40.

207. Izycka-Swieszewska, E., et al., *Prognostic significance of HER2 expression in neuroblastic tumors.* Mod Pathol, 2010. **23**(9): p. 1261-8.

208. Chen, X., et al., *Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications.* Bioinformatics, 2016. **32**(8): p. 1220-2.

209. Morinaga, T., et al., *GDNF-inducible zinc finger protein 1 is a sequence-specific transcriptional repressor that binds to the HOXA10 gene regulatory region.* Nucleic Acids Res, 2005. **33**(13): p. 4191-201.

210. Dong, G., et al., *Integrative analysis of copy number and transcriptional expression profiles in esophageal cancer to identify a novel driver gene for therapy.* Sci Rep, 2017. **7**: p. 42060.

211. Lu, S., et al., *Insights into a Crucial Role of TRIP13 in Human Cancer.* Comput Struct Biotechnol J, 2019. **17**: p. 854-861.

212. Behan, F.M., et al., *Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens.* Nature, 2019. **568**(7753): p. 511-516.

213. Durbin, A.D., et al., *Selective gene dependencies in MYCN-amplified neuroblastoma include the core transcriptional regulatory circuitry.* Nat Genet, 2018. **50**(9): p. 1240-1246.

214. Mikulasova, A., et al., *Epigenomic translocation of H3K4me3 broad domains over oncogenes following hijacking of super-enhancers.* Genome Res, 2022. **32**(7): p. 1343-1354.

215. Rosswog, C., et al., *Genomic ALK alterations in primary and relapsed neuroblastoma.* Br J Cancer, 2023. **128**(8): p. 1559-1571.

216. Bellini, A., et al., *Frequency and Prognostic Impact of ALK Amplifications and Mutations in the European Neuroblastoma Study Group (SIOPEN) High-Risk Neuroblastoma Trial (HR-NBL1).* J Clin Oncol, 2021. **39**(30): p. 3377-3390.

217. Pramod, A.B., et al., *SLC6 transporters: structure, function, regulation, disease association and therapeutics.* Mol Aspects Med, 2013. **34**(2-3): p. 197-219.

218. Chen, J., et al., *SLC34A2 promotes neuroblastoma cell stemness via enhancement of miR-25/Gsk3beta-mediated activation of Wnt/beta-catenin signaling.* FEBS Open Bio, 2019. **9**(3): p. 527-537.

219. Wood, K.A., et al., *The Role of the U5 snRNP in Genetic Disorders and Cancer.* Front Genet, 2021. **12**: p. 636620.

220. Lv, C., et al., *Over-activation of EFTUD2 correlates with tumor propagation and poor survival outcomes in hepatocellular carcinoma.* Clin Transl Oncol, 2022. **24**(1): p. 93-103.

221. Beyer, S., et al., *High RIG-I and EFTUD2 expression predicts poor survival in endometrial cancer.* J Cancer Res Clin Oncol, 2023. **149**(8): p. 4293-4303.

222. Zhu, X., et al., *The feedback loop of EFTUD2/c-MYC impedes chemotherapeutic efficacy by enhancing EFTUD2 transcription and stabilizing c-MYC protein in colorectal cancer.* J Exp Clin Cancer Res, 2024. **43**(1): p. 7.

223. Zhang, Y., et al., *A pediatric brain tumor atlas of genes deregulated by somatic genomic rearrangement.* Nat Commun, 2021. **12**(1): p. 937.

224. Rausch, T., et al., *DELLY: structural variant discovery by integrated paired-end and split-read analysis.* Bioinformatics, 2012. **28**(18): p. i333-i339.

225. Chong, Z., et al., *novoBreak: local assembly for breakpoint detection in cancer genomes.* Nat Methods, 2017. **14**(1): p. 65-67.

226. Cameron, D.L., L. Di Stefano, and A.T. Papenfuss, *Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software.* Nat Commun, 2019. **10**(1): p. 3240.

227. Dong, Z., et al., *Methylation Mediated Downregulation of TOB1-AS1 and TOB1 Correlates with Malignant Progression and Poor Prognosis of Esophageal Squamous Cell Carcinoma.* Dig Dis Sci, 2023. **68**(4): p. 1316-1331.

228. Li, R., et al., *Hepsin Promotes Epithelial-Mesenchymal Transition and Cell Invasion Through the miR-222/PPP2R2A/AKT Axis in Prostate Cancer.* Onco Targets Ther, 2020. **13**: p. 12141-12149.

229. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.