The Pennsylvania State University

The Graduate School

FROM STRUCTURE VARIATION TO CHROMATIN SPATIAL REORGANIZATION, UNCOVERING THE GENETIC AND CIS-REGULATORY ALTERATION IN PRIMARY CANCER GENOMES

A Dissertation in

Biomedical Sciences

by

Jie Xu

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

August 2020

The dissertation of Jie Xu was reviewed and approved by the following:

Feng Yue
Associated professor, Department of Biochemistry and Molecular Genetics, Northwestern University
Adjunct associate professor, Department of Biochemistry and Molecular Biology, Penn State University
Dissertation Advisor
Co-Chair of Committee

James Riley Broach
Distinguished Professor and Chair, Department of Biochemistry and Molecular Biology
Co-Chair of Committee

Zhonghua Gao
Assistant professor, Department of Biochemistry and Molecular Biology

Hong-Gang Wang
Lois High Berstler Professor, Department of Pediatrics,
Professor, Department of Pharmacology Head of the Department or Chair of the Graduate Program

Ross Cameron Hardison
T. Ming Chu Professor, Department of Biochemistry and Molecular

Barbara Ann Miller
Professor, Penn State Health Pediatric Hematology/Oncology

Ralph Keil,
Associate professor, Department of Biochemistry & Molecular Biology
Director of Biomedical Sciences Program

ABSTRACT

Structural variations (SVs), including deletion, insertion, duplication, inversion, translocation, aneuploidy and chromoplexy, can contribute to oncogenesis through a variety of mechanisms. Despite their importance, the identification of SV in cancer genomes remains challenging. G-band karyotyping, fluorescence in-situ hybridization, and microarrays have been historically widely used, but their usages are limited by low resolution, low throughput, and the requirement of prior knowledge for selecting candidates. Whole genome sequencing (WGS) is popular for throughput, but the short sequencing reads can hardly tackle genomic repeats or complex SVs. We therefore developed a method to identify SVs by high-throughput chromosome conformation capture (Hi-C), a technology invented for examining chromatin 3D structure. Hi-C presents long re-ligated DNA reads mostly from the same allele, offering great advantage for identifying and phasing complex SVs. By developing a framework that integrates WGS, Hi-C, and BioNano optical mapping, a technology that utilizes long DNA molecules, we identified and characterized SVs in 36 normal or cancer samples and cell lines, and the integration enabled us to reconstruct local haplotype-resolved map that chains a series of complex SVs. We found that each method has unique strengths in identifying different classes of SVs at different scales, suggesting that integrative approaches are likely the only way currently to comprehensively identify structural variants in the genome.

Next, we ask whether we can apply this integrative framework in a clinical setting to sensitize SV detection for improving disease diagnosis and subtyping. Acute myeloid leukemia (AML) is known for hallmark SVs recurrently disrupting genes such as *MLL, ABL1, NUP214, RUNX1, CEBPA* and *RARA*. However, the SV landscape of AML besides those well-known rearrangements is unclear. Recent TCGA's effort that characterized SVs somehow omitted AML possibly due to the lack of blood control. We hence applied optical mapping in conjunction with WGS in primary leukemia samples, which revealed on average over five thousand SVs per sample. Our computational methods determined that 5-

10% of the variants likely arose as somatic mutations, affecting 37 leukemia associated genes as well as 209 cancer driver genes not previously associated with leukemia, and at least 109 recurrently disrupted genes not previously associated with cancer. Fifteen of the genes not previously associated with AML but mutated in multiple patients' samples significantly affects survival of AML patients.

Our SV profiling in both cancer cell lines and primary leukemia samples identified that 42% of SVs disrupt functions of non-coding sequences, including the deletion of distal regulatory sequences, alteration of DNA replication timing, and the creation of novel chromatin structural domains. Moreover, AML was known for mutations on genes encoding DNA or histone modifiers, transcription factors, kinases, and large genomic variations that can lead epigenetic reshaping. It raises questions whether those mutations drive change of chromatin 3D structure, contributing to oncogenic misregulation in AML, and how intergenic SVs affects the chromatin structure.

Therefore, we performed Hi-C and whole genome sequencing in 21 primary AML and healthy donors' samples, where we identified AML-recurrent or subtype-specific alteration of compartments, TADs, and chromatin loops. To study the impact on gene regulation, we performed RNA-Seq, ATAC-Seq and CUT&TAG for CTCF, H3K27ac, and H3K27me3 in the same patient cohort. We observed that the transcriptional misregulation of many AML-related genes, represented by *MYCN*, *MEIS1*, *GATA3*, *BCL11B, WT1, ERG* and *MYC*, is intimately linked to recurrent gain of loops or compartment switch, coupled with simultaneous acquisition of enhancer or repressor on genes' distal loop anchor. To understand how SVs contribute to altered chromatin structure, we profiled SV in the patient samples using WGS and Hi-C data, and reconstructed cancer 3D genome surrounding the SV breakpoint, by which we identified hundreds of SV-induced neo-loops and enhancer-hijacking events.

We further explored what drives the chromatin misconfiguration, by profiling DNA methylation with whole genome bisulfite sequencing in our AML samples. We identified altered methylation correlated with A/B compartment switch, and hundreds of loss of CTCF insulation due to hypermethylation on the binding sites, serving as a cause of massive gain of loops. Lastly, we were able

to revert the switched compartment and dissociate the gained loops by treating AML cell lines with DNA hypomethylation agent 5-azacytidine, accompanied with significant compromised cell proliferation.

Overall, our work demonstrated that the majority of SVs in cancers or AML are understudied and this may hamper the diagnostic and prognostic efforts, underscoring the importance of accurate and comprehensive SV detection. We indicate that non-coding SVs may be underappreciated mutational drivers in AML, through altering the 3D chromatin structure and triggering enhancer hijacking events. AML subtype-specific chromatin structure changes, alongside the gain of *de novo* enhancers and repressors, contribute to the global oncogenic transcriptional misregulation. Its restoration by chemicals like 5-Azacytidine provides insights into AML treatment through therapeutically reversing the altered chromatin structure.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

SV	structure variation
CNV	copy number variation
IPM	Institute of personalized medicine
TCGA	The cancer genome atlas
FISH	Fluorescence in situ hybridization
WGS	whole genome sequencing
ecDNA	extra-chromosomal DNA
HP1	histone protein 1
PcG	Polycomb group protein
LAD	lamina-associated domain
3C	chromatin conformation capture assay
Hi-C	high throughput chromatin conformation capture assay
H3K27ac	histone 3 lysine 27 acetylation
H3K27me3	histone 3 lysine 27 tri-methylation
TAD	topologically associated domain
mESC	mouse embryonic stem cell
EM	electron microcopy
Cryo-EM	cryogenic electron microscopy
STED	stimulated emission depletion microscopy
PALM	photoactivated localization microscopy
STORM	stochastic optical reconstruction microscopy
FIRE	frequently interacting region
SNV	single nucleotide variation
SNP	single nucleotide polymorphism
IDH	isocitrate dehydrogenase
SDH	succinate dehydrogenase
T-ALL	T-cell acute lymphoblastic leukemia
AML	acute myeloid leukemia
RT	replication timing
PET-seq	paired-end tag sequencing
HMEC	human mammary epithelial cell

NHLF	human lung fibroblast
LOH	loss of heterozygosity
DEL	deletion
TL	translocation
INS	insertion
INV	inversion
ENA	European nucleotide archive
SRA	Sequencing read archive
RPM	read per million mapped reads
TPM	transcripts per million mapped reads
FPKM	fragments per kilobase of transcript per million mapped reads
VAF	variant allele fraction
FLT3-ITD	Fms Related Tyrosine Kinase 3 internal tandem duplication
DHS	DNase hypersensitive site
add	addition of origin-unknown sequence
BM	bone marrow
PB	peripheral blood
MNC	mononuclear cells
PBMC	peripheral blood mononuclear cells
PE	paired-end reads
SR	split reads
DGV	data of genomic variation
WHO	World health organization
TF	transcription factor
WGBS	whole genome bisulfite sequencing
P-P	promoter-promoter loop
P-E	promoter-enhancer loop
P-R	promoter-repressor loop
5-AZA	5-azacytidine
PC1	first principle component

ACKNOWLEDGEMENTS

First, I would like to thank my advisor and mentor Dr. Feng Yue for his guidance and unlimited support through this 6-year journey. Feng is devoted to broadening the horizon of every lab member, and he has been truly considerate over students' life-long career development. As one of his Ph.D. student, I have been offered so many opportunities for scientific communications on international stages and setting up collaboration for frontier researches. Feng's passion for science is infectious, and he has developed a wonderful environment for hatching creative and explorative scientific work. I couldn't imagine to have it done without his far-sighted vision and mentoring.

I would also like to thank Dr. James Broach for his support of my projects and that he provides us with the invaluable platform of Institute of Personalized Medicine (IPM). I want to say thanks to all the personnel who used to work or are working at IPM for their great help in my experiments, especially to Darrin Bann for his patient training of lab techniques, Royden and Lijun for always being helpful in troubleshooting computing problems, and Chris Pool for being enthusiastic and cheerful on all matters.

I would like to express my sincere gratitude to Dr. Zhonghua Gao, Dr. Barbara Miller, Dr. Hong-gang Wang, and Dr. Ross Hardison for serving as my committee members, giving me suggestions, and catching me with inspirational casual chat about my research projects and academic progress from time to time. I also want to thank Dr. Ralph Keil as I benefited a lot from his well-directed and designed curriculums in the program, and also for his constructive suggestions every now and then.

My thanks goes to my lab members, especially Qiushi Jin for managing and arranging experimental logistics, Sriranga Iyyanki for being a great classmate and lab colleague that always helps me and solves tasks cooperatively with me, and Baozhen Zhang for her dedicated time and effort in my project. My friend Wanjian Tang and Shuo Li deserve special thanks for their company and encouragement during this long voyage. I also want to acknowledge Kathy Simon and Kathy Shuey. For every year and each phase during graduate school, they made sure that students like me are always on track in every aspects.

My parents and bother have always been a source of love and support throughout my entire life. I would like to thank my parents for fostering my interest in science, encouraging me to pursue my career and dream, and always offering me unconditioned support.

Last, but not the least, I would like to acknowledge my husband, Fan Song, not only for his love and encouragement throughout my graduate school, but he is also my closest friend, my cheerleader, my best working partner, and a true soulmate. He is always a source of strength and motivation.

The work presented in this dissertation was supported by:

NIH grants R35GM124820, R01HG009906, and U01CA200060 (F.Y.), R24DK106766 (R.C.H. and F.Y.), GM083337 (D.M.G.), GM085354 (D.M.G.), DK107965 (D.M.G.), U54HG004592 (J.D. and J.A.S.), HG003143 and DK107980 (J.D.), U41HG007000 (W.S.N.), and DP5OD023071 (J. D.). This work was also supported by European Research Council (No. 615584 to D.T.O.and C.E.), Cancer Research UK (Nos. 20412 and 22398 to D.T.O. and C.E.), Wellcome Trust (No. 84459 to D.T.O. and C.E.), and Wellcome Trust (No. 106985/Z/15/Z to S.H.). St. Baldrick's foundation (G.L.M.), the Four Diamonds Children's Miracle Network (J.R.B.)

CONTRIBUTION

The following individuals contributed to the work presented in this dissertation:

Chapter 2. Integrative Detection and Analysis of Structural Variation in Cancer Genomes

Jie Xu and Fan Song led the integrative analysis of SV detection and characterization, haplotype reconstruction, and analysis of non-coding variants. Jie performed BioNano optical mapping, whole genome sequencing and PCR validation. Jesse Dixon led the Hi-C analysis and performed FISH validation. Vishnu Dileep contributed to the replication timing analysis. Yan Zhang performed Hi-C. The work is advised and supervised by Feng Yue.

Chapter 3. Whole genome optical mapping reveals previously unrecognizable structural variants in leukemia patients' samples

Jie Xu and Fan Song conducted the investigation. Christopher Pool performed the BioNano experiment for primary sample. Emily Schleicher and George-Lucian Moldovan performed T cell growth, selection and BioNano. Kathryn Sheldon performed WGS library prep. Emma Batchelder performed RNA-seq library prep. Barbara Miller, David Claxon and Hong Zheng provided the samples and resources. James. R. Broach wrote the original manuscript. Jie Xu and Fan Song led the revision and edited the writing. The project was supervised by George-Lucian Moldovan, Feng Yue and James. R. Broach.

Chapter 4. Subtype-Specific and Structure Variation-Induced Chromatin Spatial Reorganization in Acute Myeloid Leukemia

Jie Xu and Fan Song led the investigation. Jie Xu performed Hi-C, ATAC-seq, CUT&TAG, ChIP-seq, RNA-seq, gene knockdown and drug treatment related experiments, and prepared DNA for WGS and WGBS. Fan Song and Jie Xu conducted the data analysis. Baozhen Zhang contributed to the Hi-C and CUT&TAG experiments. Xiaotao Wang developed the algorithm for stripe detection. Jie Xu and Fan Song wrote the manuscript. Hong Zheng provided samples. The work is advised and supervised by Feng Yue.

Chapter 1
Literature review

1. Structure variations in cancer

1.1 Landscape of structure variations in cancer

Structural variations (SVs), including inversions, deletions, duplications, aneuploidy, translocations, and chromoplexy are a hallmark of most cancer genomes[1]. The discovery of recurrent SVs and the affected genes have greatly advanced our knowledge about oncogenesis. Numerous oncogene activation events have been identified as the product of recurrent SVs and have provided successful targets for drug therapies [2-6]. Further, SVs also provide clear diagnostic and prognostic information in the clinic[7], and the presence of certain SVs, notably gene fusion events, have clear treatment implications[8], particularly for hematopoietic malignancies.

SV are not unique to cancer genomes. Indeed, every normal individual genome harbors thousands of germline SVs, some of which also intersect with genes. Experimental deletion of those genes finds most of them dispensable to human health[9]. SVs that presented in cancer genomes mostly as somatic variants, in contrast, could have a huge impact. A recent effort of The Cancer Genome Atlas (TCGA) analyzed SVs in 2,658 whole genomes from across 38 tumor types, covering the most common tumors originated from liver, pancreas, prostate, breast, kidney, central nerve system, lung, bladder, skin, soft tissue, thyroid etc. In each individual cancer, 4-5 driver SVs were detected in average, affecting both coding and non-coding regions[10]. Some tumor types are more heavily affected by SVs like breast adenocarcinomas, whereas colorectal adenocarcinoma bears more burden of point mutations.

A set of cancer drivers can be found widely in all or most cancer types, which mostly involving tumor suppressors and a few oncogenes. Genes like *TP53, CDKN2A, KRAS, PTEN, TERT, CDKN2B, SMAD4, PIK3CA, RB1* and so on are most frequently disrupted by driver mutations. Some other genes are specifically prevalent in certain cancer types. For instance, *SETD2* mutations exclusively occur in medulloblastoma[10]. Noticeably, not all the drivers are somatic mutations. A few germline variants confers high predisposition to cancers, majorly composed of *BRCA2, ATM, EME2* and *POLR2L*[10]. Chromothripsis is a catastrophic event in cancer that involve massive genomic rearrangements at a time or within a few cell cycles. Instead of gradually accumulating mutations, one cancer can rapidly acquire hundreds of mutations. Recent TCGA work found that chromothripsis is actually so prevalent in sarcoma, adenoma, glioblastomas and lung adenocarcinoma that it occurs in more than 40%-50% of cases, and most are associated with *TP53* mutation[11]. Chromothripisis in all liposarcoma involves *MDM2* amplification and for more than 20% of cases there is co-amplification of *TERT*[10]. Similarly, around 50% of acral or mucosal melanomas have chromothripisis accompanied with amplification of *CCND1*[10]. Timing mutations by whether they are homogeneous or subclonal SVs, which respectively implies to earlier or later events, showed that chromathripsis actually took place at the very early phase of oncogenesis[10].

Screening over 2,000 cancer samples TCGA also found that around 64% of SVs reside in non-codoing regions, from untranslated regions immediately upstream or downstream of genes to up to 100Kb away from the most nearby gene. 25% of tumors bear non-coding driver SVs with one-third disrupting *TERT* promoter. Other genes with cis-regulatory altercations include *MDM2, CDK4, ERBB2, CD274, PDCD1LG2,* and *IGF[12]*, some are cancer-type specific. For example, focal deletions at the 5’ UTR of *BRD4* were found exclusively in ovarian or breast cancers[13].

1.2 Methods for detecting structure variations

Despite the importance of SVs in cancer, identifying SVs in cancer genomes remains challenging, hindering our ability to better understand oncogenesis and to develop targeted treatments for cancer. Several methods are currently used to identify structural variations in cancer genomes. G-band karyotyping has been the major method historically to detect gross structural anomalies in the genome, and it is routinely performed today clinically for certain malignancies such as leukemia[14]. However, karyotyping is an inherently low resolution and low throughput method that cannot adequately characterize extensively rearranged genomes. Microarrays are another commonly used method for detecting gains and losses of genetic material[15], but they do not provide precise localization of rearrangements. Further, microarrays are inherently limited in detecting balanced rearrangements, such as inversions or balanced translocations. Finally, targeted approaches such as fluorescence *in situ* hybridization (FISH) and PCR are also used extensively in the clinic. FISH mainly uses fluorescence labeled DNA probes to complementally bind the sites of interest, thereby illuminating the positions of these sites in multiple single cells under a light microscope, and the observer can hence tell whether there are typical translocations bringing targeted sites closer. However, FISH is low resolution and both PCR and FISH require *a priori* knowledge of the rearrangement events and hence are not suitable for *de novo* detection of structural variations.

Recently, high-throughput sequencing-based technologies have emerged as attractive methods for SV identification[16, 17]. Targeted approaches such RNA sequencing provide cost-effective means of identifying gene fusion events[18, 19], while whole genome sequencing (WGS) can identify high-resolution genomic rearrangements as well as gains and losses of genetic material [20-23]. Despite their success, they are limited by their reliance on short sequence reads (usually less than 100 bp), which cannot be effectively mapped to the repetitive regions in the genome. More importantly, these

techniques involve fragmenting the genome into approximately 500 base-pair fragments prior to sequencing, and as a result, much of the structural continuity of the genome is lost and it is challenging to resolve complex SVs that involve multiple events. Given the aforementioned limitations, it is imperative to develop alternative approaches for detecting structural variations that either use longer sequence reads or approaches that retain long-range genomic structural information.

Several technologies that utilize long DNA reads from the genome are available and most widely used in these years, such as Nanopore, Pacific Biosciences (Pacbio), Matepair-seq, and BioNano optical mapping. Although based on completely different biochemical workflows and sequencing platforms, the fundamental idea of Nanopore and PacBio is similar in that they ultimately "read through" DNA molecules nucleotide-by-nucleotide. They have been widely used for *de novo* or improved genome assembly for human or other species[24], detecting SVs on various context[25-27], and identification of new pathogenes such as the virus genome of Covid-19[28, 29]. Average read length, aside from throughput and accuracy, is a major factor that determines the performance of long-read technologies for whether they are competent for resolving large and complicated SVs. PacBio has an average read length of 10Kb and a regular maximum length of 60Kb, whereas for Nanopore it is 16Kb and 134Kb respectively[30, 31]. While those are decent size for detecting simple SVs, it is still difficult to tackle genomic repeats that are hundreds of kilobases long, and complex events which chain a series of various types of SVs in megabase scales. Mate-pair sequencing represents another solution, which uses long DNA reads but sequences them as regular paired-end reads. DNA is hence only sequenced in the 5' and 3'end for totally 100 to 300bp with the large trunk in the middle not sequenced through. The advantage is very clear: Unlike Nanopore or PacBio, the final library of mate pair can go to the regular illumine sequencing platform, which greatly reduces the cost. The shortcoming also roots from it: SV detection is only based on the difference of the actual distance and the expected distance between the 5' and 3' of

DNA molecules. SVs or SNVs without change of size is not detected, and the exact position of the SV in the middle is unknown.

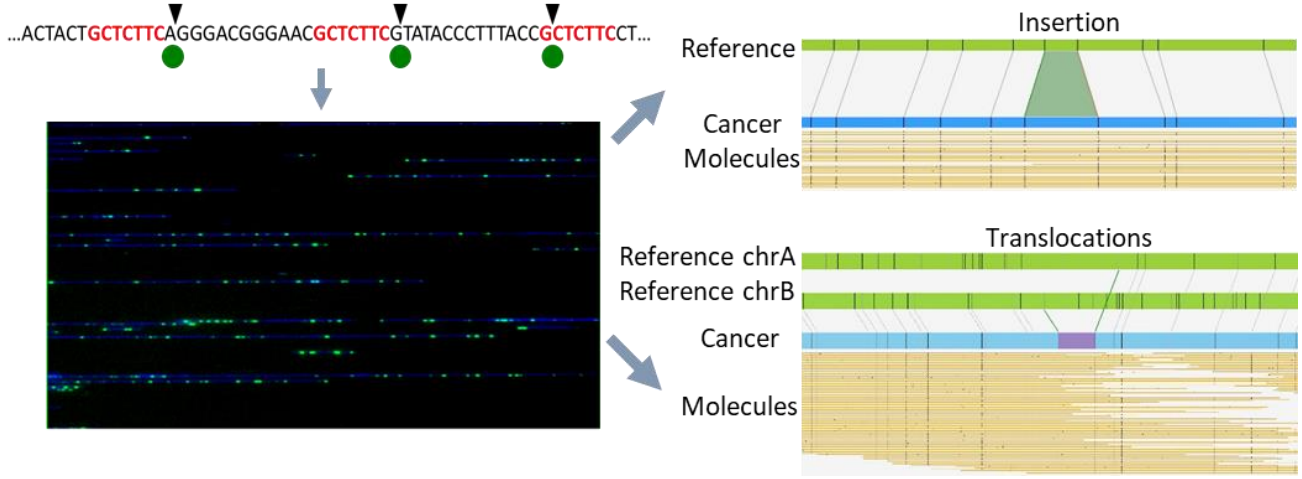


Figure 1- 1: BioNano Optical mapping identify SVs in cancer samples based on restriction enzyme nicking site or protein binding recognition motif.

Restriction enzyme BspQ1 recognizes a seven-basepair sequence and makes a nick on single strand of DNA, where green fluorophore is incorporated. The lower panel in the left shows a real picture of DAPI-stained DNA molecules, each with multiple GFP nicking labels, linearized and migrating through the a BioNano chip. The right panel shows that BioNano computationally convert the picture of DNA labels to map of molecules, which are further aligned into longer contigs. It can automatically identify SVs such as insertions and translocations by comparing the cancer contigs with genomic maps of reference genome.

In contrast to the technique mentioned above, BioNano Optical mapping does not sequence each single nucleotide but it regarded the genome as a large map of restriction enzyme nicking sites (**Figure** 1-1). The newer version of BioNano replaces nicking enzyme with motif-recognition protein, which only binds to specific sequence of DNA without nicking the DNA. As DNA molecules maintain an average length of 200-250Kb, minimally 150Kb and maximally several megabases[32], BioNano turns out a nice fit for scaffolding new genomes, especially for *de novo* assembly for genome of wild plants and crops, which are featured by large numbers of genomic repeats[33-36]. In recent years, it is more and more used for detecting SVs in the cancer by comparing the restriction enzyme labeling pattern in DNA molecule with that in the reference genome[32], including extra-chromosomal DNA (ecDNA)[37]. Since BioNano does not read detailed DNA sequences, it is usually applied in

combination with sequencing-based technologies like WGS, Nanopore or PacBio to reveal a comprehensive picture of the tested genome [38-40].

2. Chromatin spatial organization

2.1 Hierarchical nuclear architecture

If the 46 human chromatins were connected head-to-tail in a linear manner, it can stretch as long as two meters, but those long strands of DNA have to squeeze into cell nuclei that have an average radius of only 3 um[41]. Many proteins facilitate this process by binding to and folding the DNA, wiring the DNA into many loops, forming coiled structure and specialized structure at higher order of dimensions. Although DNA is highly folded and even more densely compacted during metaphase of mitosis, the way DNA is organized leaves high spatial flexibility for it to accommodate DNA replication, transcription and repair. The processes that helix of double-strand DNA wrapped into beaded threads of nucleosomes, folded at different levels to form chromatin domains, and isolated from and connected to each other to occupy specific nuclear territories as we see under the microscope, is what we refer as hierarchical chromatin 3-dimensional architecture (**Figure** 1-2).

Heterochromatin and phase separation

In 1930 two distinct forms of chromatin were observed in the nucleus of eukaryotic cells in interphase: One of them that densely packed was named heterochromatin, and the other one that loosely arranged was named euchromatin[42]. Later, heterochromatin was found to have many nucleosomes stacked closely together. While heterochromatin occurs in many different regions of a chromosome, they are mainly located at centromeres and telomeres. DNA in heterochromatic regions often contains fewer number of genes, and wrapping genes into heterochromatin is a way to downregulate or turn off

gene expression[1]. The formation of heterochromatin is thought to be accomplished by the cooperation of many non-histone proteins, including heterochromatin protein 1 (HP1) and polycomb group protein (PcG).

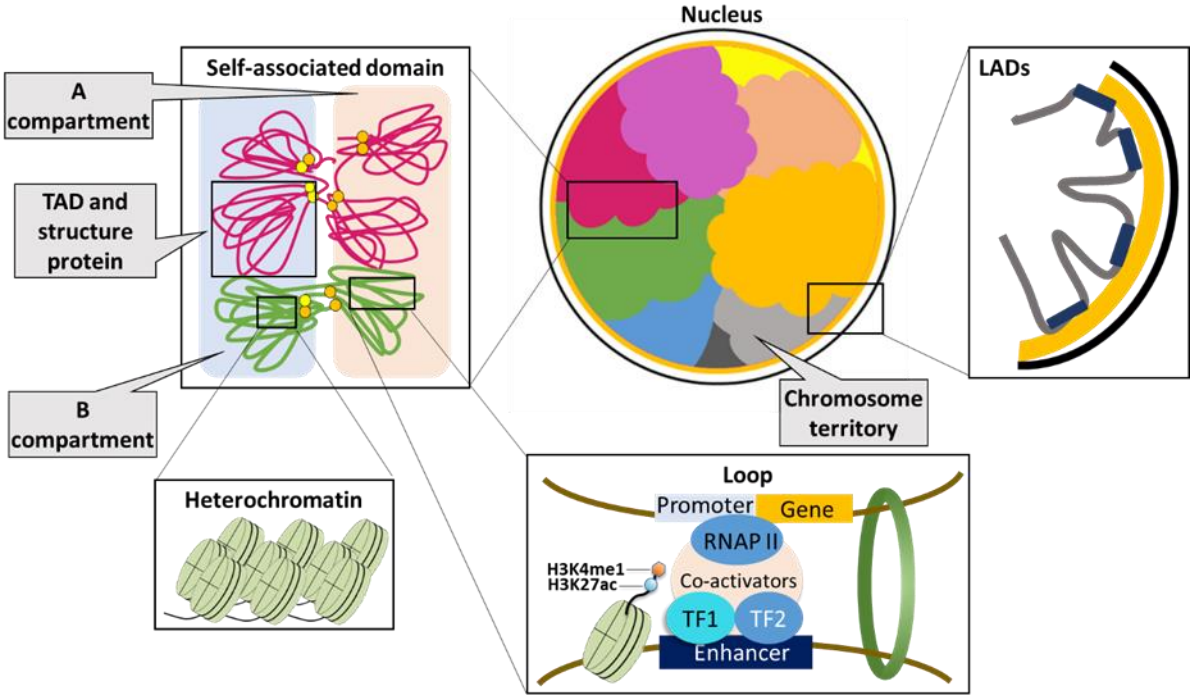


Figure 1- 2: Hierarchical nuclear architecture.

This Figure shows the nuclear and chromatin spatial organization from higher to lower order of scale. Chromatin occupies their distinct territories in the nucleus, with the periphery part often associated with Lamina and forming LADs, which is usually also heterochromatins and B compartment that are densely compacted in the form of nucleosome thread. There are A compartments more enriched in the center of nucleus with abundant chromatin loops formed.

Regarding the mechanism of how heterochromatin clustering is initiated and maintained, phase separation is one of the emerging popular theories in recent year. It has been found that the HP1 can form an oil droplet-like structure in fluid *in vitro*, and a similar structure was also seen in the nucleus of interphase cells. Due to the hydrophobic interaction between HP1 proteins, small droplets fuse together to form larger droplets over time, separated from the water phase[43]. HP1 that has the tendency to

separate from the hydrophilic phase bind to certain DNA regions, drive the relocation of chromatin and form heterochromatin structure.

Lamina-associated domain, replication timing domain and A/B compartment

DNA with similar functions and activities can move and cluster to specific areas. The 46 chromosomes in human cells each occupy a discrete territory in the cell[44]. The heterochromatin part of each chromosome is frequently located near the peripheral area in the nucleus, interacting with the lamina protein on the nuclear membrane. Chromatin regions with lamina contact is hence termed lamina-associated domains (LAD). LADs are mainly composed of inactive chromatin regions and genes with low transcription activities[45]. Nevertheless, positions of the genes and other regions on the chromatin are dynamically changing in the nucleus. For example, the FISH experiment showed that many genes reposition when their transcription is turned on, as they migrate from the periphery to the center of the nucleus[46].

DNA replication also gives rise to formation of regional structures. In the process of replication, the chromosomes of eukaryotic cells appear to be multi-centered and timely-ordered, which means different DNA regions have different replication timings. On average, every specific 400Kb to 800Kb of the genome has internally similar and continuous progression of replication time-wisely. Such a unit is called replication timing domain[47]. Interestingly, early replication domains are mostly spatially open chromatin regions, while late replication domains are mostly tightly compacted regions, and a considerable part of it overlaps with LADs[47]. Comparing the replication timing domains between humans and mice, they are highly conserved with synteny during evolution[48], indicating that this type of domain have essential biological significance in mammals.

The chromatin structure can also be comprehended by non-microscopic approaches like chromatin conformation capture assays (3C). By such method chromatin is also classified into two

majorly different compartments, termed as A and B compartment. The two compartments harbor abundant chromatin interactions within the same compartment, but much fewer connections between A and B compartments[49]. Each chromosome in the human genome have both A and B compartments, and the linear length of each compartment is about several million basepairs. The compaction of A compartment is relatively loose and open. It is rich in genes and exhibits active chromatin status, marked by higher histone 3 lysine 27 acetylation (H3K27ac). It also has a high GC content, and is usually located in the center of the nucleus. In contrast, compartment B is more densely compacted than compartment A and contains fewer genes. Often, it is located at the periphery of the cell nucleus, which usually is also where heterochromatin is located. Compartment B also has more inhibitory epigenetic modification like histone 3 lysine 27 trimethylation (H3K27me3). Gene expression in this region is usually low or silenced[49]. Interestingly, the A/B compartment most of the time corresponds to the DNA replication timing domain, where A mainly corresponds to the area where the replication fires earlier, and the B is the area where the replication occurs later[47]. Moreover, the B compartment located in the periphery of nucleus often overlaps with the aforementioned LADs[49].

Loop and Stripe

DNA looping is the basic spatial organization of chromatin[50, 51]. Looping is the process that a part of a chromatin protrudes outwards and the flanking regions keep getting gathered into the extrusion to form a ring structure. Regions at the two sides of the converging point of the loop can be originally far away on the linear DNA, but are in proximity spatially. This process is mediated by some structure proteins, mainly CTCF, Cohesin, and Mediator[52, 53], as well as various transcription factor, synergistic enhancer, and non-transcribed RNA[54]. Cohesin initiates looping and is indispensable to the process. Auxin-induced degradation of Cohesin eliminates all the loop domains in the genome without affecting compartments or histone marks[55]. CTCF binds to 20% of enhancers and participates in the

formation of some promoter-enhancer chromatin interactions. When CTCF is degraded by auxin-induced system, some chromatin interactions, such as the one between the Yamanaka factor gene SOX2 and its enhancer, will disappear or get weakened but the global chromatin organization is preserved[56]. In addition, the histone H3K4 modification enzymes MLL3 and MLL4 was also found to be involved in some chromatin loops. Knocking out these two genes in mouse embryonic stem cells (mESCs) not only eliminates some loops, but also affected the binding of Cohesin to DNA[57]. A pair of CTCF binding sites on DNA tend to be in opposite directions[51]. Looping is a way to regulate gene expression, because it can change the distance between the gene and the regulator. The gene transcription can be upregulated or downregulated, depending on what components are in the loop. Approximately 50% of human genes can participate in chromatin interaction by DNA looping[58].

Looping is a dynamic process where there is cell-to-cell variation and a timely variation within the same cell. The loops extrusion models suggests that loops are initiated by cohesion complex (including Rad21 and Mcd1) through entrapping the DNA into its lumen, followed by dynamic sliding of DNA loops until it stalls at boundary-binding protein such as CTCF[54, 59]. Although the mobility of cohesion is hard to be visualized *in vivo* in a real time manner, a “stripe” structure has been seen in Hi-C map, caused by a single anchor forming frequent contacts with a contiguous genomic interval over a population of cells. Stripes is frequently found to be overlapping with regions rich in cohesion loading. A stripe can stretch by several hundred kilobases, serving as a bridge for gene to be tethered to regulatory elements, especially super enhancer clusters: 76% of stripes are linked to enhancers than a random chance of 47%[60].

Topological associated domain

Chromatin loops are not completely dispersed in distribution or independent to each other. Instead, many loops tend to cluster into domains. In 2012, Dixon et al. discovered that the mammalian

genome is organized into specific structure named topological associating domains (TAD) with an average size of 1Mb[61], and majority of genes are regulated within TADs and insulated from elements in other TADs. An important feature of TAD is that its boundaries are always enriched in binding of CTCF, actively transcribed genes, housekeeping genes, and SINE repeat sequences. The loop extrusion model explains the TAD formation as a snapshot of Cohesin complex sliding until it encounters CTCF bound at the TAD[54]. It implies that CTCF/Cohesin anchor long-range interactions that highly represents scaffold of invariant subdomains, while mediator/cohesin sites correlates with individual interactions within or between subdomains, which are spatially dynamic and regulated during cell lineage commitment[62, 63]. This has been further proved by many studies from different aspects. For example, the single-cell Hi-C has seen highly varied individual chromatin loops in individual cell with dynamic anchors, but at scales beyond million of basepairs, these cells all exhibit stable contours of the TAD domains[64]. Not only CTCF and Cohesin are involved in TAD formation, proteins such as MLL3/4, WAPL and zinc finger protein YY1 also play a role[65, 66]. For example, gene knockout of WAPL leads to many new chromatin interactions across the TAD boundary[66].

More and more evidences suggest that TAD itself is a mode of large-scale insulation[67]. Heterochromatin with repressive properties tend to expand on the genome, and such expansion is usually blocked at the TAD boundary. TAD is also a unit of gene expression: genes in the same TAD usually show convergent regulation, so some TADs are in general more active, while others are more silent. TAD is more conservative than chromatin territories. TAD boundaries are mostly stable across different cell types and tissues, such as between human embryonic stem cells and human fibroblasts, and between mouse embryonic cells and mouse cerebral cortical cells. TAD also maintains its conservation over evolution. For example, 54% of the TAD boundaries in the human genome are also TAD boundaries in the mouse genome, and the other way round, this ratio is as high as 76%[61].

2.2 Profiling chromatin structure by microscopy or 3C-based methods

Microscopy-based technologies are historically and widely used for studying the spatial structure of the nucleus and chromatin. In general, those methods can be classified as either light microscopy-based or electron microscopy (EM) -based. For example, the nucleosome polymer structure with a diameter of 11 nm[68], and the 30-nm Z-shipped or the 33-nm solenoid-shaped DNA fibers were all discovered by electron microscopy[69-71]. Cryogenic electron microscopy (cryo-EM) is an emerging popular technology with high resolution and relatively simple operation. However, DNA *in vivo* is almost invisible to cryo-EM. ChromEMT resolves this problem by combining cryo-EM tomography with a selective DNA labeling dye DRAQ5 that photo-oxidizes and mediates polymer deposition of diaminobenzidine onto DNA, which is further visualized by OsO4 staining[72]. The *in situ* chromatin morphology is thus directly visible by EM. For the first time chromatin inside of nucleus were seen by eyes: They are particle-containing fibers with a diameter of 5nm to 24nm, in all phases of cell cycle[72]. Our understanding of the spatial organization of chromatin also comes from the extensive use of FISH. 3D FISH observes interphase DNA in intact nucleus and utilizes confocal imaging technology to label multiple sites and directly measure the spatial distance between sites[73]. FISH is further improved by a series of super-resolution fluorescence microscopy techniques, such as stimulated emission depletion (STED) microscopy[74], photon-activated localization microscopy (PALM) and stochastic optical reconstruction microscope (STORM) [75, 76]. Improved probing technologies dramatically increase the diversity of color and sites to be detected in each single experiment, such as CRISPRRainbow and HIP-map which detect

the multiple to hundreds of loci each time[77, 78].

The application of microscopy is heavily limited by low throughput and low resolution. In 2002, the first chromatin conformation capture experiment (3C) was proposed[79]. This method measures the

frequency of chromatin contacts or proximity as in a population of cells instead of measuring the direct distance between sites. 3C experiment starts with fixation of the cells to preserve the spatial structure of chromatin by formaldehyde, then the DNA is cleaved by restriction enzyme, and by applying DNA ligase, the DNA that is spatially close to each other will have a higher chance to be linked. Then, reverse cross-linking is performed to free DNA from proteins. Primers are then designed based on the two regions that are hypothesized to form contacts, and the result can be tested by PCR or qPCR[79]. Based on 3C, 4C and 5C were developed to test one-to-multiple and multiple-to-multiple chromatin contacts, respectively[80, 81].

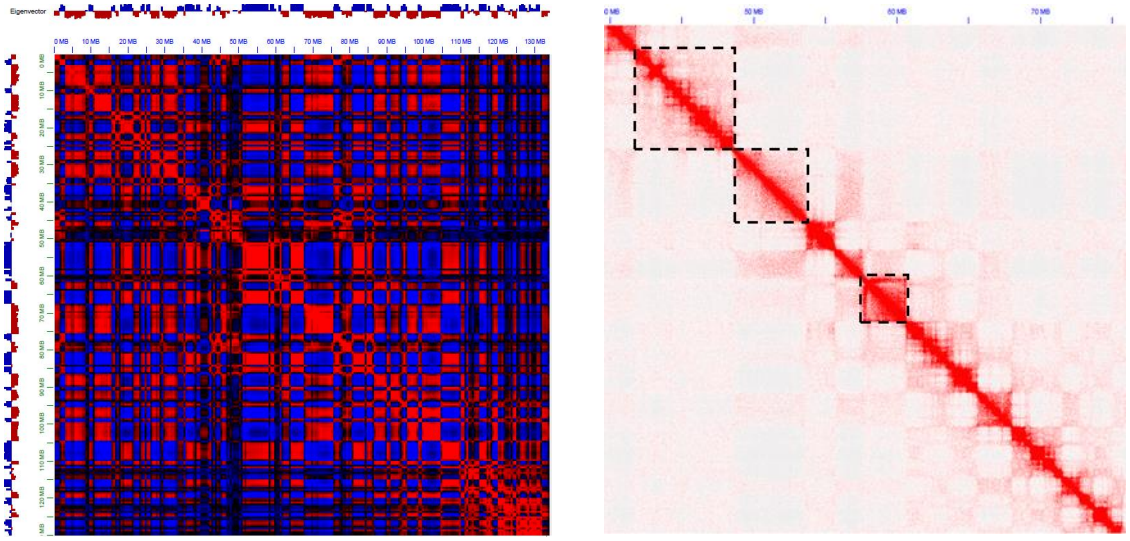


Figure 1- 3: Hi-C reveals compartment and TAD structure of chromatin.

The left panel shows A (red) and B (blue) compartment from correlation matrix of Hi-C matrix. The right panel shows the TAD structure (highlighted by dashed line) from Hi-C matrix heatmap that plots pair-wise chromatin contracts.

In 2009, Hi-C emerged as a revolutionary 3C-based method, for its capability to detect "all regions-to-all-regions" chromatin interactions, including cis- and trans- interactions[49]. Based on all the routine steps of 3C, Hi-C labels ligated DNA with biotin, which was in later steps selected to enrich reconnected DNA fragments from the whole genome. The final readout of Hi-C is sequencing based,

requiring hundreds of million to billions of sequencing reads. Re-ligated reads originated from two loci of the genome were mapped back as pairs, and numerous such pairs eventually form an n×n matrix to describe the frequency of chromatin contacts between any two given regions. This matrix is usually visualized in the form of heatmap, and the chromatin hierarchical structures like compartment and TAD were both thereby identified with specific patterns (**Figure** 1-3).

2.3 Biological significance of 3D chromatin structure

In mammals, the spatial organization of chromatin changes drastically as embryonic stem cell differentiates. In the nuclei of mouse oocyte, the chromatin is evenly distributed, lacking the common TAD or A/B compartment structure. In contrast, sperm still has TAD structure and long-distance chromatin interactions[82, 83]. Upon fertilization, the high-level spatial structure of chromatin almost disappears, and the reconstruction of chromatin 3D structure lasts until implantation of the egg. During this period, TAD and separation of A/B compartments slowly reappear, but DNA coming from the sperm and the egg is still spatially separated[84]. In the subsequent embryonic development into mesoderm, mesenchymal cells, trophoblast-like cells, or neural precondition cells, at least 36% of the human genome has undergone A/B compartment switch[84, 85]. In terminal differentiation stages, more than 56% of the human genomes show compartment switch, across various tissue types such as lung fibroblast cells IMR90, lymphocytes GM12878, prefrontal cortex cells derived from ectoderm, hypothalamic cells, small intestine, pancreas, liver cells derived from mesoderm, ovaries, left and right ventricles derived from mesoderm, cells of the pancreas and adrenal glands, etc[86]. Specifically, while B cell differentiation is accompanied with increase of B compartment and repressive epigenetic modifications[87], the terminal differentiation of pro-pre-B cells to pre-B cells correlates with a

significant B-to-A switch, involving genes with essential functions to B cell maturation, including Ebf1, Poxo1, IgK and IgI[87]. Replication timing domains also changes along compartment switch, which affects 50% of genome.

TAD are relatively stable during differentiation[85]. However, there are exceptions: the TAD boundaries observed in fat precursor cells are inconsistent with those of embryonic stem cells and cerebral cortical cells[88]. In contrast, the "strength" of the TAD, that is, the number of chromatin contacts within each TAD or across TAD, has a more significant change in differentiation, and this number can be increased or decreased. The increase of chromatin contacts is often accompanied with B-to-A compartment switch and upregulated gene expression in the TAD, and *vice versa*. Along differentiation of ESCs into adipocytes, myotubes or nerve cells, new chromatin loops have appeared, linking enhancers to the genes related to the differentiation, alongside the gaining of H3K27ac marks[88, 89]. On the other hand, more pluripotency-related interactions disappeared along ESC differentiation, and CTCF binding was greatly reduced throughout the genome[88]. It is worth mentioning that some areas in TAD are more likely to form internal chromatin interactions than with the surrounding areas, known as frequent interacting regions (FIRE)[85]. FIRE often harbors abundant enhancers and super enhancer groups, for example, more than 77.8% of the super enhancers in GM12878 reside in FIRE[85].

3. Disruption of 3D genome structure in diseases

3.1 Impact of DNA mutations and methylations on chromatin conformation

As the spatial organization of chromatin tightly associates with gene expression, when the chromatin structure is pathologically altered, it can cause dysregulated gene expression that gives rise to

developmental deformity, cancers and metabolic disorders. A variety of causes can lead to changes in the chromatin 3D structure, such as DNA mutations, including single nucleotide substitution/variations (SNV), insertion or deletion of small fragments (<50bp), and structure variations that are large deletion, inversion, duplication or translocations (⩾50bp), and epigenetic alterations like DNA hypermethylation. When these mutations disrupt essential regions like TAD borders that maintain the genomic insulation, structural instability might be seen. One model of loss of insulation is shown in the **Figure** 1-4 below:

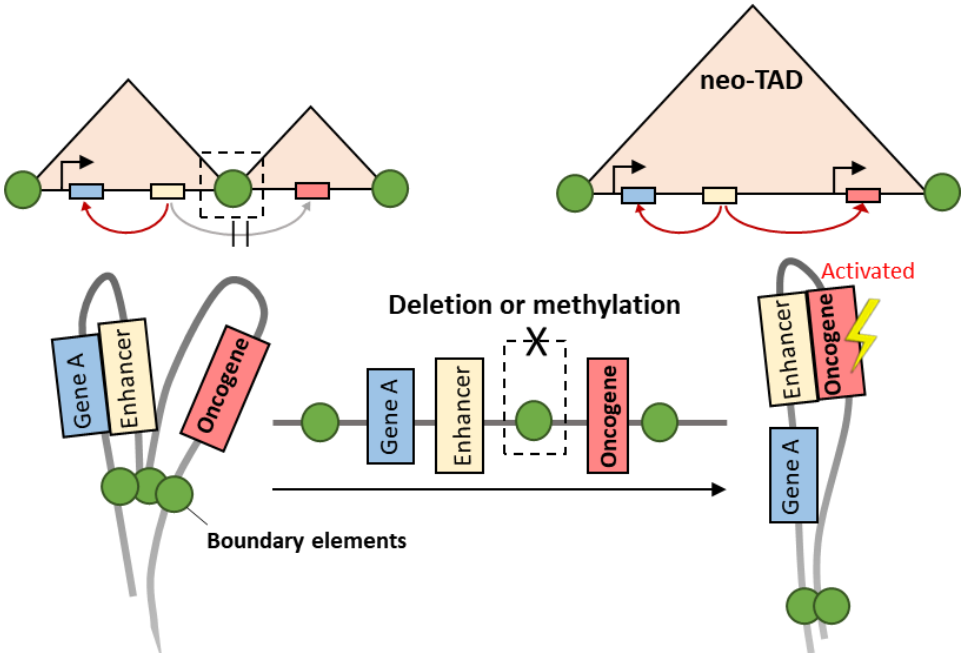


Figure 1- 4: Model of TAD fusion, loss of insulatioin, and enhancer hijacking

The green circle represents boundary elements, majorly CTCF and Cohesin. Mutation, deletion, or DNA hypermethylation of CTCF binding motif all result in prevention of CTCF binding and loss of insulation. Two separate TADs hence fuse into a larger one, and the gene in the right TAD can be activated by the enhancer in the left TAD.

two originally insulated TADs merged to form a larger new TAD (neo-TAD)[90]. In such a situation, genes that were originally silenced in one TAD can be activated by enhancers in another TAD upon fusion, which is termed as "enhancer hijacking"[91]. Opposite to the scenario above, there is also a phenomenon of TAD split. In some cancer cells, the number of TAD is increased, with the average size

of TAD significantly reduced. The smaller TADs in cancer cells often appear as sub-TADs split from the larger TADs in the normal tissue with the boundaries overlapping to each other[92]. At present, the specific mechanism of TAD split is not yet clear. A study interrogating all CTCF binding sites in 1962 samples from 21 cancer types identified 21 insulators showing positive selection for mutations. One of the mutation occurs in more than 16% of melanoma and is associated with *TGFB1* mRNA up-regulation[93]. Profiling of chromatin structure on T-ALL also identifies TAD fusion event and gain of interaction between MYC and distal enhancers[94]. CTCF depletion was also found to induce heart failure and TAD reshaping in mice, while pressure overload in the left ventricle was found correlated with loss of CTCF binding in the heart tissue from the patients[95].

Mutations do not have to be at the TAD boundary to cause disease, in fact, even a single mutation that compromises the interaction between an important gene and its enhancer can lead to severe condition. It was found that globally the genomic regions that interact with gene promoters are enriched in disease-related SNPs[96]. For example, the FIRE that contains a large number of chromatin links is extremely enriched of disease-related SNPs. Every million basepairs of FIRE contains in average 3.33-3.76 diseases SNP, and this number is closer to the upper bound when enhancers are present in the FIRE. In comparison, the reference genome contains around 1.45 disease-related SNPs per million basepairs. So far, millions of non-coding GWAS SNPs have been identified associated with diseases, but the genes they affect and their roles in the disease are mostly unclear. Through the chromatin interactions, genes related to those GWAS SNPs can be preferentially identified, and thousands of new disease-related genes were thus recognized[96].

Hypermethylation of CpG island and CTCF binding sites have been seen in many cancers. In glioma cells with isocitrate dehydrogenase (IDH) mutations, the gain-of-function mutations in IDH prevents DNA demethylation. Excessively high levels of DNA methylation blocks the binding of CTCF,

resulting in failure of insulation between two TADs[97]. In such cases PDGFR is activated by a distal enhancer. Treatment with demethylase reagents on this type of glioma cells partially restored the separation of TAD structure and reduces the expression of PDGFRA[97]. Mutation of succinate dehydrogenase (SDH) also causes DNA hypermethylation. Gastrointestinal stromal tumors with SDH mutation also exhibit loss of CTCF binding and insulations, which results in activation of FGF3, FGF4 and PDGFRA[98].

3.2 Impact of structure variations on chromatin conformation

Given that spatial structure of chromatin has undergone dramatic changes in the early embryonic development, it is conceivable that development can be heavily interfered if these specialized structures were disrupted. Genomic screening on children with chondrogenic dysplasia and polydactyly has found some large-scale deletions, inversions, and duplications downstream of *WNT6/IHH* gene and upstream of *PAX3* gene, which usually encompass a TAD boundary[99]. These mutations can place a cluster of enhancers that was near the *EPHA4* gene into the proximity of *WNT6/IHH* gene or *PAX3* gene, and ectopically activate their expression through the "enhancer-hijacking" mechanism[99]. The abnormal expression of *PAX3, IHH* and *WNT6* genes respectively leads to Brachydactyly, Polydactyly and F Syndrome. *KCNJ2* gene is also directly related to the limb development. In normal cells, this gene is insulated from a series of downstream genes and enhancers by its TAD boundary, including *SOX9*. However, a tandem duplication containing the *KCNJ2* and *SOX9* genes placed the extra copy of KCNJ2 downstream to the original copy of *SOX9*. The enhancers around *SOX9* thus ectopically activate transcription of the duplicated KCNJ2, which directly leads to limb malformation and Cook syndrome.

The alteration of chromatin structures, if created a new interaction between a proto-oncogene and an enhancer, or lost the link between a tumor suppressor and its enhancer, may be carcinogenic. Some T acute lymphoblastoma leukemia (T-ALL) highly expresses TAL1 and Lymo, which keep lymphocytes at a naive stage with uncontrolled active cell propagation. In those cells previous studies found a deletion near the *TAL1* gene, which overlaps with a binding site of CTCF and Cohesin, and abolishes insulation of the TAD[100]. Correspondingly, only in T-ALL was *TAL1* found to form a chromatin interaction with a distal enhancer. Deletion of this CTCF-binding site in normal cells by CRISPR editing recapitulates the TAD alteration, accompanied by TAL1 and Lymo1 transcription activation. [100]. Similarly, DNA mutations at the CTCF binding site is associated with activation of *BRAF* gene in esophageal cancer, and *FGFR1*, *EXT2* and *RBM15* in liver cancer[100]. Sometimes, a single structure variation is sufficient to induce cancer through enhancer-hijacking mechanism. For example, the translocation or inversion of chromosome 3, which represents a subtype with the poorest prognosis in acute myeloid leukemia (AML), places the enhancer of *GATA2* gene into the proximity of the proto-oncogene *EVI1*. The *EVI1* transcription is drastically enhanced, while *GATA2* expression is reduced because of the loss of enhancer and it is functionally haploinsuffiicent[101].

In fact, when examining the SVs and chromatin structure of many different cancer cell lines, we found that, although some cancer cell lines lack the signature mutation on certain proto-oncogene, translocation and large deletions often occur around these genes and correspondingly change the 3D chromatin structure. For example, the prostate cancer cell line PC3 does not have the most common *ETV4* mutations, but it has a translocation between chromatin 15 and 17 next to gene *ETV4*, which incorporate *ETV4* into a neo-TAD[32]. Most of glioma cell lines highly express *MYCN*, which inhibits the expression of *MYC* (c-myc). SK-N-SH and SK-N-AS are the exceptions which highly express *MYC* instead of *MYCN*, and they both have a translocation that mediates a neo-TAD formation and

incorporation of gene *MYC*. Similar examples are a large deletion near *ERBB2* gene in the pancreatic cancer cell PANC-1, and a translocation near *ZNF703* and *TERT* in breast cancer cell T47D, etc., all of which form neo-TAD[32]. Many such examples suggest that it is probably not rare incidences for oncogenes to be activated by SV-mediated change of chromatin structures. How widely and frequently are those events occurring in all types of cancers is an important question, the answer to which might offer insights into developing novel therapies to those cancers.

4. Gap of knowledge

Although we had seen great potential and advantage of BioNano optical mapping, this technology had only been used in genome assemblies back before 2016. It was unclear whether it is feasible at all to utilize the long molecules from BioNano to detect SVs in a cancer genome. Around the same time, more and more Hi-C data had been generated from ENCODE cancer cell lines, and some cell line-specific strange signals on the Hi-C map has drawn attention in the field. Researchers started to suspect that those are caused by cancer-specific SVs, but it remained clueless as what type and locus of SV can cause what specific shape and distribution of aberrant signal in the Hi-C map. Therefore, we wondered whether we can learn and summarize the association between aberrant signals and the corresponding SVs, based on the principle of Hi-C experiment, and then create a model algorithm to detect SVs, including determining their types, loci and orientations systematically based on Hi-C data. Further, we ask whether integration of those technologies would benefit the SV detection in cancer, to improve sensitivity, specificity, and resolution. Specifically, we wonder whether the clinical molecular diagnosis and subtyping of AML can take advantage of the integrative detection of SVs. Landscape of SVs have been studied in many cancer types except AML. One of the possible reason is the lack of

normal blood as control. We hence further asked whether we could utilize the publically available dataset of human genomic polymorphism to stratify germline and somatic variations.

As we showed in introduction, aberrant epigenetic modification can change chromatin 3D structure. While AML is known for carrying lot of mutations that cause loss or gain of functions in epigenetic modifiers, an important question is whether those genomic background will drive alteration of chromatin 3D structure in a subtype-specific manner, and how does the chromatin reorganization contribute the leukemogenesis. Moreover, if recurrent change of chromatin structure was identified, can there be any means to therapeutically restore the chromatin structure as a way to treat disease? Further, since SVs can result in neo-TAD formation and enhancer hijacking events, a question is whether such events also widely occur in AML, as AML is known for harboring large SVs, such as the classic subtype-defining rearrangements including t(8;21), t(t;9), and inv(16). To answer those questions, a method that can process Hi-C data in the presence of SVs is also in demand. Overall, the chromatin structure change and their association to cancer development is not well known for all types of cancers. Addressing those questions could greatly advance of understanding about oncogeneiss and potentially provide novel targets for disease treatment.

Chapter 2

Integrative Detection and Analysis of Structural Variation in Cancer Genomes

Abstract

Structural variants can contribute to oncogenesis through a variety of mechanisms. Despite their importance, the identification of structural variants in cancer genomes remains challenging. Here, we present an integrative framework for identifying structural variation in cancer genomes that applies next-generation optical mapping, high-throughput chromosome conformation capture (Hi-C), and whole genome sequencing to systematically detect SVs in a variety of cancer cells. We identify and characterize structural variants in 36 normal or cancer samples and cell lines. We find that each method has unique strengths in identifying different classes of structural variants at different scales, suggesting that integrative approaches are likely the only way currently to comprehensively identify structural variants in the genome. Studying the impact of the structural variants in cancer cell lines, we identify widespread structural variation events affecting the functions of non-coding sequences, including the deletion of distal regulatory sequences, alteration of DNA replication timing, and the creation of novel 3D chromatin structural domains. These results benchmark six structural variant detection platforms, underscore the importance of accurate and comprehensive structural variant identification, and indicate that non-coding structural variations may be underappreciated mutational drivers in cancer genomes.

Introduction

Here we propose an integrative framework to comprehensively detect SVs by using a combination of technologies, including WGS, next-generation optical mapping (BioNano Irys), and high throughput chromosome conformation capture (Hi-C). Although Irys and Hi-C have been previously used for genome assembly [40, 102-109], this is the first time that WGS, optical mapping and Hi-C technology are systematically compared and integrated for SV detection in cancer genomes. Irys optical mapping works by first introducing single-strand cuts in DNA molecules with a sequence-specific nicking endonuclease, and then repairing the nick with fluorescently labeled nucleotides [110]. Each DNA molecule is then straightened and electrophoresed through microfluidic nanochannels, through which DNA can migrate only when unfolded. Fluorescently labeled nicks are then imaged within the nanochannels. By aligning images of multiple DNA molecules at specific sites, this technology can generate high-throughput genomic maps for extremely long, single DNA molecules (~200kb – 1Mb).

In addition to analyzing Irys optical mapping data, we develop novel algorithms to use Hi-C data to systematically identify structural variations genome-wide. Hi-C technology was initially invented to investigate genome-wide chromatin interactions [49] but has been recently adopted for other purposes, such as genome assembly [104, 105] and haplotype phasing [111]. While the presence of structural variants has been observed with Hi-C datasets [51, 104, 112, 113], we have developed and validated an algorithm to use Hi-C data to find structural variation in cancer genomes. We demonstrate that Hi-C can accurately detect structural variants in cancer genomes even with modest sequencing coverage (20-100 million reads or 1-5X coverage). We compiled a list of high confidence SVs in 8 human cancer cell lines by comparing the results from each of the three technologies, and we performed validation experiments on a subset of these variants. We observed that each method can detect distinct subsets of structural variations: Irys optical mapping and Hi-C excelled at detecting large and complex structural

alterations, whereas high coverage WGS was adept at identifying insertions, deletions, and rearrangements with high resolution. Having obtained large-scale genomic structural information from Irys and Hi-C, we also investigated the consequences of these large-scale structural alterations in cancer genomes. We identified numerous instances of novel 3D chromatin structure alterations as a result of structural genome variation, such as the formation or dissolution of topologically associating domains (TADs), suggesting a critical role for structural variation in gene misregulation in oncogenesis.

Results

An integrated approach for structural variant detection

To evaluate the ability of diverse experimental methodologies to identify structural variants, we identified SVs and performed cross platform comparisons using a combination of whole-genome sequencing, optical mapping, and Hi-C in 8 cancer cell lines and one normal control (GM12878) (**Figure** 2-1a and **Table** 2-S1). We performed WGS in seven cancer cell lines with an average coverage over 30X and we downloaded the WGS data for LNCaP cells from a previous study[114] and for GM12878 cells from the Illumina Platinum Genome Dataset. We performed the initial SV detection with an in-house pipeline that uses pair-ended reads, split reads, and read depth from WGS data by integrating the results from LUMPY, DELLY, and control-FREEC software[115-117], and performed extensive filtering to remove false positive rearrangement calls (**Figure** 2-S1). Next, we performed optical mapping in the same 9 cell lines with an average coverage of ~100X, the most extensive optical mapping effort in cancer cells thus far. We used BioNano Refaligner 6119 and pipeline 6498 to conduct *de novo* assembly and SV detection, and used an in-house pipeline to perform further data filtering (**Figure** 2-S2). On average, we identified ~3,600 SVs in each cell line by optical mapping. Lastly, we

performed Hi-C experiments in 14 cancer cell lines and analyzed an additional 21 previously published Hi-C datasets (**Table** 2-S1)[51, 85, 92, 118-121]. We developed novel algorithms to identify potential re-arrangement events using Hi-C data, including translocations, inversions, deletions, and tandem duplications (**Figure** 2-S3). After comparing and merging the results from each platform, we predicted thousands of insertions and deletions (>50bp), hundreds of tandem duplications and inter-chromosomal translocations, and tens of inversions. We compiled a list of high-confidence SVs, which were predicted by at least two of the three methods. As an example, Caki2 cells carry a translocation between chromosomes 2 and 3 that was detected by all three methods. This translocation was validated by observation of dramatic shifts in DNA replication timing profiles in this region (**Figure** 2-1b). By integrating the SV calls from each method, we can also resolve the structural variants at different genomic scales, from chromosome scale alterations to base-pair sequences (**Figure** 2-1c). We visualized the high-confidence SVs as circular genome structural profiles[122], which showed that the cancer genomes displayed many more rearrangement events compared with normal cells (**Figure** 2-1d, **Figure** 2-S4).

Table 2- 1. Number of high-confidence large SVs in cancer and normal cells

	Confident calls of large intra-chr SVs ≥ 1Mb				**inter-chr TLs**
	Deletions	**Duplications**	**Inversions**	**Unclassified SVs**	
NA12878	0	0	0	0	0
T47D	4	2	6	13	30
Caki2	2	2	5	4	26
K562	4	5	6	11	33
A549	1	2	0	3	12
NCI-H460	2	1	0	0	7
SK-N-MC	2	2	6	3	9
PANC-1	3	0	0	4	14
LNCaP	3	0	0	4	9

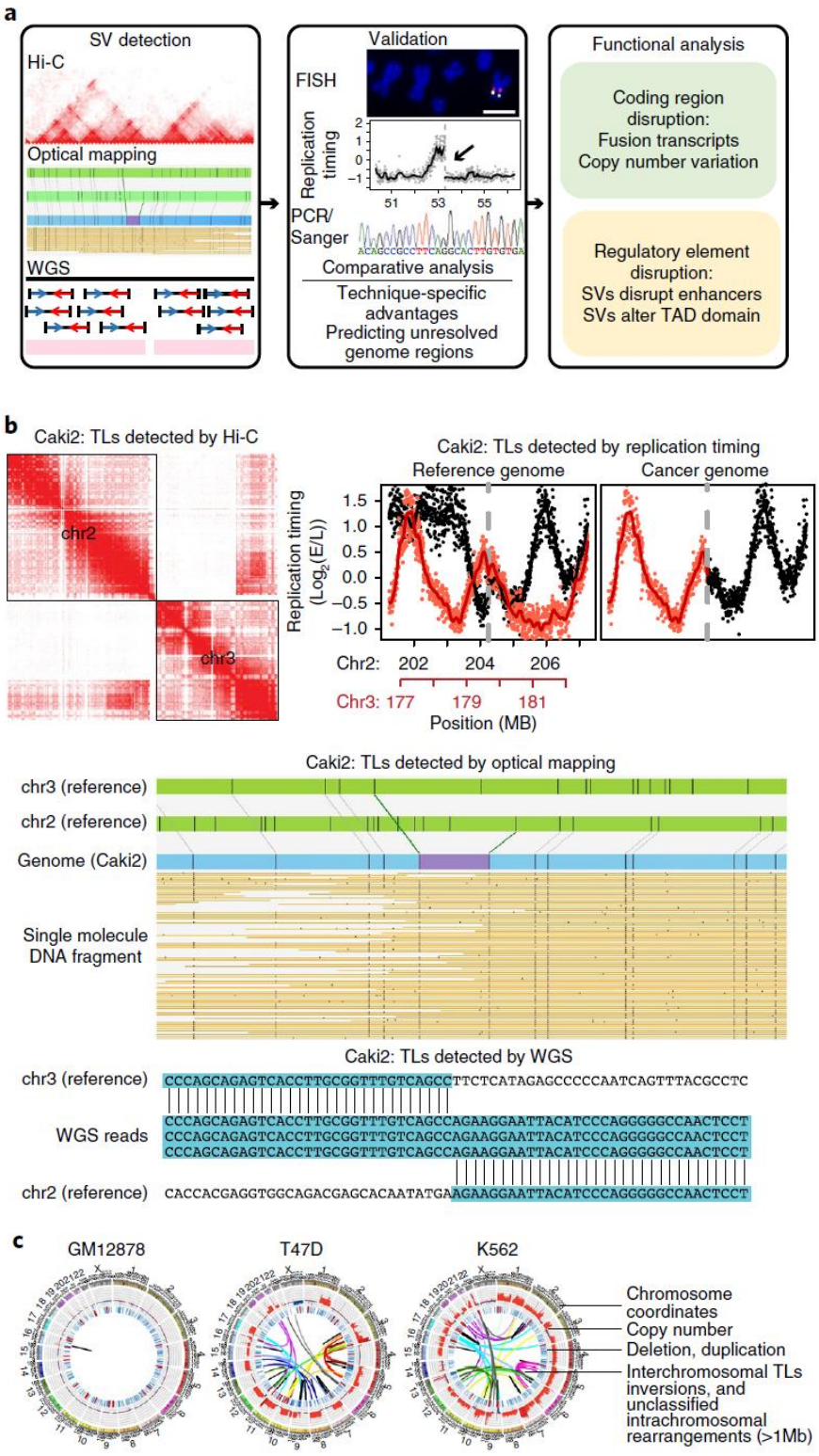


Figure 2- 1. Overall strategy of SV detection in cancer genomes.

a. The pipeline of SV detection, validation, and functional analysis.
b. An example of the same translocations detected by different technologies in Caki2 cells (hg38 coordinates: chr2:204,260,308 and chr3:179,694,900).
c. WGS, Hi-C and optical mapping detect SVs at different scales. Hi-C can detect SVs genome-wide at a scale of up to chromosomal size, while optical mapping can detect SVs and build genome maps at ~10kb resolution. Combining Hi-C and optical mapping can resolve complex rearrangements and reconstruct local genome structure. WGS detects SVs at base pair resolution.
d. Cancer genomes possess more CNVs and translocations in comparison with karyotypically normal GM12878 cells. Tracks from outer to inner circles are chromosome coordinates, copy number, duplications (red) and deletions (blue), and rearrangements including inversions, inter-chr translocations (TLs) and unclassified rearrangements. Outward red bars in CNV track indicate gain of copies (>2, 2-8 copies), and inward blue loss of copies (<2, 0-2 copies). CNVs are profiled by WGS with 50,000 bp bin size. Duplications, deletion, and TLs are detected by at least two methods from WGS, Irys, and Hi-C.

Detection of Large Scale Re-arrangements using Hi-C data

The presence of certain strong inter-chromosomal interactions observed in Hi-C interaction matrices from cancer cell lines has been suggested to be the likely consequence of SVs [51, 104, 112, 113]. Several groups have recently proposed that Hi-C can be used to find translocations in cancer genomes, but have largely relied on visual inspection of the data to identify SV breakpoints[123, 124]. Likewise, tools have recently been developed to identify copy number alterations or inter-chromosomal translocations in Hi-C datasets[123-126]. While locus specific chromosome conformation capture has been used to identify whether individual genes are re-arranged in a given genome[49], to our knowledge, no algorithm has been developed to use Hi-C for systematic, unbiased, genome-wide detection of a full range of SVs, including deletion, inversion, tandem duplication, and inter-chromosomal translocation. To address this, we developed a novel computational algorithm to detect structural variations from Hi-C interaction frequencies.

In a Hi-C experiment in karyotypically normal cells, inter-chromosomal interactions are rare (Left panel in **Figure** 2-2a). However, in the presence of structural variations, dramatic alterations from these low frequency interactions are observed. For example, in Caki2 cancer cells, we observed strong "inter-chromosomal" interactions (Right panel in **Figure** 2-2a), which are due to the fusion of chromosome 6 and chromosome 8. The challenge is to determine whether the increased signal is due to a rearrangement or due to normal variation in 3D genome organization. We have developed probabilistic models of Hi-C data to model "normal" features of normal 3D genome organization, including genomic distance between loci, TADs, A/B compartments, and the increased interactions between small chromosomes and between sub-telomeric regions (see Supplementary methods for details). In the event of a re-arrangement, the two re-arranged regions are genetically fused, altering the linear distance between loci. This leads to local clusters of deviation from the expected interaction

frequencies of the model. This signature can then be used for systematic identification of structural variation in any Hi-C dataset, including both inter-chromosomal translocations and intra-chromosomal re-arrangements (**Figure** 2-2a,b). We use an iterative approach that first identifies breakpoints using low resolution Hi-C contact maps (1Mb bin size for inter-chromosomal, 100kb for intra-chromosomal), and then progressively reduces the bin size to refine the exact breakpoints as high as 1kb in certain instances.

We tested our method with a well characterized chronic myelogenous leukemia cell line (K562) and specifically used a limited number of sequencing reads (27 million read pairs and ~1.5X coverage in replicate 1, generated as part of this study, and 22 million read pairs and ~1.3X coverage in replicate 2, generated previously [51]) to determine whether Hi-C can identify re-arrangements with limited sequencing depth. We started our analysis by comparing large scale re-arrangements identified with a Hi-C bin size of 1Mb with published karyotype data for K562[127]. We found 19 re-arrangements across the two replicates, 11 of which were known through prior karyotyping and the remaining 8 are novel re-arrangement events[127]. The 8 novel re-arrangements were found in replicate experiments performed in two independent laboratories, suggesting that these are not a product of clonal evolution. Interestingly, several of them are novel complex re-arrangements: one event is between chromosome 16 and two different regions of chromosome 6 (**Figure** 2-2c) and in another case, we observed a re-arrangement between chromosome 1, 6, 18, and 20. We performed fluorescence in situ hybridization (FISH) experiments and validated a set of novel re-arrangement events. In total, 18 of the 19 predicted translocations using Hi-C data were validated by either FISH or previous karyotyping (**Table** 2-S2, suggesting that our algorithm can identify large-scale structural variation with high specificity. Next, to estimate the precise location of breakpoints, we iteratively applied the algorithm at increasingly smaller bin sizes to determine a subset of the re-arrangements with high resolution. For example, in K562 cells,

we identified 4 re-arrangements at 1kb resolution, all of which were further validated by PCR and Sanger sequencing.

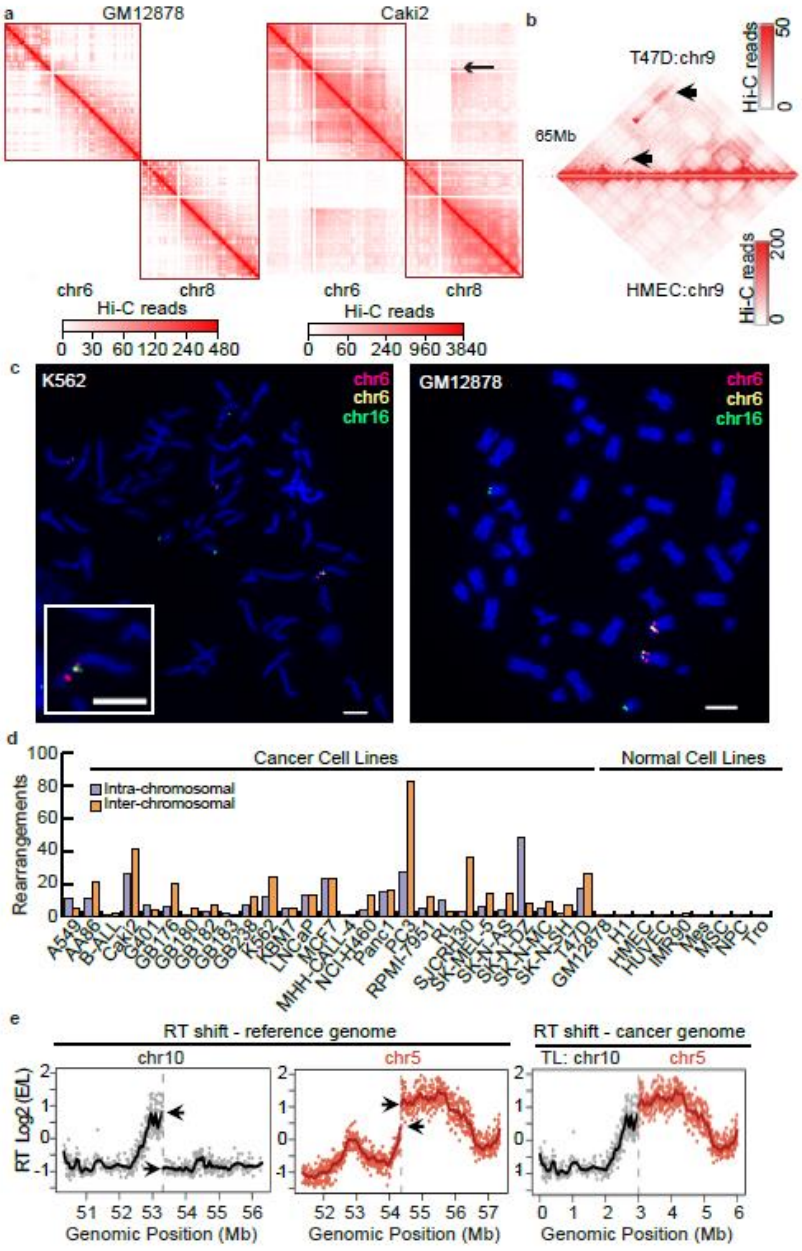


Figure 2- 2. Detection of SVs using Hi-C in cancer genomes.

a,b. Inter-chromosomal (a) and intra-chromosomal rearrangements (b) detected by using Hi-C data (marked by arrow sign). In panel a, GM12878 heat maps are shown at 100kb resolution, and Caki2 are shown at 1Mb resolution **c**. A complex translocation (TL) (chr6-chr16-chr6) in K562 cells validated by fluorescence *in situ* hybridization (FISH). Similar results for FISH validation experiments were performed using 20 independent metaphase nuclei. Scale bars (white) represent

5μM. **d.** Number of inter-chromosomal and intra-chromosomal rearrangements detected by Hi-C in 29 cancer genomes and 9 normal genomes. **e.** An example of the impact of TLs on replication timing (RT). RT profiles of chr5 and chr10 of SK-N-MC, when plotted to the reference genome, show abrupt shifts at the TL breakpoints (←, left panels), and they are smoothly connected due to their juxtaposition in the cancer genome (right panel, normal chr10 is absent in SK-N-MC). Solid black (chr10) and red (chr5) lines indicate loess smoothened RT data. As RT experiments were designed for validation purposes, one replicate was performed for RT experiments.

To further evaluate the sensitivity of our approach, we evaluated its ability to detect the previously identified breakpoints on human chromosome 21 in Tc1 mouse ES cells (**Figure** 2-S 5a). Tc1 ES cells are a mouse ES cell line engineered to carry a copy of human chromosome 21[128]. In the process of establishing this cell line, human chromosome 21 was subject to gamma irradiation [128], leading to massive genomic re-arrangements, a subset of which have been previously identified using PCR and Sanger sequencing [129]. We generated high coverage Hi-C data in Tc1 cells and identified SVs using our algorithm (**Figure** 2-S 5a). By sub-sampling the data, we evaluated the sensitivity of our algorithm at various sequencing depths. The sensitivity ranges between 40%-90% depending on the sequencing depth and method used to call overlap, and appears to plateau when using 100 million sequencing read pairs or more (**Figure** 2-S 5b). In addition, we observe high internal consistency of breakpoints calls when there is at least 50 million reads (**Figure** 2-S 5c,d). These result suggests that our method requires only modest sequencing depths to achieve high sensitivity and saturation of breakpoint calls, and that we can achieve decent sensitivity with as little as 5-10 million reads. By examining the discordant breakpoints, we observe that Hi-C may call breakpoints in identical regions as identified by WGS, but identifies a different strand as part of the breakpoint (**Figure** 2-S 5e-h). The discrepancies are usually involved with complex events, where Hi-C reported the larger scale SVs, while WGS reported the smaller SVs for the sample complex event. This suggests that Hi-C may retain more information regarding the large-scale structure of the re-arrangement. Lastly, to evaluate the effect of sample

heterogeneity, we simulated mixed tumor/normal samples by combining Hi-C reads from K562 as well as karyotypically normal GM12878 cells at various fractions. When fixed at a total sequencing depth of 100 million reads, we observe limited loss of sensitivity even with tumor fractions as low as 30%, indicating that Hi-C based SV finding is largely robust to moderate sample heterogeneity (**Figure** 2-S 5i).

Having demonstrated the sensitivity and specificity of our approach, we expanded our Hi-C analysis to 27 additional cancer cell lines and 9 karyotypically normal lines (**Figure** 2-2d). We observed on average 25 re-arrangements in cancer cells and virtually no such events in normal cells. The rare instances of re-arrangements in normal cells typically occur immediately adjacent to centromeres and therefore potentially represent anomalous or polymorphic assembly differences. In total we identified 698 rearrangements across all 27 cancer cell lines. Of these, a majority are inter-chromosomal rearrangements, with a roughly 2:1 ratio of inter-chromosomal to intra-chromosomal rearrangements identified (424 inter-chromosomal, 274 intra-chromosomal). Interestingly, in some lineages this pattern is reversed. For example, we identify 48 intra-chromosomal rearrangements and 8 inter-chromosomal rearrangements in SK-N-DZ cells. Of the intra-chromosomal rearrangements in SK-N-DZ, 46 of 48 occur within chromosome 2 alone, suggesting the presence of a complex chromosomal rearrangement in chromosome 2 in this cell line. Finally, we also analyzed the size distribution of intra-chromosomal SVs identified by Hi-C (deletions, inversions and tandem duplications). Hi-C appears to identify mostly large structural variants, with only 4.3% of intra-chromosomal SVs being less than 2Mb in size (**Figure** 2-S 5j). This is likely due to the fact that the strongest “normal” Hi-C interactions tend to be local and due to genomic features such as TADs and loops[61, 130]. As a result, Hi-C appears to have reduced signal-to-noise for finding SVs at the scale of these chromatin architectural features.

Validation of Hi-C breakpoints by replication timing

We also wanted to validate our Hi-C defined breakpoints using an independent functional test, and we chose to use altered patterns of DNA replication timing for this purpose. Eukaryotic genomes replicate via the synchronous firing of clusters of origins, which together produce multi-replicon domains each of which complete replication in a short (45-60 min) burst during S-phase[131, 132]. Genome-wide profiling of replication timing reveals that these domains can be replicated at different times during S phase, with adjacent earlier and later replicating domains punctuated by regions of replication timing transition[131, 132]. Consequently, translocations that fuse domains of early and late replication can result in earlier replication of the late replicating domain and/or delayed replication of the early replicating domain[133, 134]. When mapped to the reference genome, these changes appear as abrupt shifts in replication timing profiles that have the potential to validate breakpoints (**Figure** 2-2g). Our Hi-C pipeline identified 249 translocations (at 10kb or 100kb resolution) in 10 cell lines in which replication timing data is available, including seven new datasets generated for this study. Among them, 75 translocations were associated with an abrupt shift in replication timing. Since an abrupt shift is only expected for translocation between domains that replicate at different times, we aimed to classify the translocations based on the replication timing of the loci. However, the lack of a control cell line that represents the pre-translocation replication timing of the loci confounds this classification. To circumvent this problem, we classified the genome into regions that are constitutively early replicating (CE), constitutively late replicating (CL) and regions that switch replication timing during development (S), using 48 replication timing profiles of non-cancerous cell lines and differentiation intermediates spanning all three embryonic lineages[135, 136] (See Supplementary Methods). Among the 249 translocations detected by Hi-C, 9 were CE to CL fusions and 32 were CE to CE or CL to CL fusions. As expected, an abrupt shift in timing was identified in CE to CL with a much higher frequency (~67%)

than in CE to CE or CL to CL fusions (~13%). Translocations between CE were observed with a frequency 3 times higher than expected by chance, which is consistent with previous reports linking chromosomal breakpoints to early replication and higher transcriptional activity[137, 138]. Overall, replication timing can provide functional validation of a specific class of translocation events that fuse regions that are replicated at different times in S phase.

Cross-platform comparison and integration of SV detection

To systematically evaluate the performance of each method for detecting SVs, we compared the large SVs predicted by Hi-C, optical mapping, WGS, fusion transcripts, karyotyping[127, 139-146], and paired-end tag sequencing (PET-seq)[147, 148] (**Figure** 2-S 6). We then defined rearrangements detected by at least two different platforms as high-confidence SVs, and we compared the results from each method to this high-confidence set to assess their contribution and overlap rate as a way to approximate their sensitivity and specificity. Contribution of a given method is defined as the fraction of high-confidence SVs that are detected by this method, and overlap rate refers to the proportion of SVs from one method that overlap with high-confidence SVs.

Overall, we observed that 20% of all inter-chromosomal translocations were identified by at least 2 platforms. Compared with previously known karyotypes in each lineage, many of them are novel. For example, 14 out of 26 translocations in T47D cells found in this study have not been reported before. We selected eight of them for further validation, and all of them were confirmed by PCR. We found that Hi-C is a method with significant contribution and high overlap rate (overall 48 % and 66%), and with better performance for inter-chromosomal translocations (53% and 66%) than intra-chromosomal large SVs (43%, 71%) (**Table** 2-S 3). Integration of Hi-C, optical mapping and WGS increases the overall contribution to 90% (their individual contributions are 48%, 40% and 64%, respectively), and this observation holds true for inter-chromosomal translocations (53%, 24% and 56%, respectively, to

88%) and intra-chromosomal rearrangements (43%, 59% and 74%, respectively, to 92%). We also notice that traditional karyotyping has a high overlap rate with the high confidence calls for all kinds of large SVs (88%) and relative good contribution for inter-chr TLs (56%). However, it only contributes to 2% of the high confidence intra-chromosomal SVs. In addition, traditional karyotyping identifies SVs with low resolution (≥5Mb). Therefore, we believe that Hi-C can compensate for the performance of karyotyping for intra-chromosomal SV detection.

We further integrated the results across different platforms in each cell line into a final high-confidence SV call set by merging SV calls and refining the SV boundaries using the breakpoints of the highest resolution. For example, our Hi-C algorithm identified a large deletion on chr10 in T47D cells (~17 Mb deletion), and predicted the first breakpoint is located between chr10:17,760,000-18,300,000 and the second breakpoint between chr10:35,340,000-35,700,000. When we checked the optical mapping, it reported the same deletion between chr10: 18,304,830-35,340,945. The exact breakpoint for each end should be located between the pair of nicking enzymes surrounding it (resolution ~10kb). Finally, we checked the WGS data, which reported the same deletion at base-pair resolution at chr10: 18,307,707-35,335,171. Therefore, this is a large deletion of 17,027,464 bp that was reported by all three platforms, and we reported this event using the WGS coordinates and set its confidence level as 3 (3 platforms). Moreover, we resolved the SV type for a subset of unclassified large intra-chromosomal rearrangements detected by WGS and optical mapping, based on the orientation and the cross-platform classification. For example, Irys reported 24 unclassified intra-chr rearrangements (≥5Mb) in T47D cells. By comparing with Hi-C or WGS data, we were able to identify the types for 9 of them (37.5%), which include 3 deletions, 2 duplications and 4 inversions.

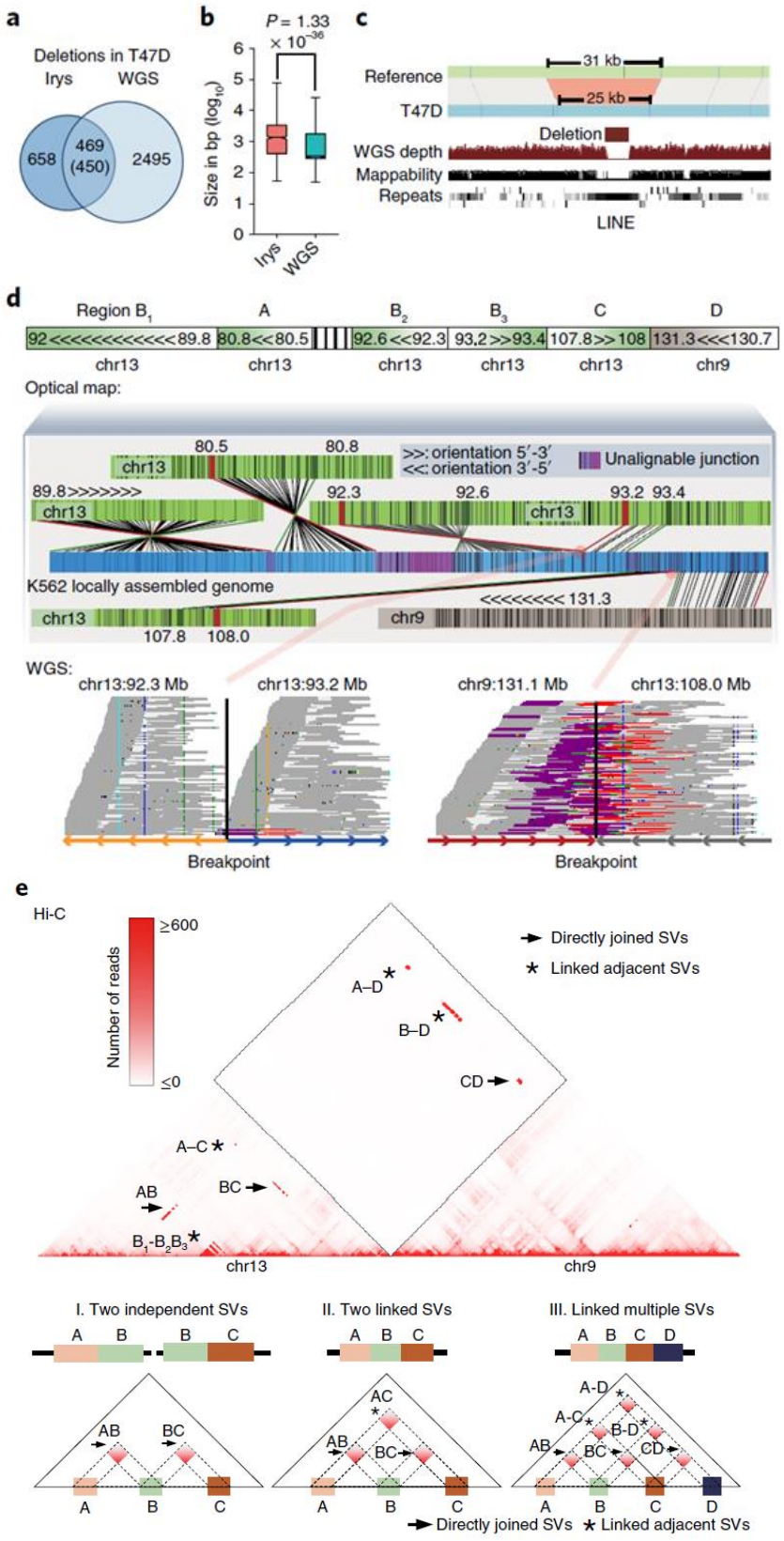


Figure 2- 3. Comparison of SVs detected by different methods.

a. Overlap of deletions in T47D cells detected by optical mapping and WGS.
b. Size distribution of deletions detected by optical mapping (n=1108) and WGS (n=2964, P = 1.33X10^{-36}, two-sided Wilcoxon rank-sum test). For boxplots, the box represents the interquartile range (IQR), and the whiskers extend to 1.5 times the IQR or to the maximum/minimum if less than 1.5x IQR.
c. Optical mapping detects a 6Kb deletion within chrX:96,041,289-96,072,340 that is missed by WGS.
d. Reconstruction of the complex local structure of a derivative chromosome in K562 cells through integration of optical mapping, Hi-C and WGS. The rearranged allele consists of 5 regions: A (chr13:80.5-80.8Mb), B (chr13:89.7-93.3Mb), C (chr13:107.8-108Mb), D (chr9:130.7-131.3Mb), and an unalignable region. Further, segment B consists of three smaller regions (B1, B2, and B3 in the Figure). We reconstructed a global view of the genome structures in this region by stitching several optical mapping contigs together (middle panel). Each junction of the optical mapping genome map can be validated by Hi-C data. WGS data can provide bp-resolution breakpoints for specific breakpoint junctions. Each line in the WGS panel represents a read pair. Reads that support the breakpoint site are parked as purple (forward strand) and red (reverse strand).
e. Strategy of using Hi-C to reconstruct SVs. Hi-C shows increased interaction frequency if two translocated regions are directly joined (→) or if they are not immediately adjacent (*), but are linked to the same rearranged allele.

We also identified thousands of gains or losses of genetic material by optical mapping and WGS in each cancer cell line (we did not include Hi-C for this comparison as WGS is inherently better than Hi-C in detecting CNVs at similar sequencing depths, as both techniques are essentially based on genomic coverage and WGS has more even coverage than Hi-C). We observe that optical mapping detects fewer but larger deletions than WGS. For example, in T47D cells, WGS detected 2,943 deletions with a median size of 552 bp, while Irys reported 1,128 deletions with a median of 1,335 bp (**Figure** 2-3a,b). 84% (2495/2943) of WGS-detected deletions are missed by Irys. Among them, 78% are smaller than 1Kb, which are likely to be missed by optical mapping, as its detection relies on changes to the distance between nicking sites and it is difficult to identify rearrangements considerably smaller than the minimum nicking frequency. Likewise, 58% of the Irys-detected deletions are not captured by WGS. This is likely because optical mapping retains long-range contiguity of large DNA fragments (>150kb) and can grasp the global view of larger deletions that can be overlooked by WGS. In addition, we found 3% of the deletions predicted by Irys overlap with multiple smaller WGS deletions, and in those cases, the summed size of these WGS deletions are very close to the Irys-detected deletion (**Figure** 2-S 7a-c). We tested a subset of deletions detected by Irys and 87.5% (14 of 16) were validated by PCR (**Table** 2-S 2). In addition, optical mapping can identify deletions within repetitive regions where WGS reads are not mapped (**Figure** 2-3c). Further analysis shows that deletions identified by WGS have a higher mappability around the breakpoints compared with those identified by Irys, indicating that Irys can more sensitively identify deletions in low-mappability regions (**Figure** 2-S 7d). Megabase-scale deletions are a signature of cancer cell lines, which is confirmed by both optical mapping and WGS read depth alteration. In contrast, the largest deletion we saw in GM12878 lymphocytes is a 700kb event associated with V(D)J recombination.

We further investigated why some other rearrangements were missed by specific methods (**Figure** 2-S 8). Our analysis indicates that, in addition to mappability (**Figure** 2-S 7d), a major factor contributing to discrepancies in SV calls between different methods is the scale with which the SV is resolved. Specifically, since both Hi-C and optical mapping rely on either restriction enzymes or nickases that recognize target motifs with a spacing greater than 1kb (Hi-C) and 10Kb (optical mapping), they will miss the smaller rearranged regions captured by WGS. On the contrary, since both Hi-C and optical mapping identify rearrangements on a larger scale, they are more capable of identifying complex rearrangements and events whose breakpoints are located in low-mappability regions. For example, one of the rearrangements in K562 cells is located near a centromere of chromosome 20, which is highly repetitive and therefore is unmappable by WGS. However, Hi-C was able to leverage the reads from nearby, mappable portions of the genome to detect the centromere-proximal rearrangement, which we subsequently confirmed by FISH. We also observed that both Hi-C and Irys are particularly powerful at detecting rearrangements with un-alignable junctions (illustrated in **Figure** 2-S 8a and b), which could come from a third chromosome that is too short to be recognized, the non-templated addition of bases to the genome, or exogenous DNA sequences such as that from viruses. By accounting for such differences in scale, it is clear that integration of Hi-C, Irys, and WGS can shed light on the global architecture of complex SVs that involve multiple rearrangements (**Figure** 2-3d). We show an example of a derivative chromosome in K562 that involves massive rearrangements among, at least, chr1, 9, 13 and 22 (**Figure** 2-3d). We use the optical map to thread the putative local structure, the WGS calls to pinpoint breakpoints, (and the Hi-C data signal to validate the linkage of several adjacent rearrangements on the same allele (**Figure** 2-3e, **Figure** 2-S 10). In summary, these results illustrate that WGS excels at detecting simple translocations and CNVs with high resolution, whereas Hi-C and optical mapping are more capable of detecting SVs near un-mappable regions and resolving large and

complex SVs (**Table** 2). Whenever possible, an integrative approach of different methods is essential to gain a more comprehensive understanding of structural variation in cancer genomes.

Table 2- 2. Comparison of three methods

		WGS	Optical mapping (Irys)	Hi-C	Optical mapping (Irys) +Hi-C	Three methods
Resolution of breakpoint	1bp resolution	√√				√√
	10kb resolution	√√	√√	√	√√	√√
Low-mappability region	Whole SV located inside a repeat	√	√√		√√	√√
	Breakpoint located inside a repeat	√	√√	√√	√√	√√
Estimate gap size	-		√√		√√	√√
SV size	Global chromosomal alteration		√	√√	√√	√√
	Deletion	≥1bp	≥100bp	≥1Mb	≥100bp	≥1bp
	Insertion	≥1bp	≥100bp	NA	≥100bp	≥1bp
	Inversion	≥1bp	≥70Kb	≥1Mb	≥10Kb	≥1bp
	Inter-chr TL	≥1bp	≥100Kb	≥10Kb	≥10Kb	≥1bp
Complex SV	Overcome un-alignable junction		√	√√	√√	√√
	Link multiple SVs		√	√√	√√	√√
	Reconstruct structure of complex SVs	√	√		√	√√

√√ *Robust performance.*
√ *Potentially capable of detection depending on variables such as coverage, contig length & label density.*

Better estimation of gaps in human genome

Interestingly, in the process of analyzing deletions detected by Bionano, we noticed that optical mapping can help refine genome assemblies, especially with respect to estimating the size of gap regions. A number of Irys-detected deletions appear in multiple samples including GM12878, and differ substantially when we profile SVs using different versions of the reference genomes (hg19 vs. GRCh38). Further investigation shows that many such “deletions” identified in the hg19 reference by optical mapping consist of gaps in the reference genome. We found that many gaps we re-estimated have been corrected in the GRCh38 build, and the corrected size in GRCh38 is very similar to our predictions (**Table** 2-S 4). For example, when we used hg19 as the reference genome, optical mapping

in 10 cell types (4 normal primary cells and 6 cancer cell lines) recurrently predicted a 143Kb "deletion" within genomic loci chr1: 3,845,268-3,995,268. This region, however, is annotated as a 150Kb gap. Therefore, optical mapping predicts that the real gap size in the human reference genome should be 6.68Kb. In the GRCH38 reference genome, the size of this gap has been corrected to 6.51Kb. However, we noticed that there remain several such "deletions" over gap regions even in the GRCh38 build that are not consistent with our findings with optical mapping, indicating that these gap sizes may still be unresolved, or that there may be heterogeneity the human population in gap sizes over these regions. The improved gap size estimation for GRCh38 is provided in **Table** 2-S 5. We compared our results with two recent studies that also re-estimated the genomic gaps in the GRCh38 reference[40, 103]. While our data show consistency to their results (**Table** 2-S 5), we do observe differences that might be due to population polymorphism. For example, while the size estimation for gap in chr11: 87,978,202-88,002,896 is 24,694 bp in hg38, we observe a range of gap estimates between 889 bp to 1,535 bp across 9 different cell lines derived from different individuals (the estimation is 1,299 bp by Pendleton et al. and 705 bp from Seo et al., respectively). Therefore, the comparison results suggest not only the consistency between our results and previous studies in closing gaps in hg38, but also the variations that potentially represent the polymorphism between cell lines of individuals and populations.

Functional consequences of structural variants in cancer genomes

We investigated the functional consequences of the genetic alterations identified in cancer cell lines. First, we examined gene fusions due to genomic rearrangements, with a goal to both confirm known events and identify novel fusions in these cancer cells. We analyzed RNA-Seq data of 11 cancer cell lines and investigated whether we can detect fused gene transcripts that are consistent with the

genomic rearrangements identified in this study. We detected RNA-Seq read pairs whose two ends are mapped to different chromosomes, crossing the identified translocation breakpoints. Some of these represent well known oncogene transcripts, such as the *BCR-ABL* gene fusion in K562 cells. Importantly, we discovered many novel fusion transcripts involving bona fide oncogenes, such as *EVI1-CFAP70* in T47D cells, whose expression from the translocated exon is over 10 folds higher than that from the non-translocated exon. How these novel gene fusions events contribute to the oncogenic potential remains to be further investigated.

Copy number alterations (CNA) also represent a well-defined class of genetic variation in cancer. Prior studies have shown the presence of recurrently amplified and deleted genes in diverse cancer types [15]. Examining the CNA that we identified and by comparing with the recent findings from WGS data of 560 breast cancer patients [23], we observed that 8 out of the top 10 frequently mutated oncogenes in breast cancer patients were also amplified in T47D cancer cells, and tumor suppressor genes such as *ATRX* and *CDKN1B* displayed loss of copies (**Figure** 2-4a), suggesting that T47D cells reflect the CNV landscape in breast cancer and our method can accurately capture these variations. We further compared the RNA-Seq data in T47D cells with those from human mammary epithelial cells (HMEC), confirming that loss-of-heterozygosity (LOH) and homozygous deletions in T47D cells indeed lead to significantly reduced gene expression correlated to the number of lost copies (**Figure** 2-S 11a). We made similar observations when comparing transcriptomes in other cancer cells (**Figure** 2-S 11b). As an example, one 18Mb deletion in T47D results in LOH of over 400 genes, and decreased transcription of the majority of this set of genes (**Figure** 2-S 11c,d). We found deletions in exonic regions of a total of 25 COSMIC tumor-related genes, and the majority (76%) showed decreased transcription (**Figure** 2-S 11e).

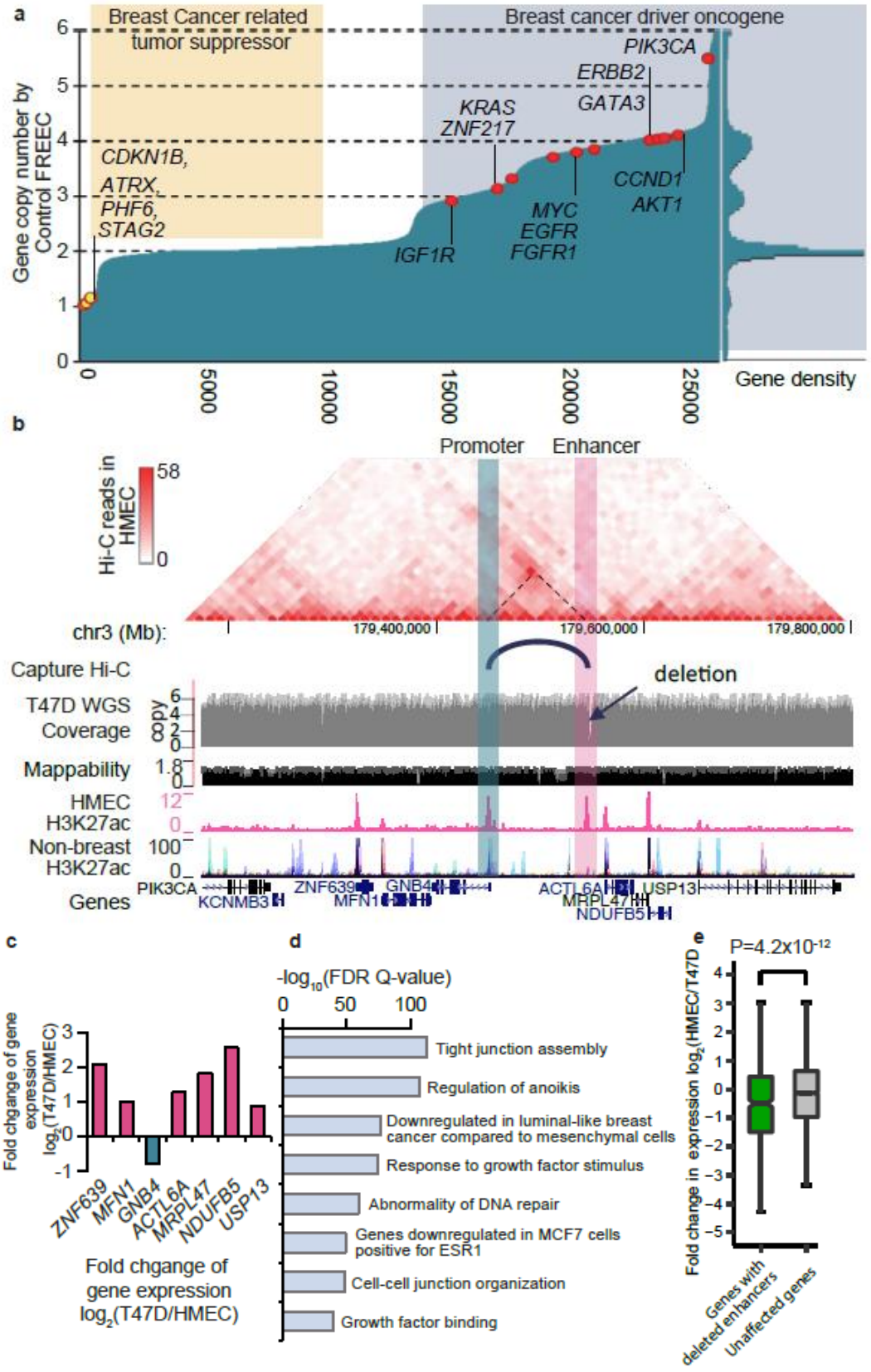

a
Breast Cancer related tumor suppressor
Breast cancer driver oncogene
Gene copy number by Control FREEC
CDKN1B, ATRX, PHF6, STAG2
IGF1R
KRAS ZNF217
MYC EGFR FGFR1
PIK3CA
ERBB2
GATA3
CCND1 AKT1
Gene density
b
Promoter
Enhancer
Hi-C reads in HMEC
chr3 (Mb):
179,400,000
179,600,000
179,800,000
Capture Hi-C
deletion
T47D WGS Coverage
Mappability
HMEC H3K27ac
Non-breast H3K27ac
Genes
PIK3CA
KCNMB3
ZNF639
MFN1
GNB4
ACTL6A
MRPL47
NDUFB5
USP13
c
Fold chgange of gene expression log2(T47D/HMEC)
ZNF639
MFN1
GNB4
ACTL6A
MRPL47
NDUFB5
USP13
Fold chgange of gene expression log2(T47D/HMEC)
d
-log10(FDR Q-value)
Tight junction assembly
Regulation of anoikis
Downregulated in luminal-like breast cancer compared to mesenchymal cells
Response to growth factor stimulus
Abnormality of DNA repair
Genes downregulated in MCF7 cells positive for ESR1
Cell-cell junction organization
Growth factor binding
e
P=4.2x10-12
Fold change in expression log2(HMEC/T47D)
Genes with deleted enhancers
Unaffected genes

Figure 2- 4. The impact of SVs on enhancers.

a. Copy number changes in T47D cells of Refseq genes, sorted by copy number. Genes that are frequently mutated in breast-cancer are labeled if they show amplification (red dots) or deletion (yellow dots). The right panel of this figure displays the density plot of gene copy numbers. b. A ~3.4kb deletion (chr3:179,546,826-179,550,207) in T47D overlaps an HMEC specific enhancer. Hi-C data from HMEC indicates that there is an interaction between the deleted enhancer and the promoter of gene *GNB4*. This enhancer-promoter linkage is also reported in GM12878 cells by the Capture Hi-C data. According to WGS data, the local region is amplified and has 6 copies in T47D cells, but the enhancer is deleted in 5 of the 6 copies. **c**. Compared with HMEC, all the genes in this region in T47D are up-regulated potentially due to the local amplification, except for *GNB4*, whose expression is reduced by ~50%. **d.** Functional pathway analysis of deleted enhancers (n=1859) by GREAT tool (P-value from two-sided Binomial test). **e.** Genes with deleted enhancers show reduced expression levels (two-sided Wilcoxon rank-sum test). Genes with exon deletions or copy number loss are excluded. 534 genes are linked by Capture Hi-C data to at least one deleted enhancer (green), and 10,677 genes are linked to enhancers that show no deletions (gray). For boxplots, the box represents the interquartile range (IQR), and the whiskers extend to 1.5 times the IQR or to the maximum/minimum if less than 1.5x IQR.

These results suggest that our integrated method can accurately capture changes in gene dosage in cancer genomes. As we extended the CNV analysis onto eight cancer cell lines, we noticed widespread amplification of known oncogenes (such as *MYC*) and loss of cell cycle checkpoint genes (such as *CDKN2A/B*, **Figure** 2-S 12). In addition, we found over 100 genes that are highly amplified (≥5 copies) or deleted in cancer cells but were not reported in COSMIC, suggesting their potential roles in cancer (**Figure** 2-S 13).

Interestingly, we found that deletions in cancer cell lines and normal cells differed in their likelihood of disrupting repetitive elements or functional elements in the genome. GM12878 cells are more enriched for deletions in repetitive elements when compared with cancer cell lines (70% vs 50%, the expected value of genome background is ~ 50%), which may be a reflection of different DNA repair mechanisms that are active in germline versus somatic tissue [149] (**Table** 2-S 6). Moreover, deletions of genes and enhancers are depleted in GM12878 cells relative to the genomic background (**Table** 2-S

7), as we found 246 instances of enhancer disruption in K562 cells versus 12 in GM12878 cells. This is not simply due to a greater loss of DNA content in cancer cells. By performing simulations that randomly distribute deletions in the genome, we found that the deletions in GM12878 overlap with significantly fewer enhancers than expectation (12 vs 60, $p<0.001$). In contrast, the cancer cell lines show no such selection against enhancer deletions and instead the number of disrupted enhancers is around the values observed by random shuffling (**Figure** 2-S 14a).

To further investigate which deletions are specific to cancer genomes, we compared the deletions detected by both WGS and Irys with the Database of Genomic Variants (DGV), which contains known polymorphic SVs identified by previous studies in healthy individuals, including the three phases of the 1000 Genomes Project. The majority (95%) of deletions identified in GM12878 cells have been previously identified in the DGV, representing polymorphisms in the population. The fraction of polymorphic deletions is lower in cancer cells at 90% (**Figure** 2-S 11f, **Figure** 2-5a), likely due to the presence of somatic mutations in addition to polymorphic germline variants. In total, cancer cells suffer a greater loss of genetic material compared with normal cells, mainly resulting from novel deletions (**Figure** 2-5b). Further analysis showed that the previously identified polymorphic deletions differ substantially from the novel deletions (somatic mutations). Specifically, polymorphic deletions are enriched for repetitive elements (70% vs 50% genomic background) and depleted of exons (1.5% vs 4% genomic background) (**Figure** 2-5 c-d). For 6 cell lines where we can find control cells lines with enhancer annotation, we found that the polymorphic deletions are also resistant to enhancer loss (empirical $P < 0.005$ in all cell lines tested, **Figure** 2-S 14b). In contrast, the novel deletions are not enriched in repeats or depleted of enhancers or exons (**Figure** 2-S 14c and **Figure** 2-5) Instead, they are enriched in COSMIC tumor related genes (**Figure** 2-S 11f)[150], suggesting that a subset of the deletions are potentially pathogenic.

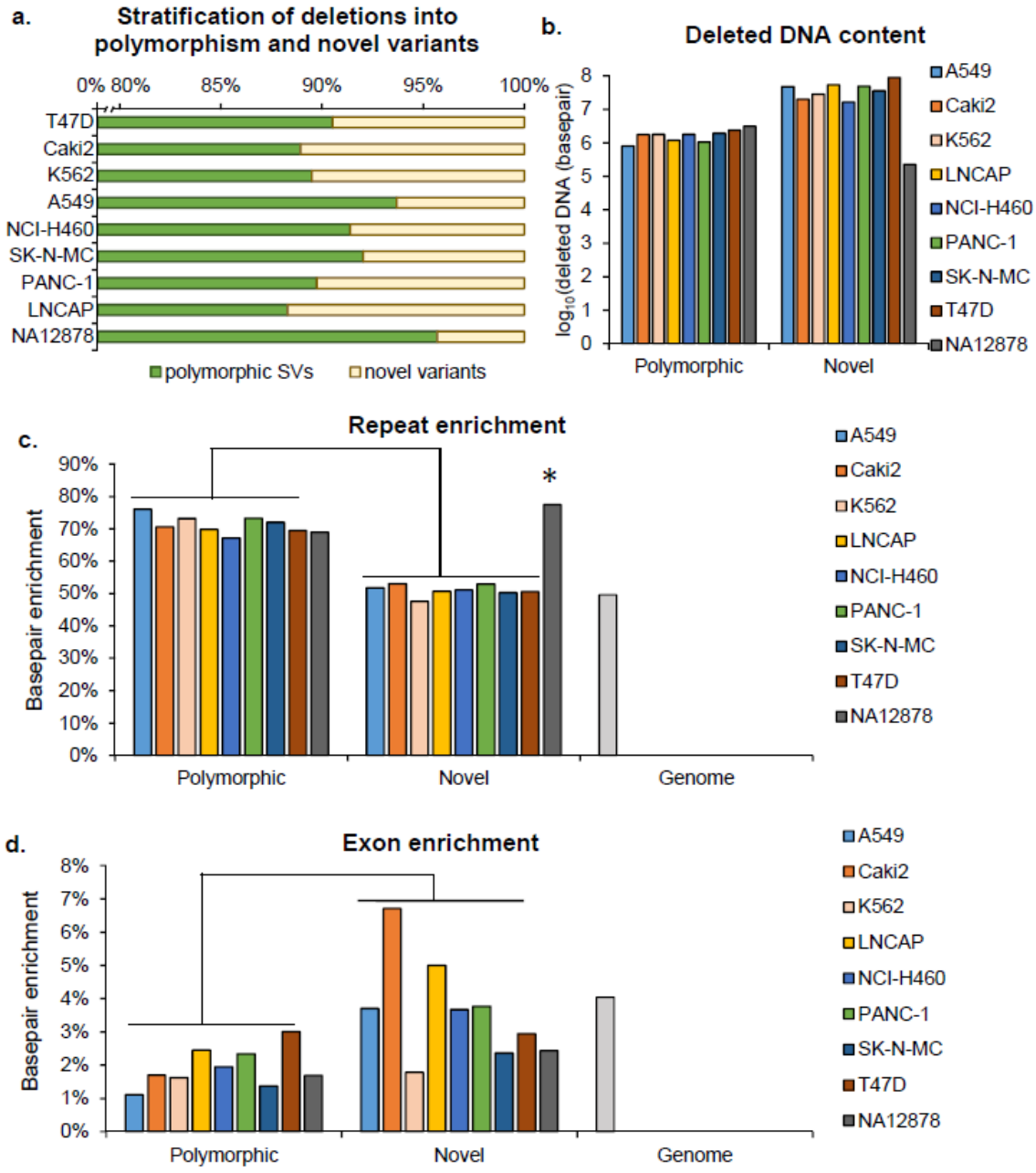


Figure 2- 5. Characterization of known polymorphic deletions and novel deletions.

We stratified the deletions into two categories by comparing with DGV database: polymorphic deletions and novel deletions. **a.** Only 5% of deletions in GM12878 cells are novel variants, whereas on average 10% of deletions found in cancer cells are novel variants. **b.** In cancer genomes, the loss of DNA content due to novel deletions is more than ten times of those induced by polymorphic deletions. **c.** Polymorphic deletions show higher enrichment of repetitive elements (70%) than genome background (50%), whereas novel deletions in cancer cells are not enriched for repeats. **d.** In general, polymorphic deletions are resistant to exon deletions, compared with novel deletions and genomic background.

We also compared the ability of WGS and optical mapping to detect global patterns of copy number variations. Individual fragments generated by optical mapping can be used to detect changes in depth of coverage over given regions, similar to WGS or microarray-based approaches for CNVs detections (**Figure** 2-S 12, 2-S 13). By analyzing chromosome-wide patterns of copy number changes, we detect strong concordance between the results of optical mapping and WGS (**Figure** 2-S 15).

Next, we investigated whether structural variants affecting non-coding regulatory elements can impact gene expression and potentially play a role in oncogenesis. For this analysis, we focused on comparing the enhancer landscape in T47D breast cancer cells and HMEC human mammary epithelial cells. We downloaded histone modification data from both cell types from the Encyclopedia of DNA Elements (ENCODE) Consortium and then predicted candidate enhancers based on H3K27ac signals. By comparing the enhancer annotations in HMEC and the deleted regions in T47D, we identified potential deleted enhancers in T47D cancer cells. We show an example in **Figure** 2-4b of a 3400bp deletion on chr3 about 100Kb downstream of the gene GNB4 (G protein subunit Beta 4) that partially overlap with a breast-tissue specific enhancer. This region has six copies, five of which carry this deletion and only one copy of the enhancer remains undisrupted. GNB4 is likely regulated by this enhancer, as we found strong Hi-C interactions between this enhancer and the GNB4 gene in HMEC cells. Recently published capture Hi-C data[151] also indicates a strong interaction between these loci. While expression of genes nearby are highly upregulated comparing to that in HMEC, possibly due to the increase in copy number of the surrounding locus, GNB4 expression is downregulated (**Figure** 2-4c). The enhancer deletion may be altering GNB4 expression in cis, as we found strong imbalanced gene expression between alleles. To investigate whether candidate enhancer elements deleted in T47D cells are broadly associated with cell growth control, specifically whether they affect any known signaling pathways, we performed Gene Ontology analysis with the GREAT tool. We found that these deleted

enhancers of T47D are located near breast cancer relevant genes, with enrichment for ontology terms such as genes downregulated in luminal like breast cancer, and genes involved in DNA repair (**Figure** 2-4d). Furthermore, we observed that genes linked to these deleted enhancers by capture Hi-C in HMEC cells show a reduced level of expression in T47D breast cancer cells (**Figure** 2-4e). Overall, these results suggest that deletions in cancer genomes may frequently remove enhancers and thereby contribute to oncogenesis. Whether these enhancer deletions represent recurrent alterations to cancer genomes remains to be further investigated in patient samples and validated by additional functional experiments.

The impact of structural variations on 3D genome organization

Having high confidence SV profiles and Hi-C data in the same set of cancer cell lines, we explored how SVs can impact 3D genome organization in cancer genomes. Our previous work in karyotypically normal cells and tissues has suggested that topologically associating domains (TADs) are fundamental features of 3D genome structure that are conserved in diverse cell types and species. Several recent reports have shown that genetic mutations can disrupt TADs and create "neo-TADs"[99, 152] that in turn can lead to mis-regulated gene expression in developmental disorders[99, 152]. Further, recent reports have also indicated that alterations that affect TAD boundaries or CTCF binding sites at specific loci can create new chromatin structural domains leading to mis-regulation of nearby oncogenes through "enhancer hijacking"[91, 100, 153]. However, the extent to which SVs alter 3D genome structures such as TADs genome-wide in cancer cells remains unclear.

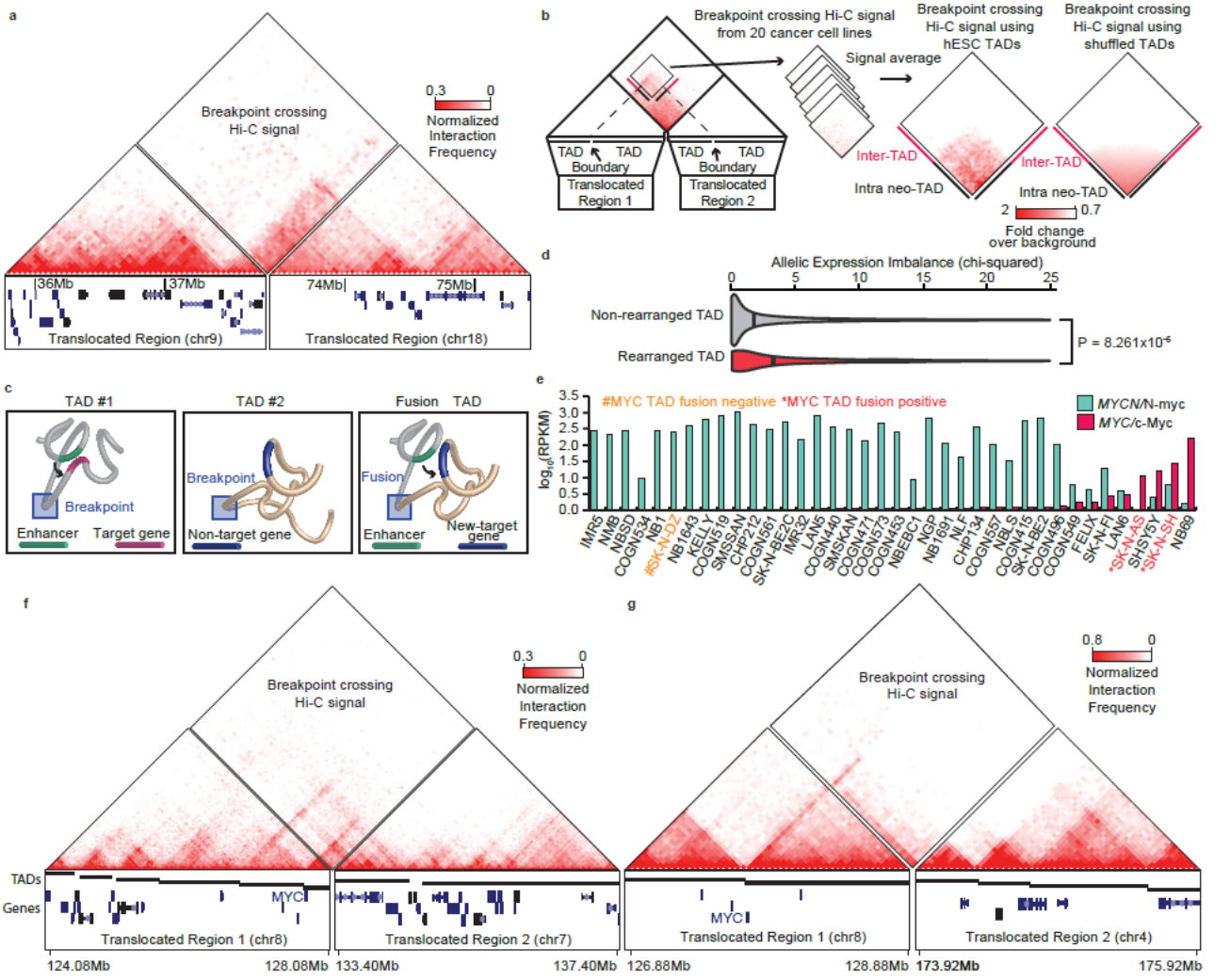


Figure 2- 6. Rearrangements and TAD fusions.

a. Fusion TAD formation as a result of a translocation in Panc-1 cells. The left box shows the rearranged region on chromosome 9, while the right box shows the rearranged region on chromosome 18. The breakpoint fusion lies in the middle. Triangle Hi-C heat maps show intra-chromosomal interactions. The diamond heat map shows the breakpoint crossing Hi-C signal, indicating the presence of a TAD fusion. **b.** Aggregate analysis of TAD fusions. Breakpoint crossing Hi-C signals were averaged and centered on bins between the nearest TAD boundaries (left) or shuffled TAD boundaries (right). The average signal shows a marked enrichment within the regions demarcated by the nearest TAD boundary for true TADs compared to random (randomization performed 1000 times). **c.** Model for neo-TAD formation. TADs are rearranged due to breaks and fusions, juxtaposing regulatory sequences with non-target genes. **d.** Violin plots showing the distribution of allelic expression bias for genes within rearranged (n=1004) or non-rearranged (n=74184) TADs. Vertical bars represent the median (p-value is from two-sided Wilcoxon rank-sum test). **e.** RNA-seq for *MYCN*/N-Myc (green) and *MYC*/c-Myc in neuroblastoma cell lines. Cell lines with TAD fusions at the MYC locus show high levels of *MYC* expression (marked in red), and the cell line that lacks a TAD fusion at the *MYC* locus lacks *MYC* expression (yellow). **f.** Hi-C data from SK-N-SH cells showing a TAD fusion at the *MYC* locus. **g.** Hi-C data in SK-N-AS cells showing a TAD fusion at the *MYC* locus

Having identified structural variants in 20 cancer cell lines with Hi-C data, we systematically investigated the consequences of structural variation on TAD structure in cancer genomes. We observed that neo-TADs are formed as the result of large-scale genomic re-arrangements in cancer cells. An example is shown in **Figure** 2-6a, where the fusion between chromosome 9 and 18 forms a neo-TAD in PANC1 cells. Furthermore, we can find evidence of neo-TAD formation as the result of known, recurrent structural variants in many of the cancer cell lines we profiled, including rearrangements that appear to create neo-TADs surrounding the *MYC*, *TERT*, *ETV1*, *ETV4*, and *ERBB2* genes. To assess whether such events may be altering gene expression in *cis,* we tested whether genes within rearranged TADs were more likely to show imbalances in gene expression between alleles. We analyzed gene expression profiles of 8 cancer cell lines in which we have both WGS and RNA-seq . Controlling for differences in copy number between alleles, we observed that genes within TADs containing a re-arrangement show greater allelic bias than genes within non-rearranged TADs, suggesting that at least a subset of these events are likely leading to altered gene expression in *cis* (**Figure** 2-6b).

To address whether neo-TAD formation is the general consequence of SV rearrangements in cancer genomes, we performed an aggregate analysis of all breakpoint crossing Hi-C signals in each cell line. We reasoned that if breakpoint crossing Hi-C interactions were largely random, they should mainly reflect the distance dependent decay properties of Hi-C data. Alternatively, if breakpoint crossing interactions were non-random and were resulting in fused TADs, we would expect to see the "peak" of a TAD-like "triangle" in aggregated Hi-C interaction plots. When we aggregate breakpoint crossing Hi-C signal, we see a "triangle peak" that is limited by the nearest TAD boundaries (**Figure** 2-6c), indicating that, in general, the nearest TAD boundaries are being fused together into a fusion-TAD as a result of the rearrangement (see supplemental methods for details). This pattern was not observed when performing the same analysis of breakpoint crossing Hi-C interactions using randomized TAD

calls. Under this randomized scenario, interaction frequencies primarily reflect distance dependent decay of interactions from the breakpoint (**Figure** 2-6c – "shuffled TAD calls"). This suggests that, on average, the nearest "normal" TAD boundaries appear to be fused together on each side of the breakpoint creating neo-TADs. These results indicate that structural variations in cancers can re-wire TAD structure to create novel domains in cancer genomes and potentially lead to altered regulatory environments within the domain (**Figure** 2-6d).

We explored one of these neo-TADs which occurred near the *MYC* gene in neuroblastoma cell lines in greater detail. We obtained Hi-C data for 3 neuroblastoma cell lines as part of this study. While most neuroblastoma cell lines express high levels of the *MYCN*/N-myc gene, a subset do not express *MYCN*, but instead express high levels of *MYC*/c-Myc, a pattern of largely mutually exclusive expression that has been described in the literature before[154]. In our study, we profiled one cell line with high *MYCN*/N-myc expression (SK-N-DZ) and two cell lines with high *MYC*/c-Myc expression (SK-N-SH and SK-N-AS) (**Figure** 2-6e). Copy number segmentation from the Cancer Cell Line Encyclopedia indicates that both of these cells lines lack any *MYC* amplification, indicating that the expression changes are not the product of local copy number alterations[155]. Remarkably, in both of the two neuroblastoma cell lines that had high *MYC* expression (SK-N-AS and SK-N-SH), we identified the presence of balanced translocations in the vicinity of the *MYC* gene. In examining the 3D genome structure near this re-arrangement, we observe the formation of neo-TADs that encompass the *MYC* gene in both cases (**Figure** 2-6f,g). Notably, the rearrangement breakpoint occurs 300kb downstream from *MYC* in the SK-N-SH cells, and nearly 1Mb away from *MYC* in SK-N-AS. These results suggest that neo-TAD formation can result in activation of oncogenes even at great linear genomic distances from the precise breakpoint. Determining whether any individual neo-TAD represents a re-current alteration in a given cancer cell type, or how neo-TADs may ultimately contribute to oncogenesis,

remains to be elucidated. However, our analysis suggests that creation of neo-TADs is a common consequence of re-arrangements in cancer genomes.

Discussion

Detecting structural variations in cancer genomes remains a challenge for geneticists and cancer biologists. Here, we developed an integrative approach that employs a combination of WGS, optical mapping, and Hi-C to detect structural variations. We have developed tools for identifying different types SVs from Hi-C data on a genome-wide level for the first time. We tested a selected subset of them by PCR and FISH in three cancer genomes and the results show a high success rate for our integrative approach. No single method identifies all structural variants, and each approach has its own strengths and weaknesses. Hi-C is sensitive for detecting inter-chromosomal translocations and intra-chromosomal rearrangements greater than 1Mb in size. Furthermore, the algorithm we developed can successfully detect rearrangements with as little as ~1X coverage of the genome. However, our algorithm currently has limited power in detecting alterations less than 1Mb in size. On the other hand, optical mapping can readily detect both intra-chromosomal and inter-chromosomal alterations, including rearrangements less than 1Mb in size. Furthermore, optical mapping can be used to detect CNVs, similar to what is commonly done with WGS- or microarray-based approaches. However, compared with WGS, optical mapping cannot identify small deletions and insertions (< 1kb). Finally, WGS has the highest resolution in detecting structural variation. However, WGS is less successful in detecting SVs in poorly mappable regions of the genome or in resolving complex structural variants. One interesting implications of our findings is that tools such as Hi-C can have a dual utility in interrogating cancer genomes, providing information on both genetic and cell biological alterations to cancer genomes.

In examining regions affected by structural variants identified in this study, we observed well-characterized functional consequences, such as the creation of gene fusions and changes in gene dosage in cancer genomes. In addition, we detected extensive deletions of distal enhancer elements. These deletions are enriched for proximity to genes known to be mutated in cancer and important for pathways in cancer biology, including DNA repair and signal transduction. To what extent such distal non-coding mutations are re-current in cancer genomes remains unclear, but this represents an important largely unexplored aspect of cancer genomics. Lastly, by analyzing the 3D genome structure surrounding the structural variants, we observed the creation of new TADs as a result of genomic rearrangements in cancer genomes. We have developed a web-based tool for users to visualize neo-TADs as well as upload their own Hi-C data to evaluate the presence of neo-TADs (available at 3D Genome Browser). TADs appear to be an invariant organizational principle of metazoan genomes, and alterations that disrupt TAD structure have already been shown to underlie certain rare disorders of limb development. There is ample evidence that the juxtaposition of active regulatory sequences to known oncogenes can contribute to tumorigenesis. These results indicate that at least part of this effect may result from the creation of novel structural domains in cancer genomes. Whether all SVs generate fusion TADs, and the extent to which TAD fusion events are recurrent and act as driver mutations in cancer genomes will be an important question for future studies to address.

Supplementary Figures and Tables

Table 2-S 1. List of cell/tissue types with performed experiments and analysis

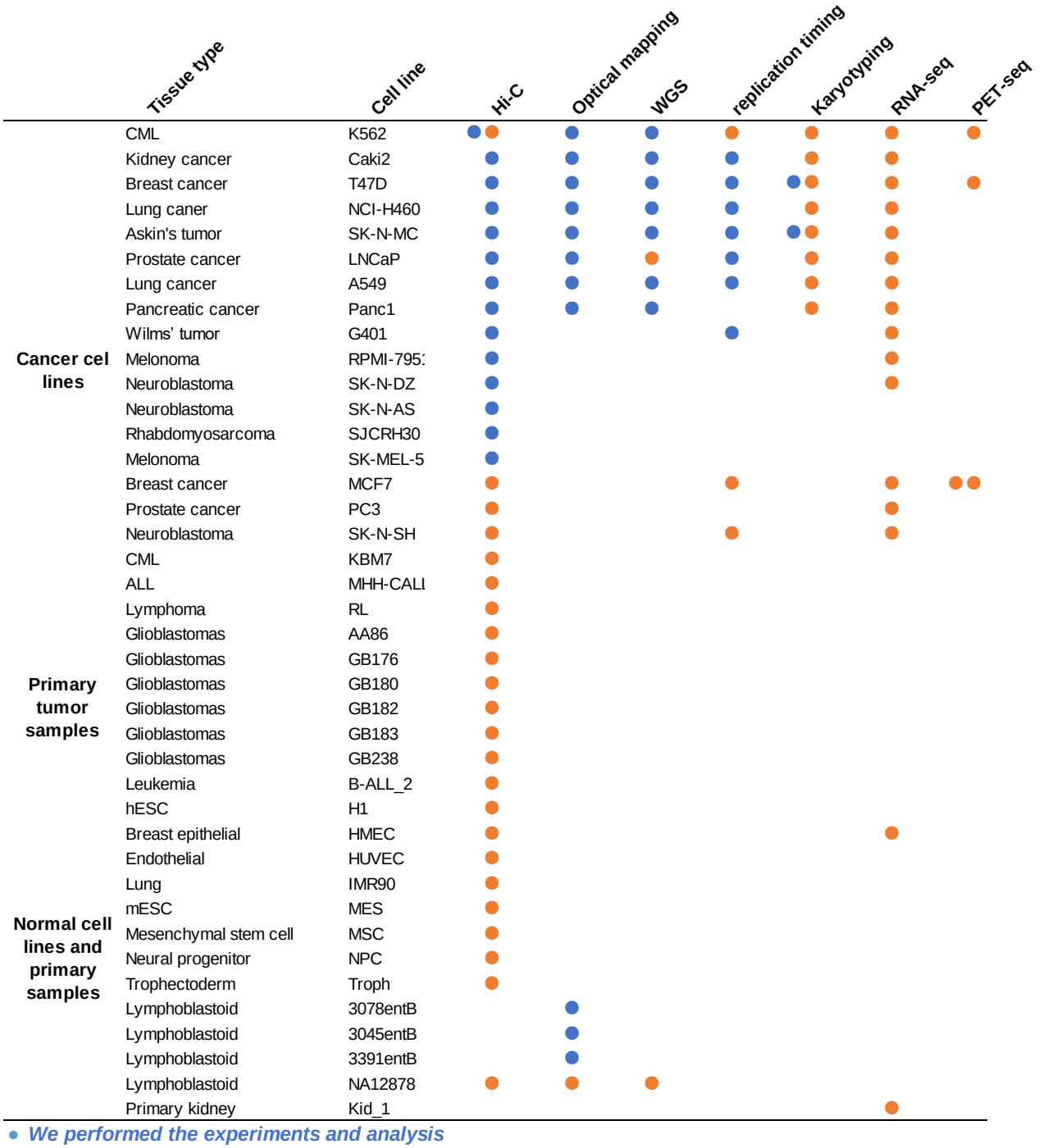


Table 2-S 2. Validated translocations and deletions in K562, CAKI2 and T47D cells

Hi-C source	Region I				Region II				Known/novel	Validation	Result
Validating translocations detected by Hi-C in K562 cells											
Rao et al. Dixon et al.	chr13	+	107854000	108009000	chr9	+	131176000	131280000	Known*		
Rao et al. Dixon et al.	chr13	-	19000000	47000000	chr9	-	27000000	39000000	Known		
Rao et al. Dixon et al.	chr13	+	107800000	108000000	chr22	+	22000000	23300000	Known		
Dixon et al.	chr17	-	27000000	29000000	chr9	+	0	21000000	Known		
Rao et al. Dixon et al.	chr17	-	51000000	57000000	chr9	+	0	21000000	Known		
Rao et al. Dixon et al.	chr17	+	19000000	23000000	chr10	-	43000000	51000000	Known		
Rao et al. Dixon et al.	chr9	-	130731000	131000000	chr22	+	22958000	23291000	Known		
Rao et al. Dixon et al.	chr3	+	48147000	48186000	chr10	+	86065000	86089000	Known		
Rao et al. Dixon et al.	chr5	-	51084000	51094000	chr6	+	37789000	37856000	Known		
Rao et al. Dixon et al.	chr22	-	22500000	22700000	chr2	+	150400000	150900000	Known		
Rao et al. Dixon et al.	chr12	-	22621000	22633000	chr21	-	24258000	24281000	Known		
Rao et al. Dixon et al.	chr3	-	138000000	162000000	chr18	-	26000000	27000000	Novel	FISH	Confirmed
Rao et al. Dixon et al.	chr3	+	138000000	150000000	chr18	+	4000000	8000000	Novel	FISH	Confirmed
Rao et al. Dixon et al.	chr1	-	107000000	112000000	chr20	-	30000000	35000000	Novel	FISH	Confirmed
Rao et al. Dixon et al.	chr1	+	54500000	54800000	chr18	+	24400000	25900000	Novel	FISH	Not Confirmed
Rao et al. Dixon et al.	chr1	+	106780000	106820000	chr18	-	27260000	27450000	Novel	FISH	Confirmed
Dixon et al.	chr1	+	115000000	120000000	chr6	-	135000000	140000000	Novel	FISH	Confirmed
Rao et al. Dixon et al.	chr16	-	85528000	85548000	chr6	-	16766000	16770000	Novel	FISH	Confirmed
Rao et al. Dixon et al.	chr18	-	27000000	27300000	chr6	-	135400000	136200000	Novel	FISH	Confirmed
Validating translocations in T47D											
	chr3	-	136170000	137100000	chr5	+	171830000	171430000	Known		
	chr3	-	169130000	170110000	chr10	+	73240000	73280000	Known		
	chr3	-	193000000	193620000	chr12	-	15100000	15580000	Known		
	chr6	+	46000000	58000000	chrX	+	36000000	64000000	Known		
	chr7	+	86890000	87700000	chr15	-	29660000	30080000	Known		
	chr8	+	36770000	38090000	chr14	-	24870000	25550000	Known		

	chr9	-	68000000	101000000	chr17	-	19000000	36000000	Known		
	chr10	+	53650000	56020000	chr20	-	56170000	56580000	Known		
	chr10	-	56020000	58280000	chr20	+	54080000	56170000	Known		
	chr12	-	21300000	22100000	chr13	-	78800000	79300000	Known		
	chr12	-	15150000	15860000	chr16	-	67060000	67360000	Known		
Hillmer et al.	chr4	+	6590000	6800000	chr5	+	900000	1380000	Reported	PCR	Confirmed
Hillmer et al.	chr6	+	71240000	72130000	chr22	+	16920000	17230000	Reported	PCR	Confirmed
Hillmer et al.	chr9	+	15500000	17340000	chr15	-	27200000	28140000	Reported	PCR	Confirmed
Hillmer et al.	chr5	+	1640000	1750000	chr5	-	40600000	40870000	Reported	PCR	Confirmed
Hillmer et al.	chr9	+	75040000	75340000	chr9	-	103600000	104790000	Reported	PCR	Confirmed
	chr3	+	45740000	46390000	chr9	+	89250000	89420000	novel	PCR	Confirmed
	chr3	-	169130000	170110000	chr10	+	79230000	79650000	novel	PCR	Confirmed
	chr10	+	18080000	18280000	chr10	+	36210000	36880000	novel	PCR	Confirmed
Validated translocations in Caki2											
	chr12		66571831		chr4		64330748			PCR	Confirmed
	chr9		85978709		chr19		45733773			PCR	Confirmed
	chr6		56750050		chr8		58550779			PCR	Confirmed
Validating deletions detected by optical mapping in T47D cells											
	chrX		42652746		chrX		42656304			PCR	Not confirmed
	chr2		212590110		chr2		212720073			PCR	Confirmed
	chr2		97188517		chr2		97190465			PCR	Confirmed
	chr14		104948976		chr14		104951429			PCR	Confirmed
	chr3		58586154		chr3		58586217			PCR	Confirmed
	chr4		165081464		chr4		165083902			PCR	Confirmed
	chr2		28466613		chr2		28469693			PCR	Confirmed
	chr7		6861596		chr7		6887316			PCR	Confirmed
	chr1		207523594		chr1		207546536			PCR	Confirmed
	chr12		58325913		chr12		58339245			PCR	Confirmed
	chr11		107361838		chr11		107374676			PCR	Confirmed
	chr7		97762466		chr7		97773481			PCR	Confirmed

	chr7	70969523	chr7	70979773	PCR	Confirmed
	chr6	85998091	chr6	86007304	PCR	Confirmed
	chr1	53126296	chr1	53129986	PCR	Not confirmed
	chr13	69400712	chr13	69404714	PCR	Confirmed

Table 2-S 3. Contribution by each method and their overlapping percentage with high-confidence SVs

All large SVs (inter-chromosomal TL and intra-chromosomal SVs ≥1Mb)

SV detection methods	Average contribution	Average overlap with high confidence SVs
Hi-C	48%	66%
Irys	40%	43%
WGS	64%	22%
3 Methods	90%	23%
Karyotype	23%	88%
Transcript fusion	18%	NA
PET-seq	73%	12%

Inter-chromosomal translocations

SV detection methods	Average contribution	Average overlap with high confidence SVs
Hi-C	53%	66%
Irys	24%	28%
WGS	56%	15%
3 Methods	88%	18%
Karyotype	56%	88%
Transcript fusion	24%	NA
PET-seq	61%	7%

Intra-chromosomal large SVs (≥1Mb)

SV detection methods	Average contribution	Average overlap with high confidence SVs
Hi-C	43%	71%
Irys	59%	62%
WGS	74%	42%
3 Methods	92%	36%
Karyotype	2%	50%
Transcript fusion	12%	NA
PET-seq	88%	23%

Table 2-S 4. Optical mapping predicts the size of unresolved genome gap in hg19

GAP location in hg19			Gap type	hg19 size (Kb)	Optical mapping prediction (Kb)	hg38 size (Kb)
Prediction consistent with GRCH38						
chr1	3845268	3995268	contig	150	6.68	6.51
chr1	29878082	30028082	contig	150	3.43	3.67
chr1	103863906	103913906	clone	50	-27.22*	-27.07*
chr1	144710724	144810724	clone	100	-101.17	-128.94
chr1	223747846	223797846	clone	50	37.76	35.64
chr1	235192211	235242211	clone	50	23.99	22.43
chr1	248908210	249058210	contig	150	-33.48	-48.49
chr10	47792476	47892476	contig	100	74.46	72.30
chr10	128616069	128766069	contig	150	47.60	40.31
chr10	133381404	133431404	clone	50	11.34	15.93
chr11	69089801	69139801	clone	50	2.65	2.70
chr11	69724695	69774695	clone	50	19.02	18.63
chr11	96287584	96437584	contig	150	12.73	12.16
chr12	7189876	7239876	contig	50	4.15	4.71
chr12	109373470	109423470	contig	50	5.99	5.97
chr12	122530623	122580623	contig	50	3.43	3.36
chr12	132706992	132806992	contig	100	14.96	13.70
chr13	114639948	114739948	contig	100	33.98	33.56
chr15	29159443	29209443	contig	50	3.59	2.94
chr16	8636921	8686921	clone	50	7.07	6.15
chr16	88389383	88439383	contig	50	18.27	17.20
chr18	52059136	52209136	contig	150	5.36	9.14
chr18	72283353	72333353	clone	50	6.45	5.28
chr18	75721820	75771820	clone	50	2.76	1.95
chr19	7346004	7396004	contig	50	0.74	0.00
chr19	8687198	8737198	contig	50	12.60	-0.10
chr19	20523415	20573415	clone	50	-20.00	-22.00
chr2	3529312	3579312	contig	50	6.73	6.18
chr2	5018788	5118788	contig	100	7.94	7.46
chr2	16279724	16329724	contig	50	8.37	8.85
chr2	21153113	21178113	contig	25	1.51	1.89
chr2	110109337	110251337	contig	142	0.85	0.88
chr2	149690582	149790582	contig	100	1.14	1.06
chr2	239801978	239831978	contig	30	19.72	16.95
chr2	240784132	240809132	contig	25	8.59	7.28
chr20	34897085	34947085	clone	50	11.02	9.52
chr20	61091437	61141437	clone	50	29.94	27.85
chr20	61213369	61263369	contig	50	16.36	15.86

chr21	42955559	43005559	contig	50	1.68	1.77
chr22	50364777	50414777	contig	50	6.15	5.22
chr3	66170270	66270270	contig	100	34.86	35.25
chr4	1423146	1478646	contig	55.5	50.73	47.56
chr5	91636128	91686128	contig	50	9.78	10.11
chr5	138787073	138837073	contig	50	5.28	6.10
chr5	155138727	155188727	contig	50	1.52	2.55
chr7	232484	282484	clone	50	12.84	10.03
chr7	50370631	50410631	contig	40	12.01	11.90
chr7	74715724	74765724	clone	50	-164.41	-165.17
chr7	130154523	130254523	clone	100	55.32	55.57
chr7	139379377	139404377	contig	25	9.55	9.95
chr7	154270634	154370634	contig	100	5.44	5.38
chr8	142766515	142816515	clone	50	-15.38	-21.26
chr8	145332588	145432588	contig	100	-57.90	-68.78
chr9	133073060	133223060	contig	150	37.47	36.89
chr9	137041193	137091193	contig	50	22.58	23.25
chr9	139166997	139216997	contig	50	47.92	47.39
chrX	7623882	7673882	clone	50	0.20	0.00
chrX	10738674	10788674	clone	50	-0.22	-0.08
chrX	76653692	76703692	contig	50	15.08	14.97
chrX	148906424	148956424	clone	50	3.00	2.85
chrX	149032062	149082062	contig	50	12.05	10.56
chrX	152277099	152327099	clone	50	-45.33	-59.07
chrY	20143885	20193885	clone	50	-1.15	-0.01
Inconsistent regions						
chr1	205922707	206072707	contig	150	-47.82	445.77
chr1	206332221	206482221	contig	150	-47.82	445.78
chr11	87688378	87738378	clone	50	4.28	27.59
chr13	86760324	86910324	contig	150	24.45	-0.10
chr13	114325993	114425993	contig	100	1.86	51.34
chr15	22212114	22262114	contig	50	-32.44	402.90
chr17	34675848	34725848	contig	50	9.02	21.59
chr17	79709049	79759049	contig	50	9.50	59.15
chr4	8799203	8818203	contig	19	0.52	19.00
chr4	9274642	9324642	clone	50	2.03	50.00
chr4	31820917	31837417	contig	16.5	3.96	16.50
chr4	59739333	59789333	contig	50	8.90	50.49
chr4	75427379	75452279	contig	24.9	-32.77	-0.10
chr5	17530657	17580657	clone	50	29.50	50.00
chr6	157559467	157609467	clone	50	-18.18	-49.90
chr6	157641300	157691300	clone	50	-18.18	-49.90
chr6	167942073	168042073	clone	100	65.45	111.51
chr7	100556043	100606043	clone	50	37.13	-0.12

chr7	143347897	143397897	clone	50	-25.73	50.00
chr8	86576451	86726451	contig	150	137.60	50.00
chr9	92343416	92443416	clone	100	-44.61	13.47
chr9	92528796	92678796	clone	150	5.39	89.17
chrX	37098256	37148256	contig	50	16.56	208.82
chrX	49242997	49292997	contig	50	-42.53	141.86
chrX	49974173	50024173	contig	50	24.49	71.65
chrX	115682290	115732290	contig	50	13.89	47.19
chrX	120013235	120063235	clone	50	-26.06	50.00
chrX	143507324	143557324	contig	50	4.98	51.63
chrY	8914955	8964955	contig	50	30.15	80.43
chrY	9241322	9291322	contig	50	22.29	50.00

Table 2-S 5. Optical mapping predicts the size of unresolved genome gap in hg38 verified by literature

	Gaps in GRCh38			**Predicated gap size from this study**			**Pendleton et al.**	**Seo et al.**		**Genes overlap with gaps**
chrome	start	end	size	median	*range of variation across individuals*		assembly size	Status	added sequence	
chr5	155760324	155761324	1000	0	*0*	*44*	1	Span	0	.
chrX	37099262	37285837	186575	0	*0*	*0*	1	Span	8,498	.
chr4	32833016	32839016	6000	0	*0*	*290*	125	Span	0	.
chr5	139452659	139453659	1000	356	*314*	*356*	264	Span	263	ECSCR
chr12	7083650	7084650	1000	250	*211*	*289*	563	Span	231	C1R
chr13	113673020	113723020	50000	342	*342*	*342*	658	Span	1,206	GRK1
chr4	8797477	8816477	19000	334	*177*	*803*	745	Span	750	.
chr1	223558935	223608935	50000	1285	*439*	*1789*	837	Span	832	CAPN8
chr11	87978202	88002896	24694	1355	*889*	*1538*	1299	Span	705	.
chr7	237846	240242	2396	1864	*1864*	*1864*	2398	Span	1,992	FAM20C
chr2	16145119	16146119	1000	899	*899*	*899*	2541	Span	2,563	.
chr6	95020790	95070790	50000	3588	*3423*	*5719*	3347	Span	1,730	.
chr11	70955696	71055696	100000	4014	*3522*	*4176*	3642	Span	3,647	SHANK2
chr4	1435794	1441552	10606	*3386	**1965*	**6926*	NA	Span	197	.
chr4	1429358	1434206	10606	*3386	**1965*	**6926*	NA	Span	21	.
chr1	16799163	16849163	50000	0	*0*	*0*	NA	Span	0	CROCC
chr10	133690466	133740466	50000	250878	*0*	*40471*	NA	Span	84,551	.
chr13	86202979	86252979	50000	3348	*3267*	*3429*	NA	Span	16,222	.

chr17	81742542	81792542	50000	338	*293*	*383*	NA	Span	359	.
chr21	43212462	43262462	50000	6244	*4841*	*11841*	NA	Span	5,790	.
chr4	31819295	31832569	13274	192	*0*	*794*	NA	Span	0	.
chr4	58878793	58921381	42588	1092	*681*	*1515*	NA	Span	630	.
chr6	167591393	167641393	50000	2110	*1738*	*5242*	NA	Span	686	.
chrX	50228964	50278964	50000	3866	*3708*	*4114*	NA	Span	3,646	CCNB3
chrX	114281198	114331198	50000	5857	*5627*	*6162*	NA	Span	5,703	.
chrX	116557779	116595566	37787	5326	*5206*	*5709*	NA	Span	4,456	.
chrX	144425606	144475606	50000	3775	*3472*	*3956*	NA	Span	3,379	.
chrY	9057608	9107608	50000	0	*0*	*0*	NA	Span	4,965	.
chr14	19511713	19611713	100000	0	*0*	*0*	NA	right	168	.
chr16	33392411	33442411	50000	0	*0*	*36390*	NA	right	18267	.
chr2	97439618	97489618	50000	1143	*1017*	*1270*	NA	right	3450	.
chr22	18659564	18709564	50000	0	*0*	*0*	NA	right	4860	.
chr7	143650804	143700804	50000	0	*0*	*0*	NA	right	483	TCAF2
chrX	115738949	115838949	100000	82287	*80843*	*83730*	NA	right	3571	.
chr22	18239129	18339129	100000	0	*0*	*0*	NA	left	749	.
chrX	120879381	120929381	50000	0	*0*	*0*	NA	left	1557	CT47A8
chr2	89685992	89753992	68000	57497	*53450*	*59960*	NA	Extension	7,215	.
chr18	46969912	47019912	50000	32390	*8358*	*49997*	NA	Both	2850	TCEB3CL, KATNAL2
chr22	18433513	18483513	50000	0	*0*	*0*	NA	Both	2338	.
chrY	9403713	9453713	50000	1945	*1945*	*1945*	NA	Both	4105	.
chr12	37185252	37235252	50000	1768	*1768*	*1768*	NA	.	.	.
chr17	26735204	26735774	570	13	*9*	*17*	NA	.	.	.
chr20	29412507	29413577	1070	0	*0*	*0*	NA	.	.	.
chr6	61357029	61363066	6037	0	*0*	*0*	NA	.	.	.
chrX	49348394	49528394	180000	91174	*91174*	*91174*	NA	.	.	.

* The summed size of two gaps next to each other, when they are too close and the reduced size of gap cannot be resolved for each single one

Table 2-S 6. Summary of genes, repetitive elements and insulators overlapping with high-confidence deletions

Deletions of genes and repetitive elements

Cell lines	No. of confident deletion	No. of genes disrupted by confident deletions	Percentage of repetitive elements (basepair enrichment)
CAKI2	404	586	54.38%
T47D	454	1097	50.98%
K562	435	398	48.99%
A549	237	624	52.16%
NCI-H460	405	414	52.58%
PANC-1	320	558	53.26%
LNCAP	281	658	51.07%
SK-N-MC	487	965	51.34%
NA12878	535	273	69.02%

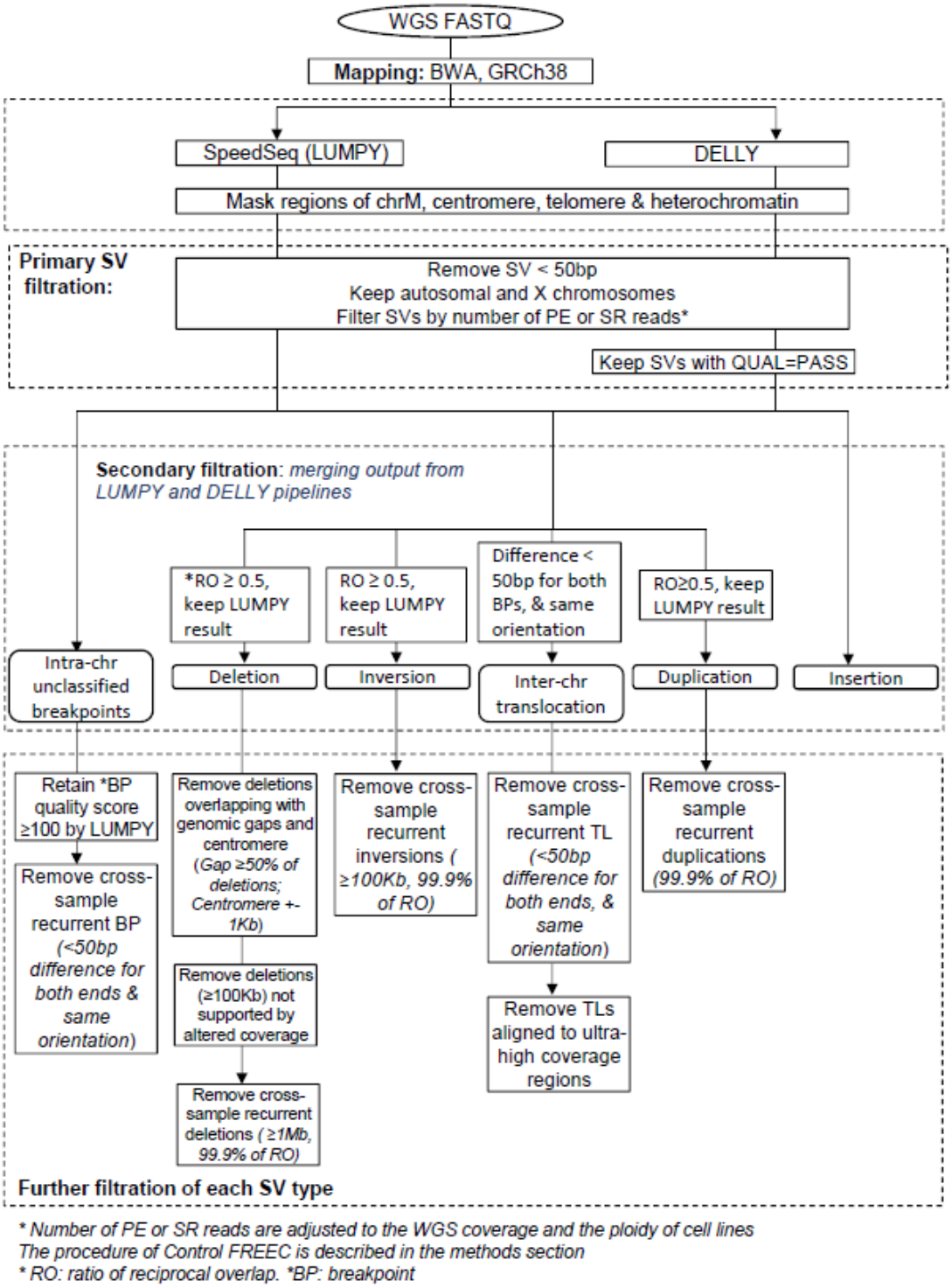
Deletions of insulators

Cell lines	Tissue for enhancer annotation	No. of all confident deletion	No. of deleted insulator annotated by CTCF binding sites (Tissue for annotation)
T47D	HMEC	454	1019 (HMEC)
K562	Primary blood mononuclear cells	435	228 (NA12878)
A549	NHLF	237	2125 (NHLF)
NCI-H460	NHLF	405	663 (NHLF)
PANC-1	Primary pancreatic tissue	320	457 (Primary pancreatic tissue)

Table 2-S 7. Frequency of enhancer deletions versus simulated expectation in cancer cells and normal cells

Cancer cell lines	Number of deletions	Total deleted base pairs	Control tissue	Number of enhancers in control tissue	Number of enhancers deleted in cancer cells VS expectation (P value)	Number of deleted enhancers per 100 Kb deletion	Number of deleted enhancers per 100 Kb deletion, normalized to 100,000 total enhancers
T47D (breast cancer)	454	91,072,914	HMEC (human mammary epithelium cells)	66,066	1859 : 1928 (p = 0.440)	2.04	3.09
K562 (chronic leukemia)	435	29,756,292	Primary blood mononuclear cells	51,862	246 : 484 (p = 0.099)	0.83	1.6
A549 (lung carcinoma)	237	48,725,609	NHLF (lung fibroblast)	91,440	1643 : 1398 (p = 0.696)	3.37	3.68
NCI-H460 (lung carcinoma)	405	18,082,136	NHLF (lung fibroblast)	91,440	467 : 556 (p = 0.373)	2.58	2.82
PANC-1 (pancreatic cancer)	320	49,875,753	Primary pancreatic tissue	78,896	931 : 1237 (p = 0.213)	1.86	2.35
NA12878 (lympho-blastoid)	**535**	**3,359,296**	Primary blood mononuclear cells	51862	0.541666667 (p < 0.001)*	**0.36**	**0.69**

* From left to right, the second and third columns show the total number of deletion incidence and the sum of deleted DNA content (basepair). The fourth column shows the control tissue/cell lines that are in close developmental relationship to the tested cell lines, and we use the H3K27ac marks in the control cell line to annotate the enhancers for that tissue type. The fifth column indicates the total number of annotated enhancers from each control tissue. The sixth column shows the number of deleted enhancer in each cell line, and we also include the result of simulation to approximate how many deletions of enhancers are likely to occur if the deletions are stochastically distributed in the genome. In the seventh and eighth columns we calculate the number of enhancer deletion per 100Kb deletion, and that value normalized to the total number of enhancers from that tissue type.



** Number of PE or SR reads are adjusted to the WGS coverage and the ploidy of cell lines*
The procedure of Control FREEC is described in the methods section
** RO: ratio of reciprocal overlap. *BP: breakpoint*

Figure 2-S 1. Pipeline of structural variants detection and filtration by WGS.

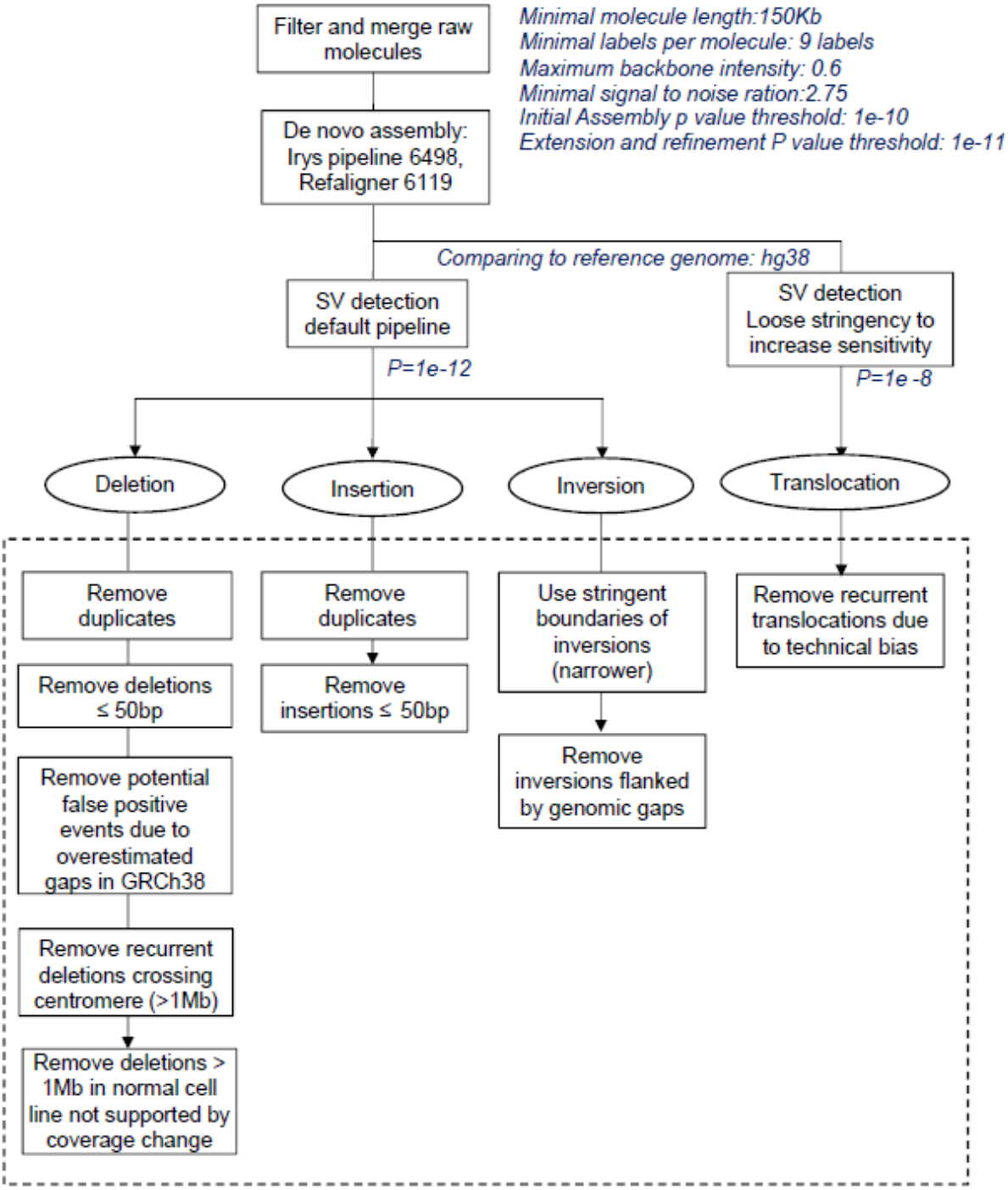


Figure 2-S 2. Pipeline of structural variants detection and filtration by optical mapping.

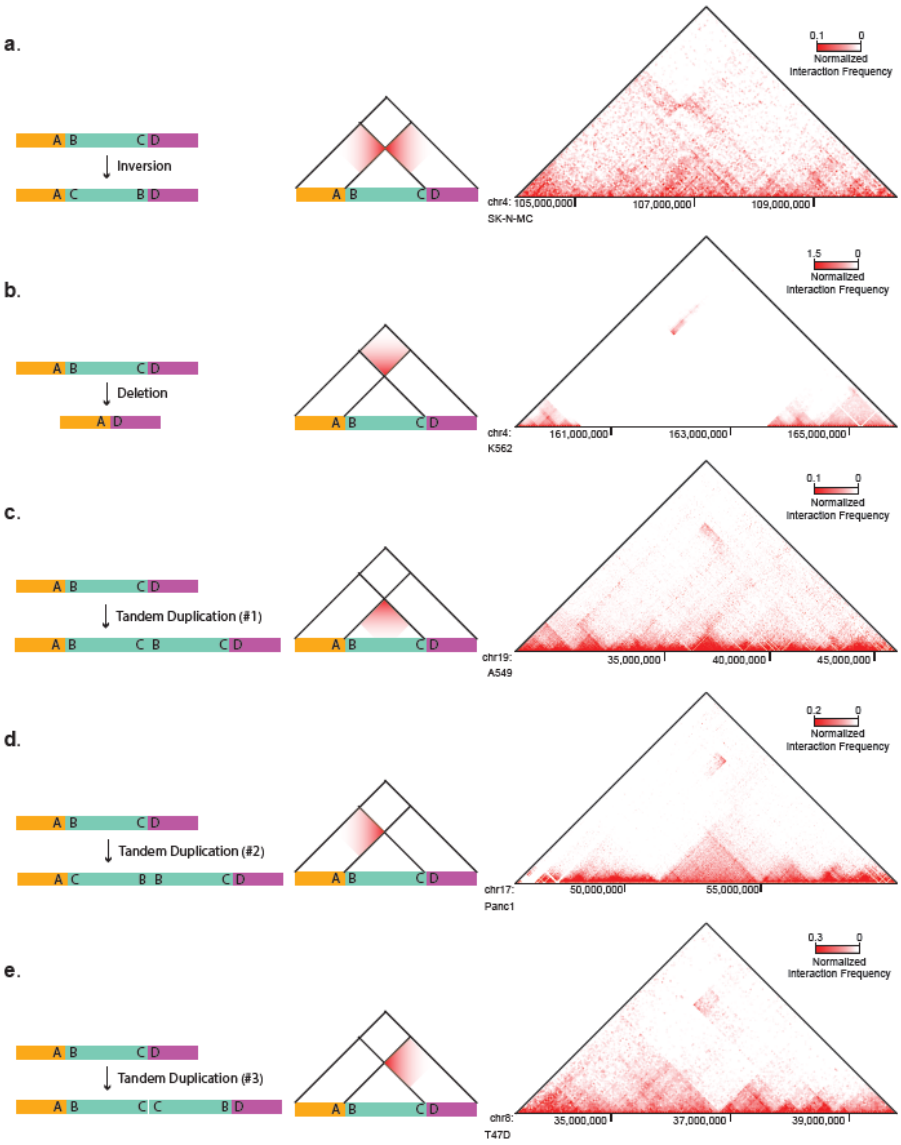


Figure 2-S 3. Hi-C identifies inversions, deletions and tandem duplications.

a. Example of an inversion identified by Hi-C. The left hand cartoon shows the example of the genetic structure of an inversion, juxtaposing regions A and C as well as B and D. The cartoon in the middle depicts the expected alteration to chromatin interaction frequencies by such an event, showing increased interaction frequencies between regions A and C or between B and D as a result of the altered linear proximity of these regions (a "butterfly pattern"). The right-hand panel shows an example of an inversion identified in SK-N-MC cells by Hi-C, optical mapping, and WGS. b. Example of a deletion and its effects on Hi-C data. The deleted region (left panel), removes the B-to-C region in the diagram, and results in the juxtaposition of the A and D regions, which would result in an increase in the interactions between regions flanking the deleted region (middle panel). The right panel shows a deletion in K562 identified by Hi-C, optical mapping, and WGS. c-e. Examples of tandem duplications in Hi-C data, with different orientations of the duplicated region (left-hand diagram), and their expected changes in interaction frequencies (middle panel). The right-hand examples in panel c shows tandem duplications identified in A549 cells by Hi-C, optical mapping, and WGS. The right-hand panel in d shows a tandem duplication in Panc1 cells identified by Hi-C and WGS. Panel e shows a tandem duplication identified in T47D cells by Hi-C and WGS.

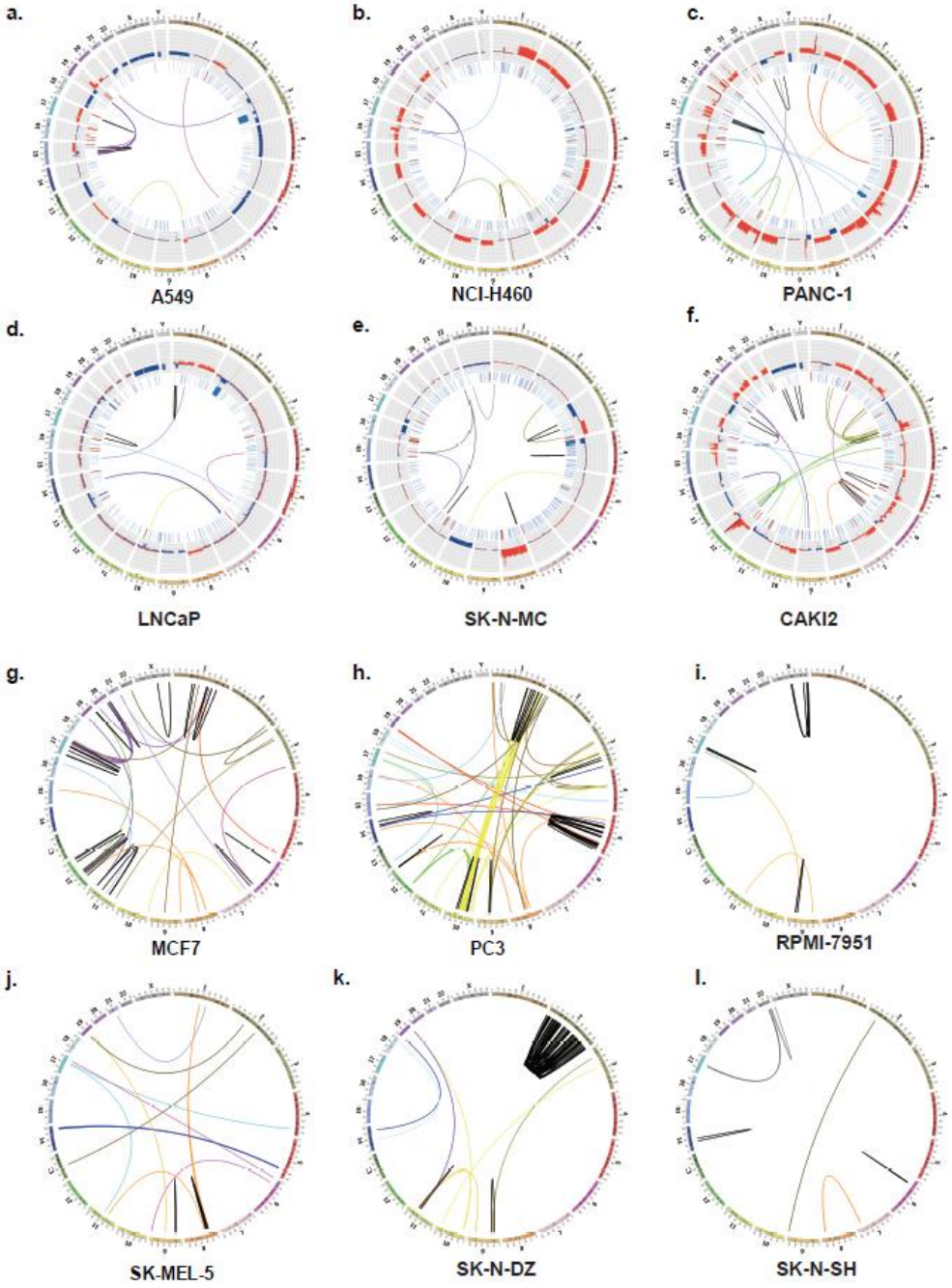


Figure 2-S 4. Cancer genomes possess extensive CNVs and translocations.

a-f. Genome profiles of 6 cancer cell lines. All SVs are detected by at least two out the three methods (Hi-C, optical mapping and WGS). Tracks from outer to inner circles are chromosome coordinates, CNVs, deletions (blue) and duplications (red), and positional rearrangements including inversions, unclassified rearrangements and inter-chromosomal translocations. Outward red bars in CNV track indicate gain of copies (>2), and inward blue loss of copies (<2). CNVs are profiled based on WGS data binned at 50-kb resolution. **g-kl.** Large intra-chromosomal rearrangements and inter-chromosomal translocations detected by Hi-C in 6 cancer cell lines.

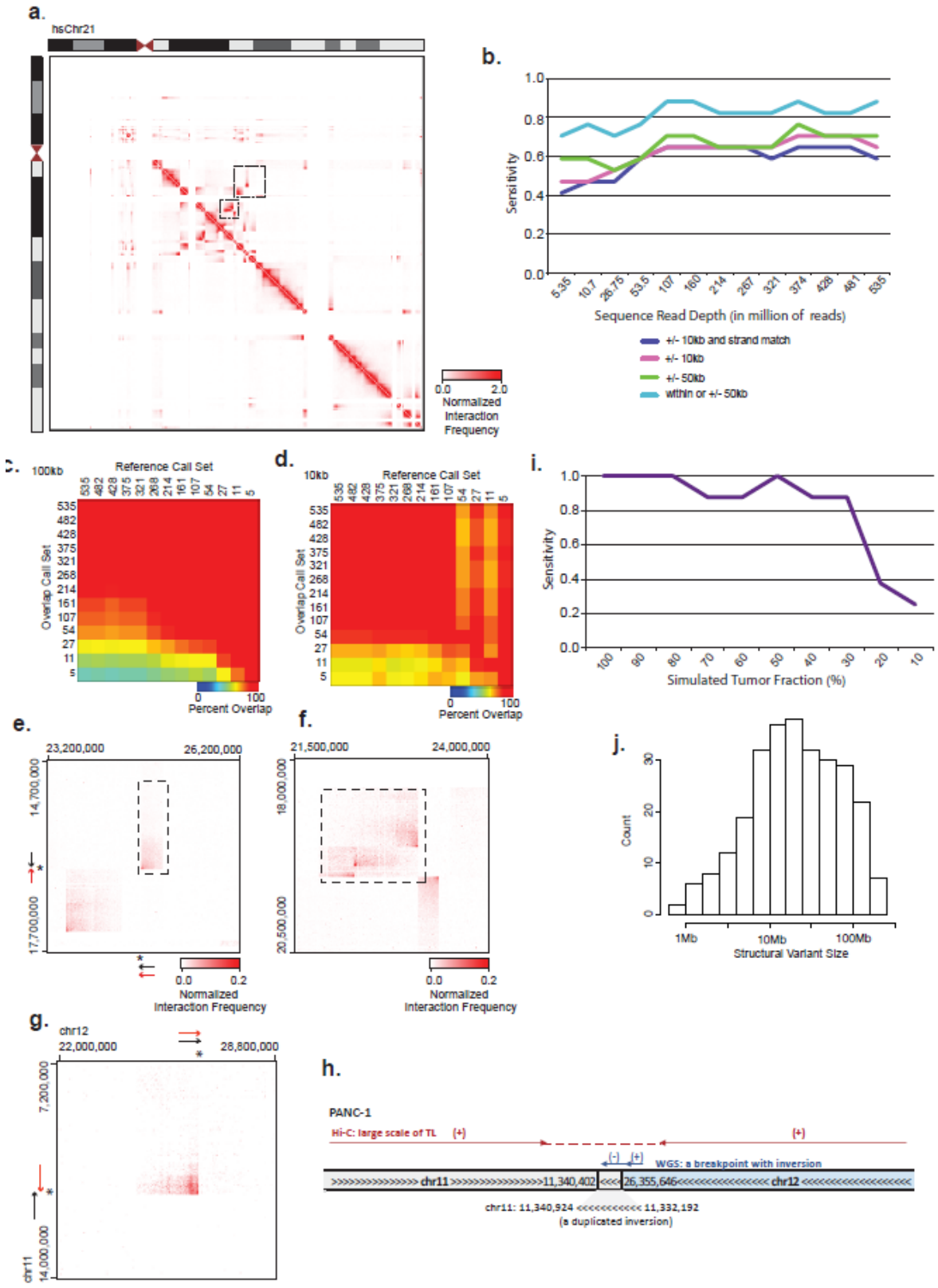

a.
hsChr21
0.0
2.0
Normalized
Interaction
Frequency
b.
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
5.35
10.7
26.75
53.5
107
160
214
267
321
374
428
481
535
Sequence Read Depth (in million of reads)
+/- 10kb and strand match
+/- 10kb
+/- 50kb
within or +/- 50kb
c.
100kb
Reference Call Set
Overlap Call Set
535
482
428
375
321
268
214
161
107
54
27
11
5
0
100
Percent Overlap
d.
10kb
Reference Call Set
Overlap Call Set
0
100
Percent Overlap
i.
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
100
90
80
70
60
50
40
30
20
10
Simulated Tumor Fraction (%)
e.
23,200,000
26,200,000
14,700,000
17,700,000
0.0
0.2
Normalized
Interaction Frequency
f.
21,500,000
24,000,000
18,000,000
20,500,000
0.0
0.2
Normalized
Interaction Frequency
j.
Count
30
20
10
0
1Mb
10Mb
100Mb
Structural Variant Size
g.
chr12
22,000,000
28,800,000
7,200,000
14,000,000
chr11
h.
PANC-1
Hi-C: large scale of TL
(+)
(+)
(-)
(+)
WGS: a breakpoint with inversion
chr11
11,340,402
26,355,646
chr12
chr11: 11,340,924
11,332,192
(a duplicated inversion)

Figure 2-S 5. Sensitivity and internal reproducibility of rearrangements identified by Hi-C.

a. Hi-C data from human chromosome 21 Tc1 cells. Dotted lines indicate regions shown in panels e and f. **b.** Sensitivity of Hi-C to detect gold standard SV calls at different sequencing depths. SVs were considered as matched if within 10kb with identical strandedness (purple line), within 10kb (pink), within 50kb (green), and within 50kb or internal to the rearrangement region (light blue). **c, d.** Internal consistency of sub-sampled calls at a resolution of 100kb (c) and 10kb (d). The number of reads sub-sampled reads is shown on the axes. **e.** Example of an SV where the breakpoint site (*) matches but the strandedness does not. Hi-C strandedness is "+/-", while gold standard is "-/-" (red arrows).**f.** Example of a region where Hi-C merged multiple rearrangements together. **g.** Example of an SV with strand discrepancy between Hi-C and WGS in Panc-1 cells (breakpoint marked with an asterisk). Hi-C indicates strandedness as +/+ (red arrows), while WGS indicates -/+ (black arrows). **h.** Diagram of the breakpoint shown in panel g. WGS identifies a small inversion (8kb) on chromosome 11 near the translocation breakpoint, such that the breakpoint lies within the inverted region. As a result, the global structure of the translocation is "+/+" (consistent with Hi-C), while the exact fusion is "-/+". **i.** Sensitivity to detect SVs using K562 (tumor) and GM12878 (normal) Hi-C data mixed at various fractions. **j.** Histogram of SV sizes detected by Hi-C.

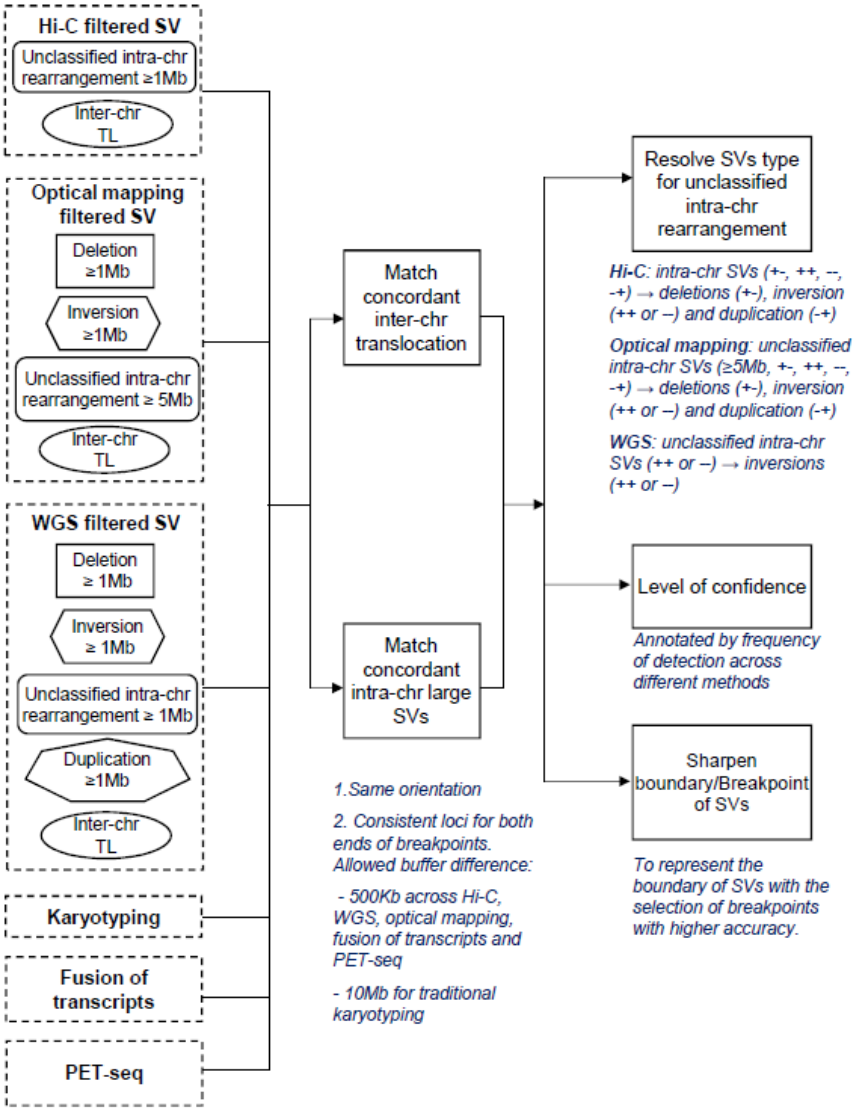


Figure 2-S 6. Comparison and integration of inter-chromosomal translocations and large intra-chromosomal SVs (≥1Mb).

We compared the SV calls by Hi-C, optical mapping and WGS, and we also included SV calls from additional methods, including karyotyping, fusion transcripts, and paired-end tag sequencing (PET-seq) when available from the same cell line. For the comparison, we first converted the strand orientation for SVs detected from different methods to a unified system, in which "+" indicates the breakpoint locates at the 3' end of the joined arm, and "-" indicates the breakpoint at the 5' end of the joined arm. For WGS data, this dictates that SV originally classified as deletions are given the strand orientation of "+-, inversions as "++ and - -", duplications as "-+" and unclassified intra-chromosomal rearrangement as "++" or "- -". Optical mapping originally reports deletions, which are assigned a strand orientation of "+-", inversions as "++" or "- -", and also intra-chromosomal rearrangements >5Mb as "unclassified intra-chromosomal rearrangements" for which the software reports the strand orientation. The same SV from distinct methods is considered a match when they have the same orientation and loci for both ends of breakpoint. The confidence level for each SV is represented by the times that the SV is independently reported by different methods. Further, the breakpoint/boundary of each SV is sharpened by choosing loci determined by the highest-resolution method. Finally, unclassified intra-chromosomal variants from WGS or optical mapping can be re-classified if resolved by an alternative method.

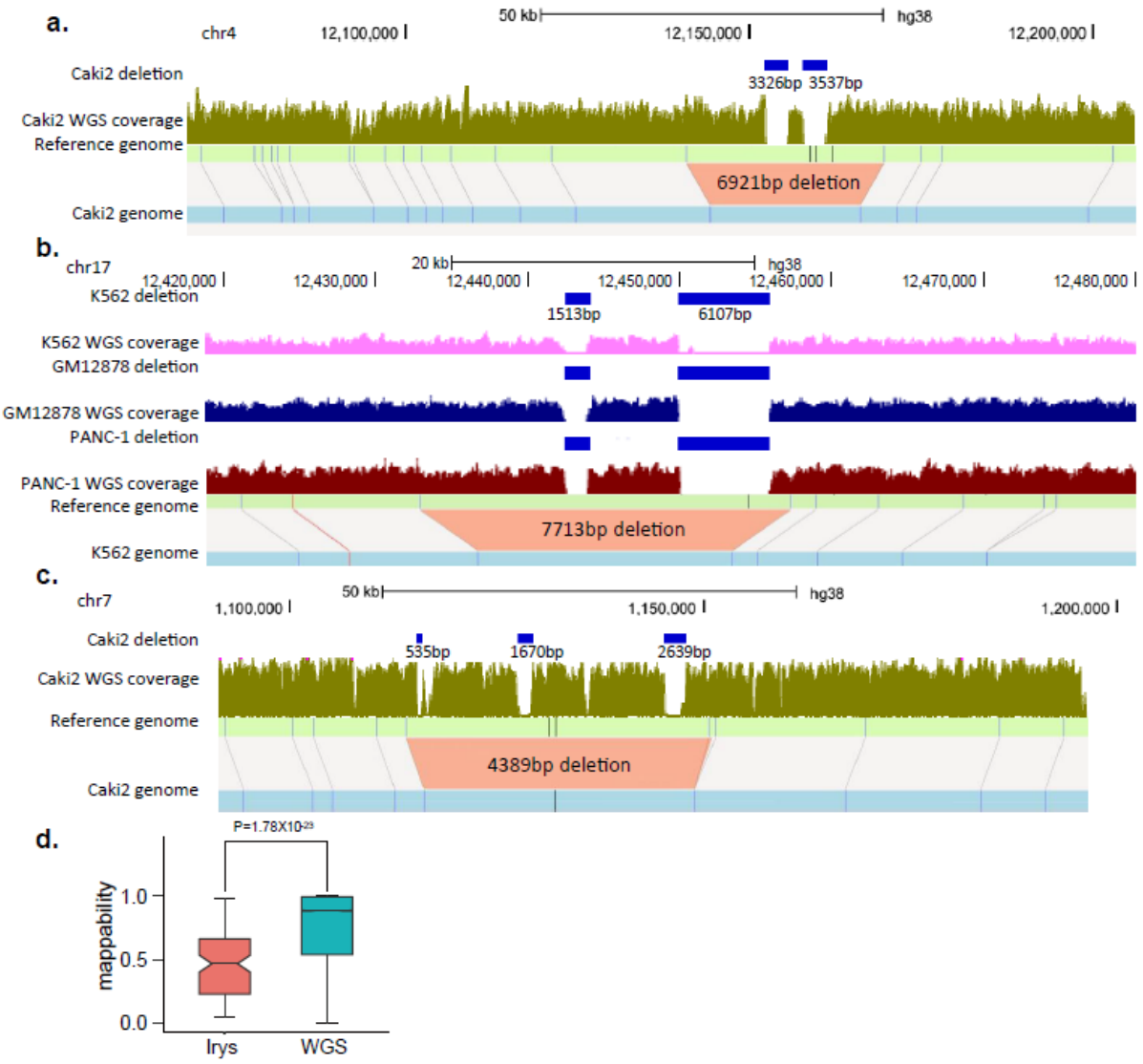


Figure 2-S 7. Deletions predicted by Irys overlap with multiple smaller WGS predicted deletions.

a. Optical mapping detects a 6,921 bp deletion within chr4: 12,140,782-12,169,591 in Caki2 cells. In the same region, there are two deletions reported by WGS (*Del1*: 12152,224-12,155,550, *Del2*: 12,157,718-12,161,255). The sum of their sizes is 6,863 bp, which is similar to that of the Irys predicted deletion. **b**. Similar as in **a**, optical mapping detected a shared polymorphic deletion of 7,713bp within chr17:12,432,762-12,457,176 in K562, GM12878 and PANC-1 cells. Again, this deletion can be supported by two smaller deletions detected by WGS (*Del1:* 12,442,344-12,443,887, *Del2*:12,449,829-12,455,936), whose summed size is 7,650bp. **c.** An Irys-detected 4,389bp deletion within chr7:1113898-1151045 in Caki2 cells overlaps with three WGS-detected deletions (*Del1*:1,115,577-1,116,112, *Del2*:1,127,730-1,129,400, *Del3*:1,145,442-1,148,018), whose summed size is 4,781 bp. **d.** Deletions detected by Irys have overall lower mappability compared to deletions detected by WGS (by two sided Wilcoxon rank-sum test). For WGS deletions, we computed the average of mappability scores for the 500bp regions upstream and downstream of the deletions (immediately outside the two breakpoints, n=26,255). For Irys-detected deletions, we computed the average mappability score between the two nicking enzymes (labels). We also require the size of deletions to count for at least 80% of the genomic distance between the two labels (n=103). For boxplots, the box represents the interquartile range (IQR), and the whiskers extend to 1.5 times the IQR or to the maximum/minimum if less than 1.5x IQR.

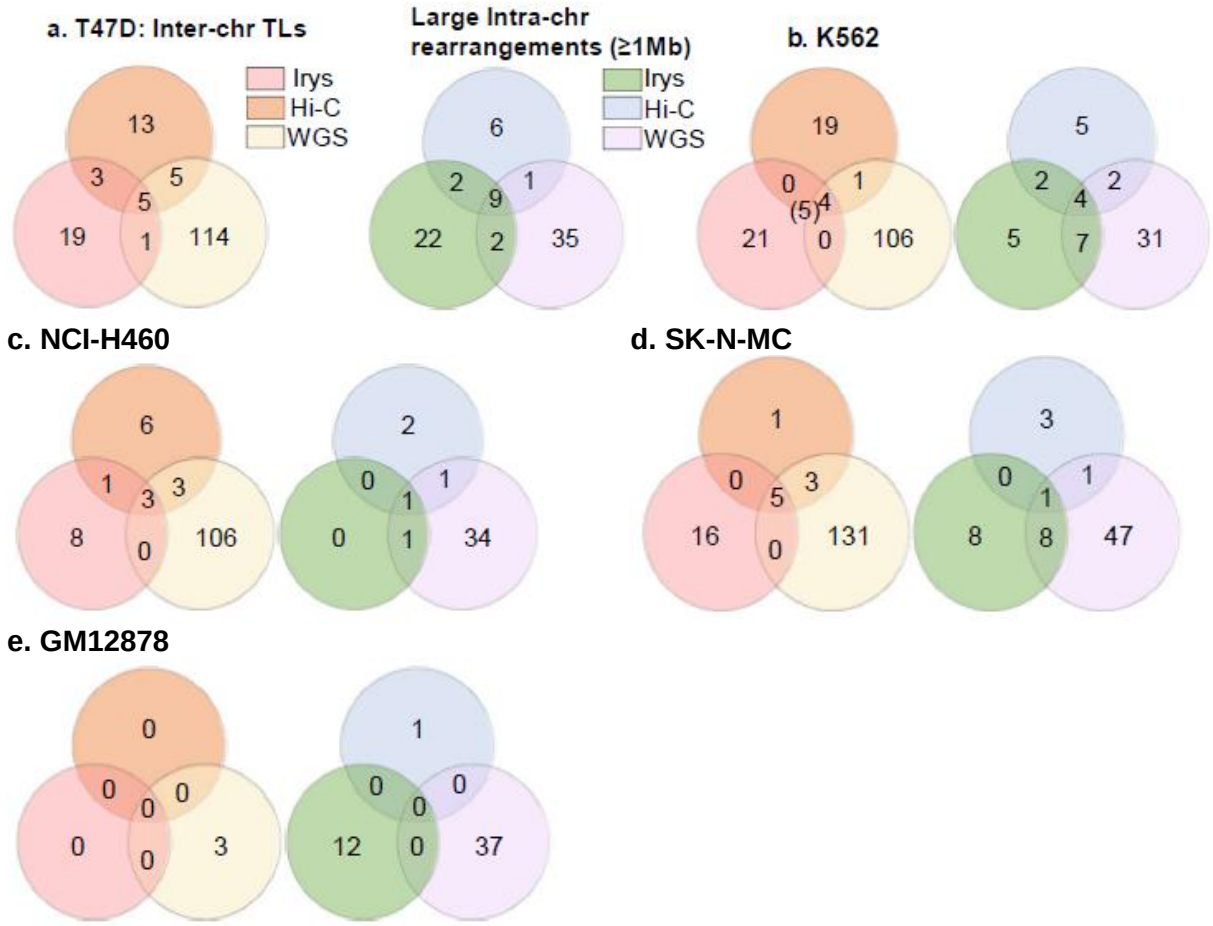


Figure 2-S 8. Overlap of large SVs detected by Hi-C, optical mapping, and WGS.

Number of inter-chromosomal translocations (left panel) and large intra-chromosomal rearrangements (≥1Mb, right panel) detected by optical mapping, Hi-C, and WGS in T47D (a), K562 (b), NCI-H460 (c), SK-N-MC (d), and GM12878 (e).

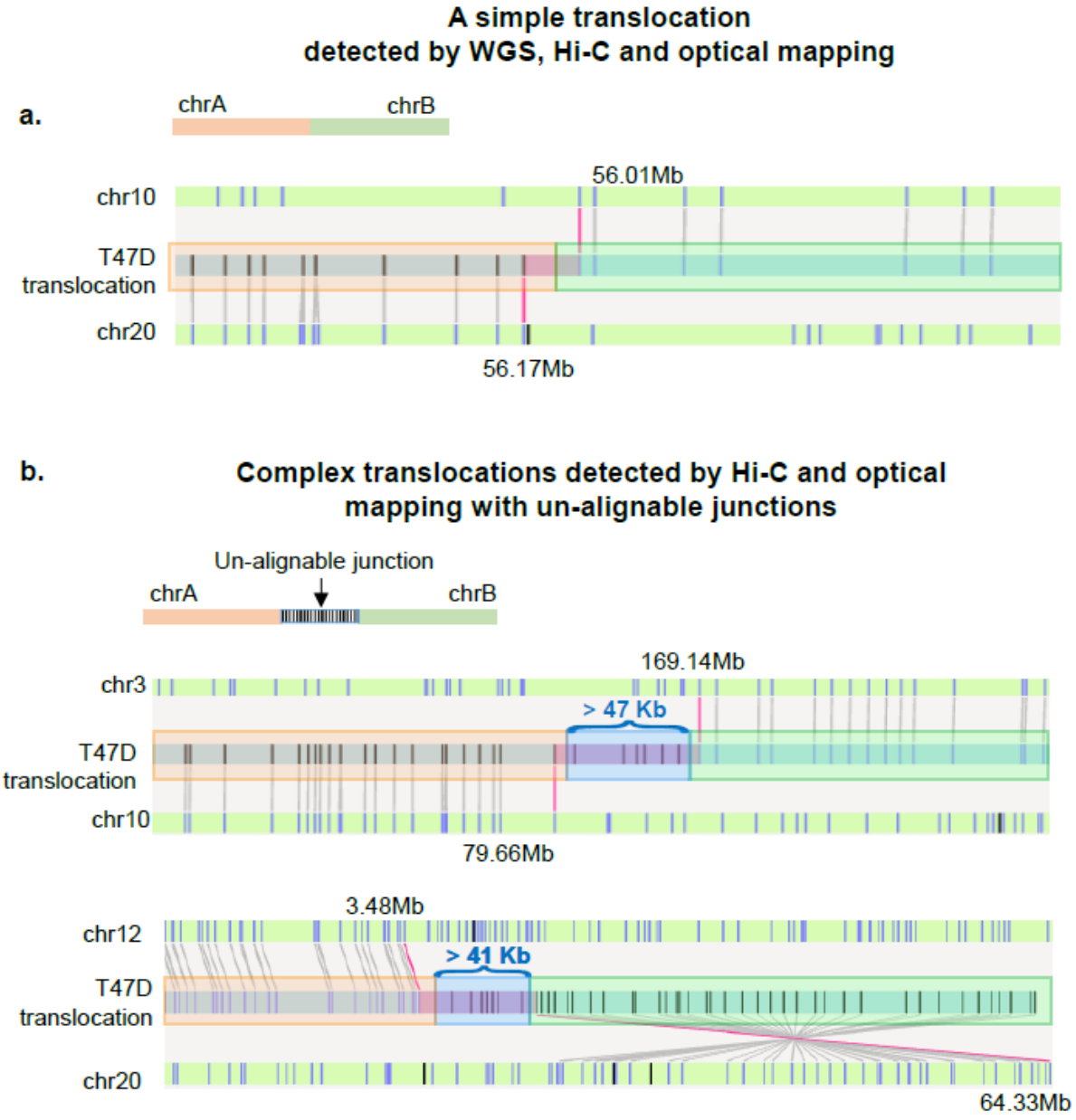


Figure 2-S 9. Hi-C and optical mapping detect translocations with unalignable junctions.

a. An example of a simple translocation detected by WGS, Hi-C, and optical mapping. The predicted breakpoint is located between the two labels (nicking enzymes) and there is no unalignable region between them.
b. Two examples of complex translocations with unalignable junctions detected by Hi-C and Irys but missed by WGS. In both scenarios, the large DNA fragments (> 40kb) between the two translocated arms were not mapped to human reference genome.

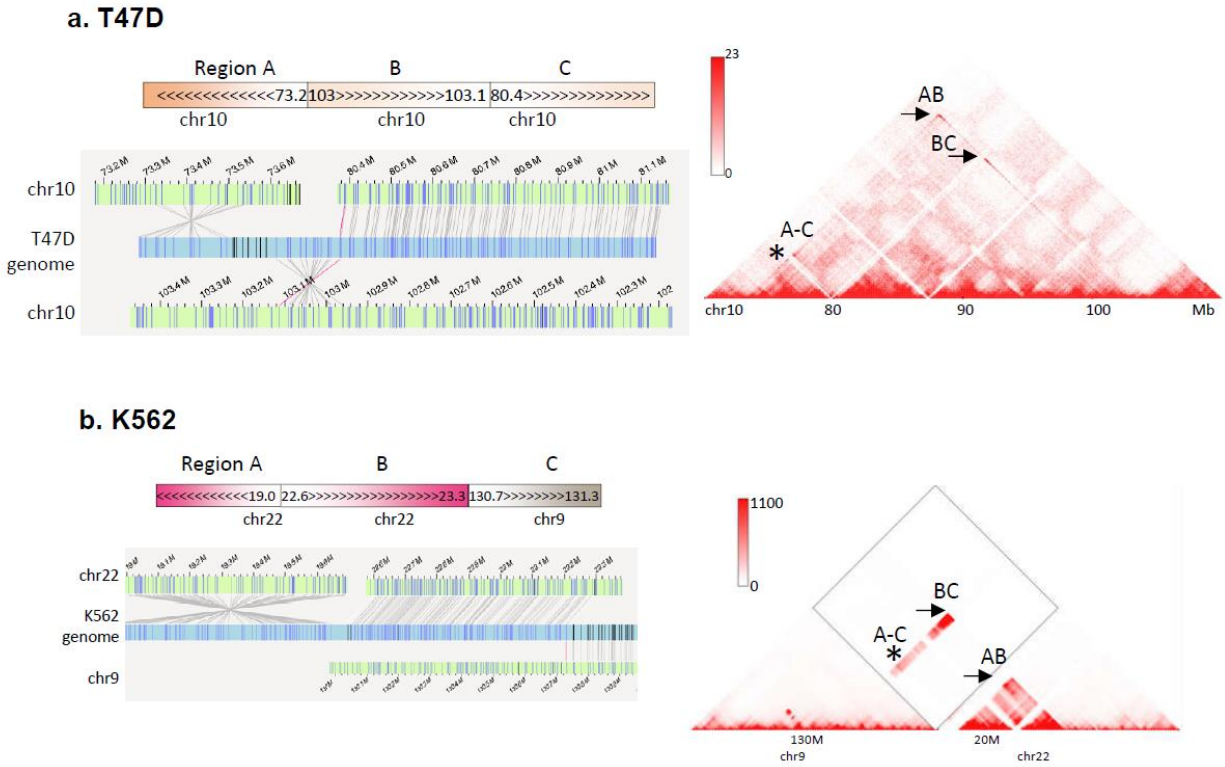


Figure 2-S 10. Examples of using Hi-C and optical mapping to reconstruct the overall structure of complex translocations.

Similar to Figure 2-3d, Arrow (->) indicates directly jointed translocation and asterisk (*) marks the linked adjacent SV. **a.** Schematic of the local chromosome structure in T47D cells, which consist of 3 translocated regions: A (chr10:73.5-73.5M), B (chr10:80.4M-81.1M), and C (103-103.1M).
b. Another example of locally resolved SV in LNCaP cell line. A ~8mb region on chr 7 (A) is inversely inserted between regions B and C on chr14.

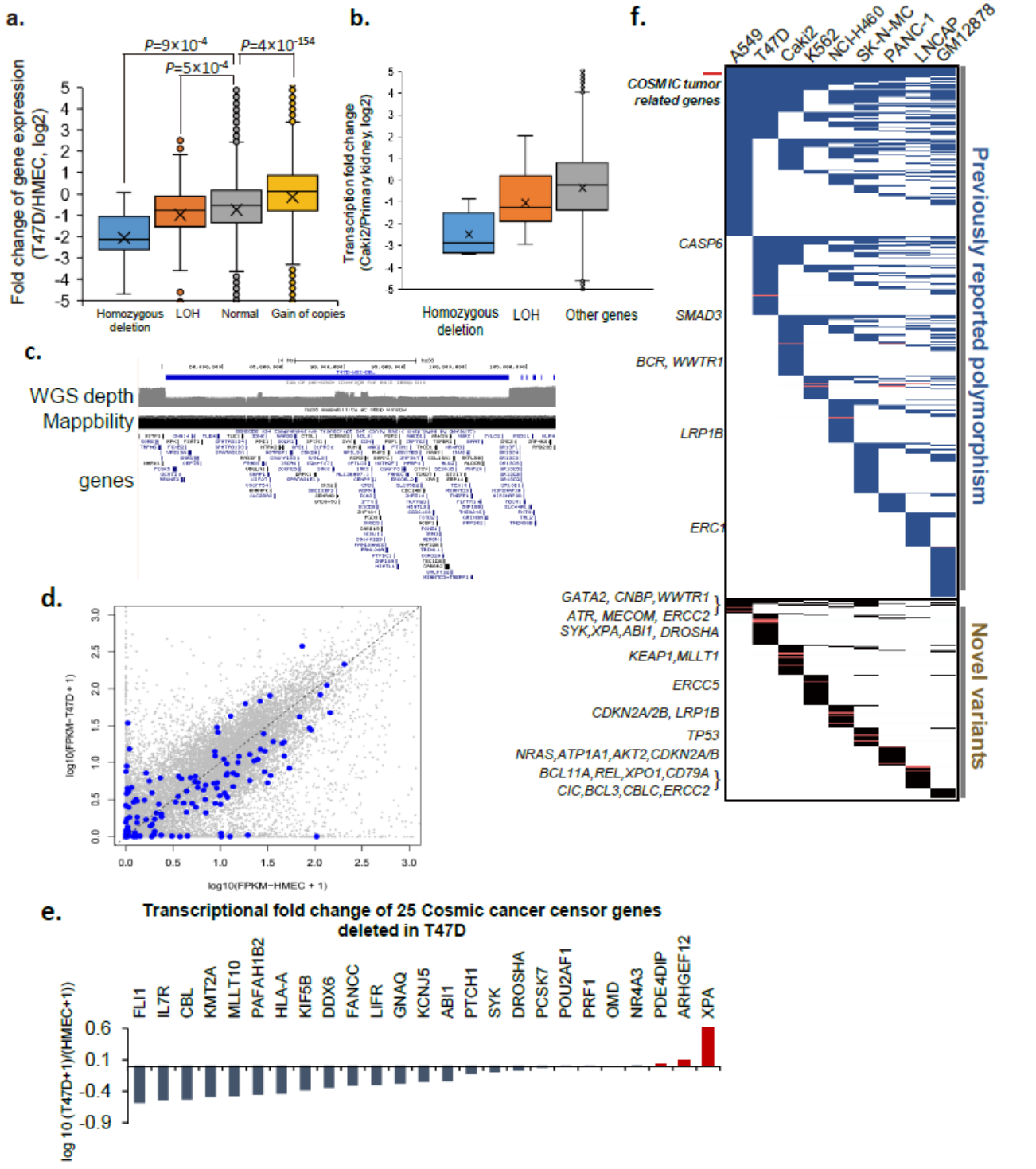


Figure 2-S 11. Impact of exon deletion and copy loss on gene expression.

a. Compared with HMEC cells, expressed genes (FPKM>1 in HMEC cells) with homozygous deletions (n=10) and LOH (n=325) in T47D cells show reduced expression compared to copy-neutral genes (n=5113, P = 0.009 and 0.003 respectively, two-sided Wilcoxon rank sum test), and compared to gain of copy genes (n=6413,

$p=4\times10^{-79}$, two sided Wilcoxon rank sum test). For all boxplots in the Figure the box represents the interquartile range (IQR), and the whiskers extend to 1.5 times the IQR or to the maximum/minimum if less than 1.5x IQR. **b.** Expressed genes (FPKM>1 in primary kidney epithelium cells) with homozygous deletions (n=5) or LOH (n=28) in Caki2 show reduced expression relative to non-copy number reduced genes (n=13859). **c.** A 28Mb deletion (chr9:75,335,996-103,526,867) in T47D cells causing LOH of over 400 genes. **d.** Deleted genes in T47D show reduced transcription. **e.** 25 COSMIC tumor-related genes have deletions overlapping with exons and the majority show reduced transcription. **f**. Cancer-specific novel deletions are enriched in COSMIC cancer-related genes. High-confidence deletions are classified as either known polymorphisms (from DGV database) or novel variants. In karyotypically normal cells (GM12878), 95% of deletions are polymorphic and 5% are novel, while in cancer genomes, over 10% of the deletions are novel. Novel deletions in cancer genomes are enriched for tumor related genes annotated by COSMIC database (red bars in the heat map).

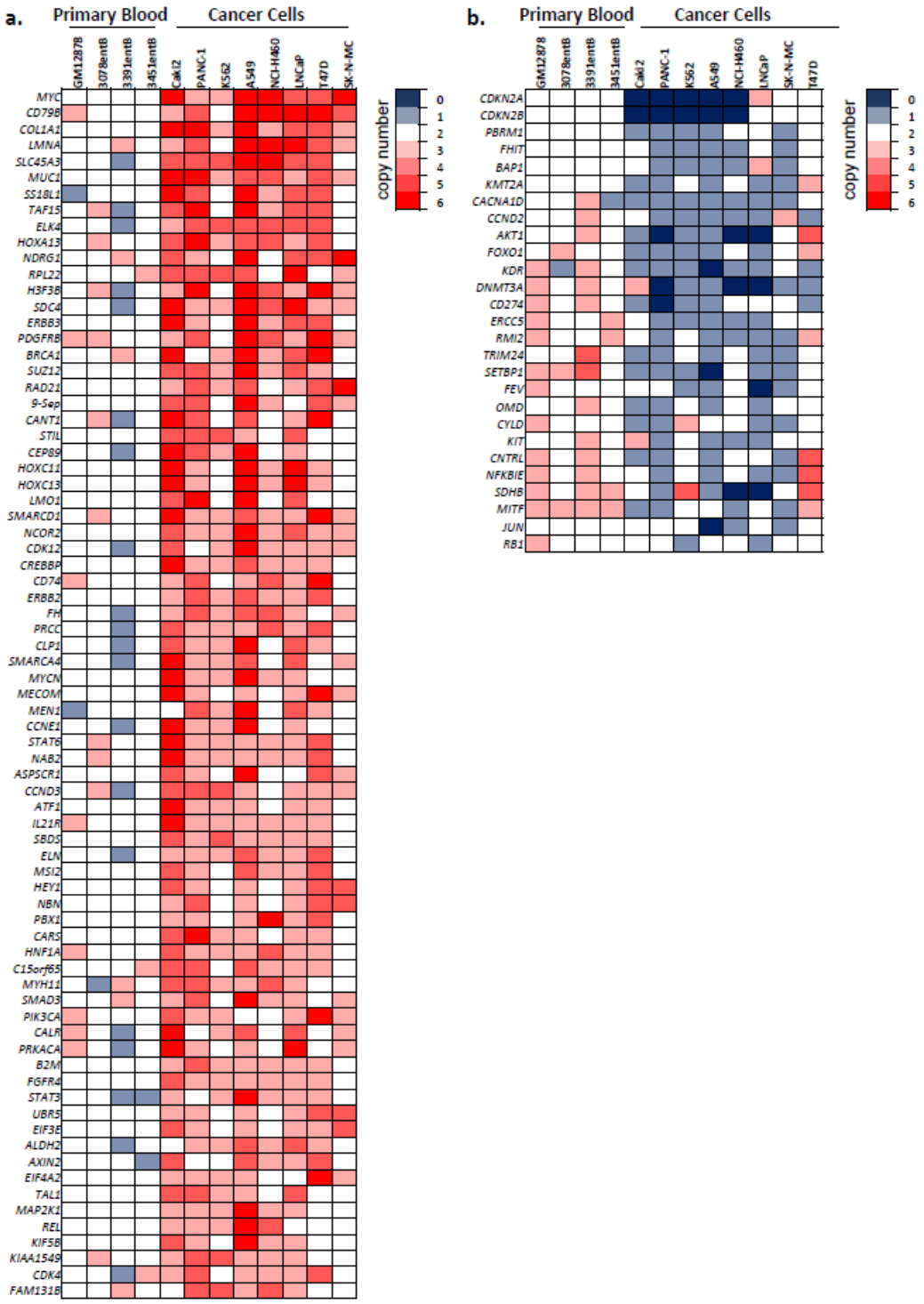


Figure 2-S 12. Copy number alterations of COSMIC tumor-related genes, which are computed based on its surrounding 50 kb regions by optical mapping.

a. COSMIC tumor-related genes with extensive gain of copies in cancer cell lines. **b.** COSMIC tumor-related genes with significant loss of copies in cancer cell lines.

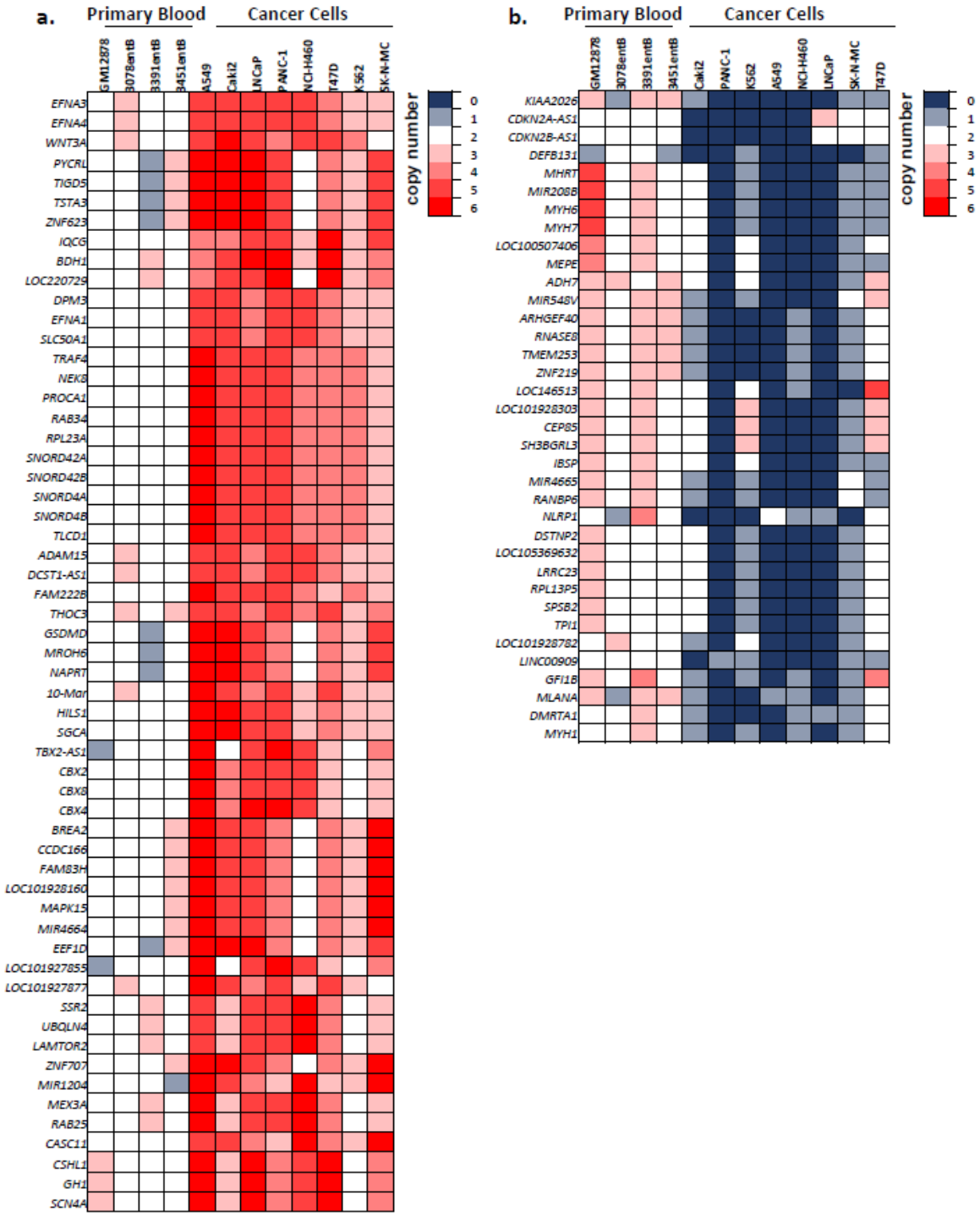


Figure 2-S 13. List of non-COSMIC tumor-related genes that have significant copy number changes.

Copy number is computed based on the surrounding 50Kb regions by optical mapping**. a.** 58 Genes with most significant amplifications. **b.** 37 genes with most significant loss of copies.

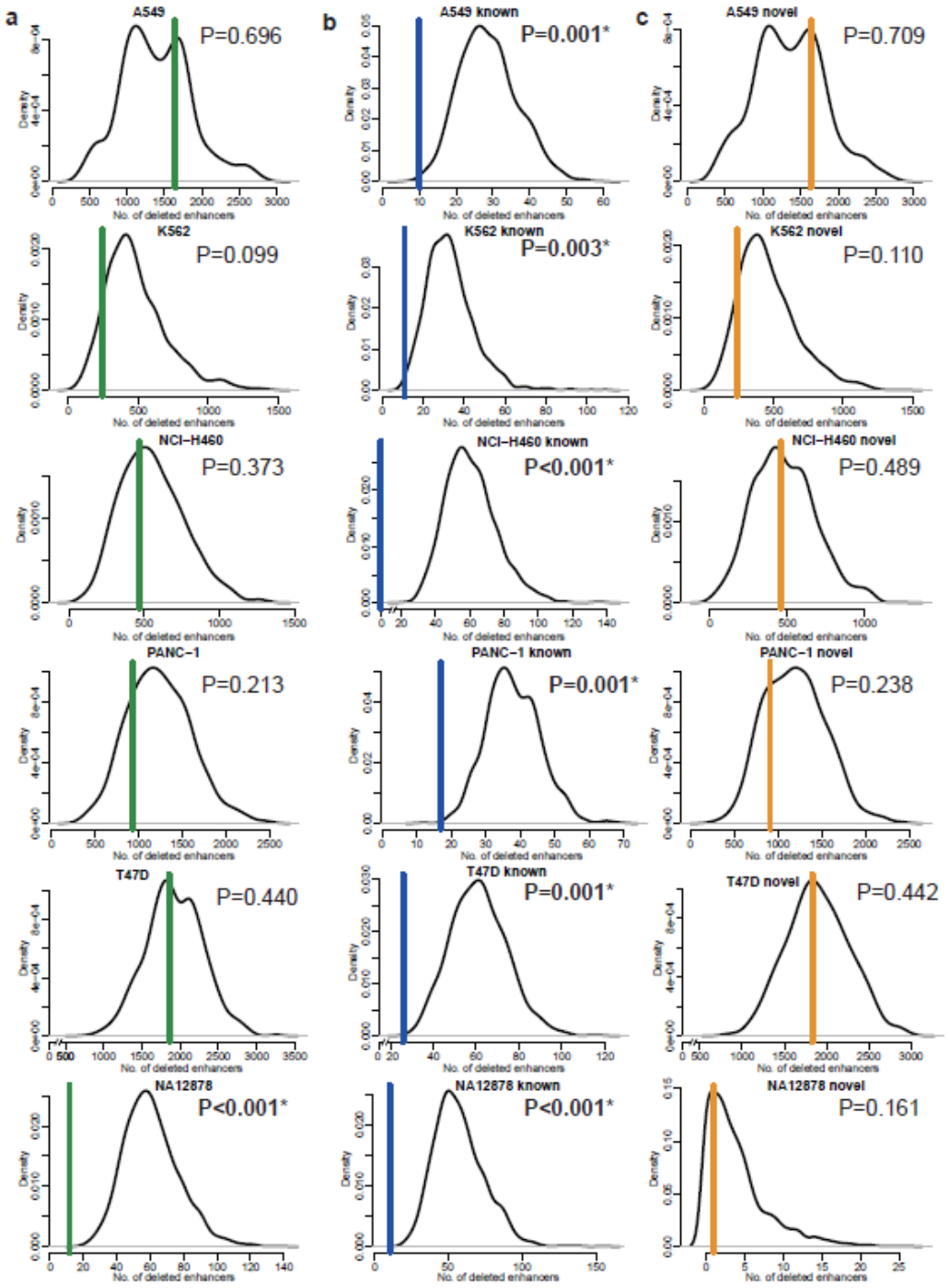


Figure 2-S 14. Comparison of the frequency of enhancer disruptions versus expectation.

a. Overall, we found that deletions in normal cell types (GM12878) are less likely to delete enhancer that would be expected at random, while the enrichment level of deletions in enhancers in cancer cells are close to the values expected at random. For this analysis, we matched each cancer cell line with a control normal cell type that is developmentally from the same/similar tissue type: T47D vs. HMEC, K562 vs. mononuclear cells, PANC-1 vs. primary pancreatic tissues, A549 and NCI-H460 vs. NHLF cells. We used the H3K27ac peaks in the normal

cell/tissue type as enhancer set. Then, we randomly shuffled the deletions in the cancer genomes 1,000 times and overlapped them with the enhancer set to compute the expected value (number of deletions: A549=237, K562=435, NCI-H460=405, PANC-1=320, T47D=454, NA12878=535). The curve shows the distribution of simulated results and the vertical line shows the observed value. The empirical P value is then calculated based on how many times the simulated number is smaller than the observed value (P<0.001 means no such incidence was observed in the 1000 simulations). **b,c.** We stratified the deletions into two categories by comparing them with DGV database: polymorphic deletions (A549=223, K562=392, NCI-H460=372, PANC-1=289, T47D=411, NA12878=513) and novel deletions (A549=14, K562=43, NCI-dH460=33, PANC-1=31, T47D=43, NA12878=22). We found that polymorphic deletions are less likely to delete enhancer, while novel deletions are reflect the genome wide distribution of enhancers.

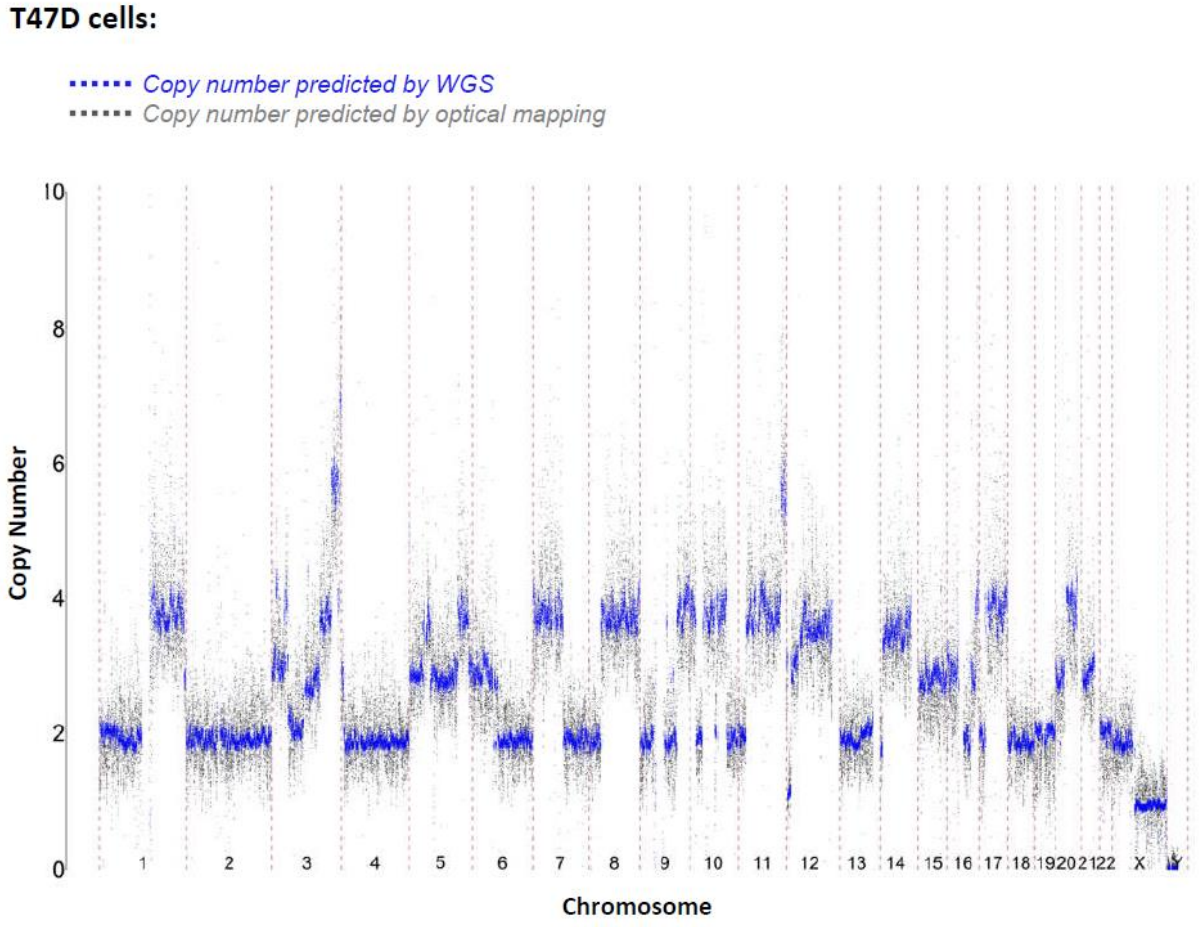


Figure 2-S 15. Genome-wide CNVs predicted by optical mapping and WGS are consistent.

Materials and Methods

Materials and Experiments

Cell culture

K562 cells (ATCC CCL-243) were cultured in Iscove's Modified Dulbecco's Medium supplemented with 10% FBS and antibiotics. T47D cells (ATCC HTB-133), NCI-H460 cells (ATCC HTB-177), A549 cells (ATCC CCL-185), LNCaP (ATCC CRL-1740), and GM12878 cells (Coriell) were cultured in RPMI-1640 supplemented with 10% FBS and antibiotics, or 15% FBS and antibiotics (GM12878). Caki2 cells (ATCC HTB-47), G-401 cells (ATCC CRL-1441) were cultured in McCoy's 5a Medium Modified supplemented with 10% FBS and antibiotics. PANC-1 cells (ATCC CRL-1469) were cultured in Dulbecco's Modified Eagle's Medium supplemented with 10% FBS and antibiotics. SK-N-MC (ATCC HTB-10), RPMI-7951 (ATCC HTB-66) cells were cultured in Eagle's Minimum Essential Medium supplemented with 10% FBS and antibiotics. SK-N-AS cells (ATCC CRL-2137) were cultured in Dulbecco's Modified Eagle's Medium supplemented with 10% FBS, 0.1mM Non-Essential Amino Acids (Gibco) and antibiotics All cell lines cultured as part of ENCODE data generation (A549, Caki2, G401, LNCaP, NCI-H460, Panc1, RPMI-7951, SJCRH30, SK-MEL-5, SK-N-DZ, SK-N-MC,T47D) were cultured using standardized protocols, the details of which can be found through the ENCODE consortium website (https://www.encodeproject.org/).

Optical mapping experiments

10 million cells of T47D, Caki2, K562, SK-N-MC, A549, NCI-H460, PANC-1, and LNCaP were pelleted and then washed three times with PBS. Cells equivalent to 600ng of DNA were embedded in 2% Agarose (Bio-rad), solidified at 4°C for 45 minutes. Cells within plugs are lysed in 2ml cell lysis

buffer (BioNano Genomics) containing 167ul proteinase K (Qiagen) for 48 hours, and washed twice with Tris-EDTA, pH 8 (TE) for 15 minutes per wash. DNA plugs were purified with 2ml 5% RNAase (Qiagen) for two hours, washed in TE for 15 minutes × 6 times, melted and equilibrated on 43℃ for 45 minutes with 2ul of GELase (Epicentre). DNA was transferred onto a membrane floating in TE and concentrated by dialysis for 135 minutes. DNA was then equilibrated at room temperature overnight. 900ng DNA was digested by 30U nicking enzyme BspQ1 (New England Biolabs) in 1× buffer 3 (BioNano Genomic), 37 ℃ for 4 hours, and labeled with 1× labeling mix (BioNano Genomics) and 15U Taq polymerase (New England Biolabs) in 1× labeling buffer (BioNano Genomics) at 72℃ for 60 minutes. Nick-labeled DNA was repaired in 1X repair mix (BioNano Genomics), 1× Thermo polymerase buffer (NEB), 50uM NAD+ (New England Biolabs), and 3ul 120U Taq DNA ligase (New England Biolabs) at 37℃ for 30 minutes. DNA staining was finally performed with the final solution containing 1× flow buffer, 1× DTT (BioNano Genomics), and 3ul DNA stain (BioNano Genomics), in room temperature overnight. Optical mapping data collection: Each sample underwent in average 7 rounds of data collection on BioNano Irys platform to reach 100X reference coverage. For each round, 160ng prepared DNA was loaded to a BioNano Irys chip that contains two flow-cells, and each r. Hi-C experiments and sequence read alignment

Whole genome sequencing

1ug DNA was respectively collected and purified from each sequenced samples using DNeasy Blood & Tissue kit (Qiagen), including T47D, Caki2, K562, NCI-H460, SK-N-MC, and PANC-1. The DNA library was further prepared according to the Illumina TruSeq DNA PCR-free library preparation guide. DNA was fragmented by covaris system into 300-400bp dsDNA with 3’ or 5’ overhangs, repaired to blunt end and selected by size. DNA was then adenylated at the 3’ end, indexed by adapters

ligation, and validated by quality control. 150bp Paired-end sequencing was performed to reach in average 30X of genome coverage on platform HiSeq XTen.

Hi-C

Hi-C in K562 and SK-N-AS cells was performed using the in situ Hi-C protocol [78] from 5 million cells using the MboI enzyme. Hi-C experiments in all ENCODE cells lines was performed using the original Hi-C protocol using the HindIII enzyme [79]. Hi-C experiments were performed as biological replicates to ensure experimental reproducibility. Hi-C libraries were sequenced using Illumina HiSeq 2000 and HiSeq 2500 sequencing machines and processed to FASTQ files using standard processing pipelines. Read pairs were aligned independently using BWA-MEM to a custom GRCh38 genome assembly. The base for this assembly is available through the 1000 genomes consortium ound contains 30 cycles of data collection.

Breakpoint PCR

PCR across predicted breakpoints was performed using the Qiagen Long-Range PCR kit. PCR products amplified from K562 template were cloned into TOPO-XL cloning vectors and sequenced using conventional Sanger sequencing. In the event that the breakpoint did not fall within the Sanger sequenced regions, primers were re-designed and the process was repeated. Primers are as below:

Cell	SV type	Name	Sequence
K562	Translocation	K_chr9_22_F	AAAGAGCCTTTTGTTGGCTATGTTGTT
K562	Translocation	K_chr9_22_R	CAGAAGGAAGAGCTATGCTTGTTAGGG
K562	Translocation	K_chr3_10_F	CTGCCATAAAGAGTTCACAAACACACC
K562	Translocation	K_chr3_10_R	CTGAGACCTGGAAAACAGAGCAAGAC
K562	Translocation	K_chr5_6_F	AGCAATTTTAGAGGCACTTCTCCTTGT
K562	Translocation	K_chr5_6_R	AGGCATTTGGGATCTTGCTGGATTATG
K562	Translocation	K_chr9_13_F	TTGAGATGTCTGTTTCATTTCCCGACT
K562	Translocation	K_chr9_13_R	GAACCACTGCTCCTGGACTTCATCTT
T47D	Translocation	T_[chr6_chr22]_F	CACATAACCAAGGGAGAGTT
T47D	Translocation	T_[chr6_chr22]_R	GTGAGGTGAATTCAAATGTT

T47D	Translocation	T_chr4]_chr5]_F	TTGCACACCGGCTCCATGAG
T47D	Translocation	T_chr4]_chr5]_R	GATCTCTACTTAATCTGCAT
T47D	Translocation	T_9]_[15_F	TAAAAGATAAAGGCATCTGT
T47D	Translocation	T_9]_[15_R	ACCAACCAAAAAAAGCCCAG
T47D	Translocation	T_5]_[5_F	CTTCCCGTCTAAGCAGACCT
T47D	Translocation	T_5]_[5_R	CTTTCATCATGTTAGTCATG
T47D	Translocation	T_9]_[9_F	GGTTTGGGCATTCTATTTTC
T47D	Translocation	T_9]_[9_R	GCCTTCAGAAAGTTCTCAGT
T47D	Translocation	T_chr10]_[chr10_F	ATATAAATGCGATGCTTTTTCCT
T47D	Translocation	T_chr10]_[chr10_R	GAGTTGTTTTGAGTTCCTTGGAG
T47D	Translocation	T_chr10]_[chr3_F	GCAAAGTTCTTCTTAAGAATGT
T47D	Translocation	T_chr10]_[chr3_R	ACAGATTAATTGACTCCCTTC
T47D	Translocation	T_chr3]_[chr9_F	GTGCTAGGATTACAGGAATGAGC
T47D	Translocation	T_chr3]_[chr9_R	GGAAACCCTTGTACACTATTGGT
Caki2	Translocation	C_chr12]_[chr4_F	TTCCCTTTAAAAGCACAATGCCC
Caki2	Translocation	C_chr12]_[chr4_R	ATTTCCTATAATTGGGTTTTCCT
Caki2	Translocation	C_chr9]_[chr19_F	AGTCAGTCTTGTACCTTGGGATG
Caki2	Translocation	C_chr9]_[chr19_R	AGAAAGCTTCCAGTCACAAAACT
Caki2	Translocation	C_ [chr6_[chr8_F	GGTATGGAGATGATCAACCCAAG
Caki2	Translocation	C_ [chr6_[chr8_R	TTGACAAAAGAATAAACAAATAGAT
T47D	Deletion	T_chr2_212590110_F	GTGGGATAAACAAGTGACTAACC
T47D	Deletion	T_chr2_212720073_R	ACCACGAAGCCACCAGAAGGAAG
T47D	Deletion	T_chr2_97188517_F	AATTAACTCCTAAAATGGTAATT
T47D	Deletion	T_chr2_97190465_R	ATCAATGTGGATATGCCGAGTGA
T47D	Deletion	T_chr14_104948976_F	GCATCTGCAGCTTGGGCAGGTGC
T47D	Deletion	T_chr14_104951429_R	AAAGTGGACCTCAAGGGCCCCCA
T47D	Deletion	T_chr3_58586154_F	TTTCCTGAATAGAAAAGAAACAC
T47D	Deletion	T_chr3_58586217_R	CAATCCTCACGTCATTCTTTTTA
T47D	Deletion	T_chr4_165081464_F	CCACCTAGGAACCTCCCACTCTT
T47D	Deletion	T_chr4_165083902_R	GAAAAAAACATGACTGGGCGCGG
T47D	Deletion	T_chrX_42652746_F	CCACTGCAAAAACATGCCAA
T47D	Deletion	T_chrX_42656304_R	AGTTTTCAAAGGGAATGCTT
T47D	Deletion	T_chr2_28466613_F	AATTATAAAAGTATCATGGG
T47D	Deletion	T_chr2_28469693_R	CCAGGCAAATCAGAGGTGTC
T47D	Deletion	T_chr7_6861596_F	CTTTACTGGTGTTGGACTCG
T47D	Deletion	T_chr7_6887316_R	ATTAAAGCAGTTGGATTTTT
T47D	Deletion	T_chr1_207523594_F	AAAAGCAATAGGACAAAGGC
T47D	Deletion	T_chr1_207546536_R	GCTCATCTCCTTTCAAGTCT
T47D	Deletion	T_chr12_58325913_F	TGAGTTCCCTTAGTATTTAT
T47D	Deletion	T_chr12_58339245_R	ATAGGTGGGGATTATGGGAG
T47D	Deletion	T_chr11_107361838_F	GAAGCCTCAGGAGCTGATGA
T47D	Deletion	T_chr11_107374676_R	GTCACCAATCTTGTCTTCCT
T47D	Deletion	T_chr7_97762466_F	ACTGGATCCCTTCCTTACAG
T47D	Deletion	T_chr7_97773481_R	GGCAAGCTGCTGAATTGCCT
T47D	Deletion	T_chr7_70969523_F	TGAGCCAATTAAACCTCTAT
T47D	Deletion	T_chr7_70979773_R	GTATTCATGCTTCAAAGAAG
T47D	Deletion	T_chr6_85998091_F	TGCAGTGTTTGGTTTTCTAT
T47D	Deletion	T_chr6_86007304_R	AAAAAGTGGGCAAAGGATAT

T47D	Deletion	T_chr1_53126296_F	GGACTACAGGTGCCCACCAT
T47D	Deletion	T_chr1_53129986_R	CCAGTGGTGGCTTCATCTGT
T47D	Deletion	T_chr13_69400712_F	CTACAGAAAGACTGAATAGC
T47D	Deletion	T_chr13_69404714_R	ATTATATTTGGGGAATCTAC

Informatics analysis

Detection of structural variants based on optical mapping

de novo assembly and SV detections

Cell line or sample-specific genomic maps are generated through *de novo* assembly of DNA optical reads using BioNano Refaligner 6119 and pipeline 6498. We require that DNA reads be no shorter than 150Kb with at least 9 labels per molecule, and the signal to noise ratio no less than 2.75, while the maximum backbone intensity is 0.6. The assembly pipeline was applied with the following parameters: iterations: 5; initial assembly P value threshold: 1e-11; extension and refinement P value threshold: 1e-11. De novo assembly noise are specifically: False positive density/100Kb:1.0; False negative rate:0.1; SiteSD:0.15; ScalingSD:0; RelativeSD: 0.03; ResolutionSD: 0.25.

SV detection is performed after the completion of de novo assembly by comparing assembled contigs to the GRCh38 reference genome GRCh38 using the built-in module *runSV*. All centromere regions are skipped during SV identification. Deletions, insertions and inversions are detected with the default settings using a p-value threshold of 1e-12. In the default output, any intra-chromosomal SVs larger than 5Mb are defined as “unclassified” intra-chromosomal rearrangements. Unclassified intra-chromosomal rearrangements and inter-chromosomal translocations are detected using a less stringent P-value threshold of 1e-8.

Filtration of detected SVs

Duplicates of SVs can be generated during SV detection from different contigs mapping to the same region, and such duplicate SV calls are merged into a single SV call. For deletions and insertions, we further remove small indels with a size smaller than 50bp. Many of the deletions we detect overlap with genomic gaps. This is most likely the result of overestimation of gap sizes. In this sense, these are not true deletions but instead assembly errors (or regions with polymorphic gap sizes). We classify deletions as gap errors if the deletion recurrently appears in different cell lines and at least 30% of the deletion overlaps with gaps, and at least 80% of the gap overlaps with the deletion. We remove these "gap errors" from the list of deletions and use them for gap size re-estimation analysis.

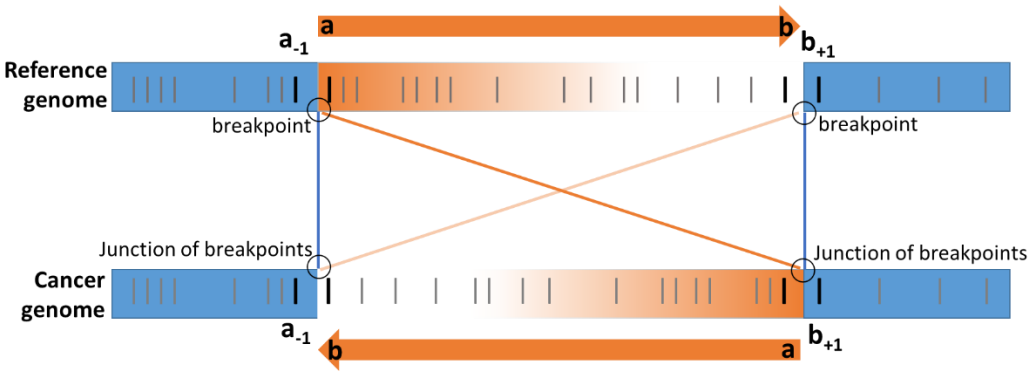
We also developed strategies to filter SVs in close proximity to the centromere. In peri-centromeric regions, we noticed that contigs can have ambiguous alignments to multiple regions due to redundant labeling patterns, which result in the appearance of deletions that cross the centromere. We therefore remove recurrent large deletions (80% reciprocal overlap, >1Mb) crossing centromeres. We further stratify deletions larger than 100kb into two categories, one where sequences within the deleted region show reduced mappability and one where the sequences are mappable. We then filter deletions over mappable regions that are not supported by a loss of coverage in WGS data. Deletions that are supported by valley of WGS coverage are annotated as "High confidence".

Defining inversions by Irys

A simple inversion involves two breakpoints and each breakpoint is represented by a pair of loci. **Figure** A below shows an example: the left breakpoint of this inversion occurs between nicking sites a_{-1} and a, and the right breakpoint occurs between nicking sites b and b_{+1}. The orange sequence in the middle is inverted and forms two breakpoint junctions: the left junction between (a_{-1} and b) and the right junction between (a and b_{+1}). We use the distance between sites a and b to approximate the size of this inversion (distance=b-a). To compare with inversions detected by other methods such as WGS and

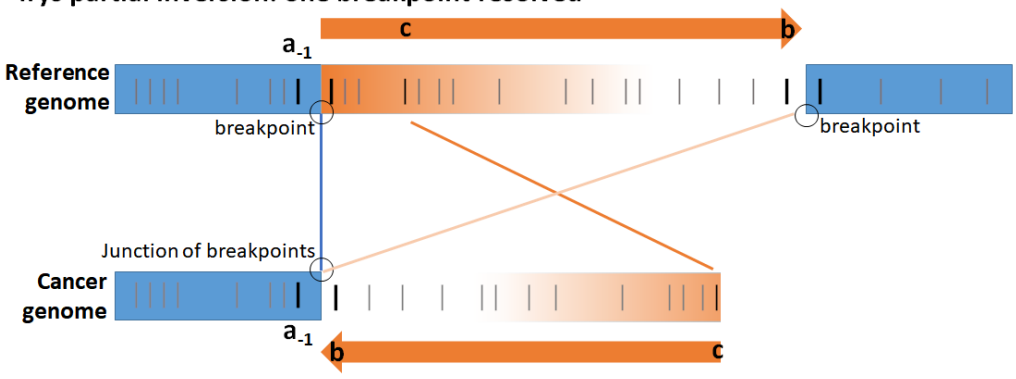
Hi-C, we used the junction of breakpoints (a_{-1}, b) and (a, b_{+1}). Such inversions with both junction of breakpoints resolved and four loci available are called “paired inversions”.

Figure A
Irys paired inversion: both breakpoints resolved



Due to technical limitations, Irys may also detect an incomplete inversion in cancer genomes. As shown below, at the left end, Irys detects the junction of breakpoints (a_{-1}, b) in the cancer genome, but at the other end, its contig stops at loci c and cannot reveal the real junction of other breakpoint. Such inversions with only one junction of breakpoint resolved are named “partial inversions” by Bionano Iry. In this scenario, we use the distance between loci b and c (b-c) to calculate the minimal size of this inversion. To compare with WGS and Hi-C, we only use the resolved breakpoint junction (a_{-1}, b). Therefore, the two columns of positions reported Supplementary **Table** 8 only represent the breakpoint junctions and cannot be used to estimate the size of inversion.

Figure B
Irys partial inversion: one breakpoint resolved



According to a recent study from Pendleton et al. [103], some inversions detected against hg19 were no longer detected against hg38, and they turned out to be inverted assembly of contigs in reference genome hg19, which were corrected in hg38. Those inversions have a unique feature that they are flanked by genomic gaps at each side. We scanned through inversions detected by our pipeline against hg38 and also found inversions flanked by gaps, which could represent inverted assembly of genome contig in hg38, or variations across populations that could be in both orientations in human genome. We thus remove such inversions to ensure we focus on genomic rearrangements and not genome assembly anomalies.

We observed that regions of the reference genome that harbor similar sequences distributed across multiple regions appear to harbor recurrent translocations in many samples. This is most likely due to misalignment of optical DNA reads leading to fixed false detection of translocations. A list of recurrent false-positive translocations was hence generated by comparing translocations detected across ten samples (with difference less than 1Mb away for both breakpoints), and the calls matching the list were removed. This list of recurrent translocation did not match any translocations detected by Hi-C or WGS, confirming that these are likely false positives.

Structural variant detection and filtration from whole genome sequencing

SV detection

Structural variants were detected by three independent pipelines. In the first pipeline, paired-end sequencing reads were first aligned by BWA-MEM (v0.7.15-r1140) to a GRCh38 human reference genome (version GCA000001405.015) with alternate haplotypes removed. Duplicate reads were removed by Picard. Reads with a mapping quality of at least 20 were retained for SV detection. SV calls were generated from this mapped data using Delly (v0.7.7) with default parameters (-q 20). Delly detects deletions, inversions, tandem duplications, insertions, and inter-chromosomal translocations.

In the second pipeline, paired-end reads were processed by the Speedseq framework. Paired-end reads were aligned to the GRCh38 reference genome using BWA-MEM in the same manner as the first pipeline. Duplicated reads are removed by SAMBLASTER (v0.1.24). Discordant and split reads were extracted by SAMBLASTER for SV detection. SV calls were generated using Lumpy (v0.2.13) with default parameters (speedseq sv -g -t 64 -x). Lumpy reports SVs as deletions, inversions, duplications, inter-chromosomal translocations, and unresolved break ends. In both pipelines, telomeric, centromeric, and 12 heterochromatic regions are masked for SV detection using blacklisted regions provided by the Delly software.

Copy number profiles were generated using Control-FREEC [61](v11.0). For all cell lines, we used a set of common parameters (ploidy = 2 for normal cells NA12878, pseudodiploid cells SK-N-MC, and hypotriploid cells T47D, A549, LNCaP, NCI-H460 and Caki2; ploidy = 3 for triploid cells K562 and hypertriploid cells PANC-1), breakPointThreshold = 0.8, coefficientOfVariation=0.062, mateOrientation = FR). For A549, Caki2, LNCAP, NCI-H460, and PANC-1, sex was set to “XY”; for K562, NA12878, SK-N-MC, and T47D, sex was set to “XX”. Predicted copy number for each 50,000bp bin was used for making Circos plots. Regions with copy loss (copy number equal to 0 or 1) that are not

captured by SV detection using Delly or Lumpy (by exclusion of those reciprocally overlap by at least 50% with deletions called by Delly and Lumpy) were included in the set of detected deletions.

SV filtration

To reduce false positive calls, the following filtration steps were applied for Delly and Lumpy SV calls. First, we require all SV calls to be supported by at least three split reads (SR) or three spanning paired-end reads (PE). Insertions or deletions less than 50 bp are removed, as are SV that map to chromosome Y or to the mitochondrial genome. SV calls from Delly and Lumpy are then merged, and only SVs that are identified by both methods are retained. We used separate criteria to call SVs overlapping between the two methods depending on the type of SV. For deletions, calls were merged between the two pipelines if they had an reciprocal overlap (RO) ≥ 50%. We used the coordinates provided by Lumpy for this merged deletion set. For inversions, calls identified by both Lumpy and Delly were merged if they had an RO ≥ 0.9. The final merged coordinates were based on the coordinates from the Lumpy calls. Translocations were merged between the two pipelines if the paired break ends mapped within +/- 50 bp of each other and if the strand of the break ends matched. The final coordinates were based on the calls from Lumpy. Regions annotated as insertions were identified by Delly alone, since Lumpy does does not annotate SVs as insertions. No specific filtration for insertion was applied.

Additional filtration was applied to specific types of SVs. For deletions, we removed deletions that have at least 50% reciprocal overlap (RO ≥ 50%) with known gap regions (+- 50bp), or at least 1bp overlap with centromere regions (+/- 1kb). Recurrent deletions that are larger than 1Mb and present in more than one cell line with an RO ≥ 99.9% are removed. Large deletions (≥100kb) that do not show consistent decrease of read depth compared with adjacent regions are also removed (less than one difference of read depth between deletions and flanking 10Kb regions). For inversions, recurrent

inversions that are longer than 100Kb and are present in more than one cell line (defined by RO ≥99.9%) are removed. For translocations, recurrent translocations that are present in more than one cell line (defined by both break ends being within +/- 50pb) are filtered out.

We required a minimal number of supporting reads (SR+PE) for translocation calls that we varied according to the sequencing depth and the ploidy of the WGS sample. (Cells with polyploidy can harbor an SV in only one copy of the DNA so that the SV is only present in a small fraction WGS reads.) Due to high sequencing coverage (~80X) in LNCAP sample, we only keep translocations with at least 15 supporting reads (PE+SR). For GM12878 cells (coverage of 50X), since they are diploid, we use a more stringent filter of 20 supporting reads, with at least two being split reads. For all other cell lines, which have similar read depth and ploidy, we require at least five supporting reads to call a translocation. We further compiled a list of high-coverage regions (coverage > 500X) in NA12878 which are largely characterized by repetitive genomic elements. In our initial analysis, we observed that such regions have high rates of translocation calls. However, given their extreme outlier coverage and association with repetitive elements, these are most likely simply anomalous alignments. We filtered out translocations whose breakpoint ends are located in those regions. In addition, for unclassified intra-chromosomal rearrangements called by Lumpy, we removed calls with a quality score less than 100. Finally, for tandem duplications, we require 10 supporting reads for LNCAP and 5 for GM12878 and 3 supporting reads for all other samples.

Cross-method comparison and integration of structural variants

The methods that we use to identify SVs appear to have different sensitivity for detecting SVs of different sizes. Specifically, Hi-C only rarely identifies SVs smaller than 1Mb. Therefore, we perform comparisons of SVs by dividing SVs into three different categories, namely, 1) inter-chromosomal

translocations identified by Hi-C, WGS and optical mapping, 2) large intra-chromosomal SVs (≥1Mb) identified by Hi-C, WGS, and optical mapping 3) and intra-chromosomal SVs < 1Mb that involves WGS and optical mapping. For the first two groups, we also included SV calls from additional methods, including karyotyping [127, 139-146], fusion transcripts, and paired-end tag sequencing (PET-seq)[147, 148]. Data from all six methods are available only for the T47D and K562 cell lines, we hence perform the cross-six-method comparisons in these samples. For six cell lines (Caki2, A549, NCI-H460, PANC-1, LNCaP and SK-N-MC), we have data from Hi-C, WGS, optical mapping, karyotyping, and RNA-seq, therefore we perfom a five-method comparison. For MCF7 cells, we have Hi-C, PET-seq (from two separate studies), and RNA-seq data, so we compared between these three methods in MCF7 cells. Finally, we have Hi-C data and fusion transcript data for PC3, SK-N-SH, SK-N-DZ, RPMI-7951 and G401 cells lines. Finally, we have Hi-C, optical mapping, and WGS data for the karyotypically normal cell line NA12878 that we use as a non-cancer cell line control.

We converted the strand orientation for SVs detected from different methods to a unified system, in which "+" indicates the breakpoint locates at the 3' end of the joined arm, and "-" indicates the breakpoint at the 5' end of the joined arm. For WGS data, this dictates that SV originally classified as deletions are given the strand orientation of "+-", inversions as "++ and - -", duplications as "-+" and unclassified intra-chromosomal rearrangement as "++" or "- -". Optical mapping originally reports deletions, which are assigned a strand orientation of "+-", inversions, which are assigned as "++" or "- -". Optical mapping also reports intra-chromosomal rearrangements >5Mb as "unclassified intra-chromosomal rearrangements" for which the software reports the strand orientation.

To determine whether the SVs detected by different methods reflect the same event, we set criteria for SV matching when comparing inter-chromosomal translocations and large intra-chromosomal SVs: 1) They have the same loci for both ends of the breakpoint. 2) They have the same

strand orientation. Because the different methods have very different resolutions for SV detection, we use variable criteria for determining whether two methods identify SVs at the "same loci". This overlap is set such that break ends within +/- 500Kb are considered as overlapping when comparing Hi-C, WGS, optical mapping, fusion transcripts and PET-seq. For karyotyping, an overlap of +/- 10Mb was set to accommodate for its low resolution. For specifically comparing deletions smaller than 1Mb, for calling to deletions as overlapping, we require that at least 50% of deletion defined by WGS must overlap with the deletion defined by optical mapping, and the size of the deletion detect by optical mapping must be within 80-120% of total length detected by WGS.

After identifying matched SVs between methods, we can resolve some unclassified SV types. Since we require SVs to have the same orientation, we can confirm certain Hi-C-detected intra-chromosomal SVs to be deletions, insertions or inversions if the same event was specified by optical mapping or WGS. Likewise, we can resolve unclassified intra-chromosomal variants from WGS to be inversions detected by optical mapping or Hi-C, and we can determine the SV type for unclassified large intra-chromosomal SVs identified by optical mapping as deletions, inversions and duplications if the orientation and SV type are determined by WGS or Hi-C. In addition, in our comparison of smaller scale of SVs, we found that insertions detected by optical mapping may be resolved as duplications in WGS, which we annotate as duplications.

We then calculated confidence levels for each SV and refine the SV coordinates based on the integration of different methods. Confidence levels are presented as the number methods by which each SV is detected. For refining the SV breakpoint coordinates, we choose loci determined by the highest resolution method for final breakpoint refinement. We consider WGS as the highest resolution method, followed by optical mapping, fusion transcripts, PET-seq, Hi-C, and then karyotyping.

Circos genome profiling

Genome profiles of cancer cell lines and GM12878 were generated using Circos [71]. Copy number is plotted according to the normalized CNV predicted by Control-freec for each 50Kb region. Duplications and deletions plotted if identified as high-confidence calls detected by at least two methods between Hi-C, WGS and optical mapping. Plotted rearrangements includes inter-chromosomal translocations, intra-chromosomal inversions and unclassified intra-chromosomal rearrangements, all of which are high-confidence calls that are identified at least twice between Hi-C, WGS, optical mapping, karyotyping, fusion transcripts, or PET-seq.

Size distribution of deletions and un-mappable translocations transitions:

Deletions

The size of deletion detected by WGS is simply the distance between the start and end of a deletion event. The size of deletion detected by optical mapping is calculated as: $Size_{deletion} = Size_{reference} - Size_{sample} = (Reference_{end} - Reference_{start}) - (Contig_{end} - Contig_{start})$. The size of final merged deletions detected by both WGS and optical mapping was defined by the size from WGS. Then we performed Wilcoxon rank sum test to examine the difference of deletion size detected by WGS and optical mapping.

Translocation un-mappable transition

In the detection of translocations, certain SVs will include a "transition" region between the two resolved portions of the rearrangement. The size of the un-mappable transition of a translocation detected by WGS is the number of basepairs that fail to align to either of the two rearranged regions. For a translocation detected by optical mapping between two chromosomes, chrA and chrB, is the distance between the closest two labels (L_A, L_B) that map to chrA and chrB respectively. There may be multiple

un-mappable labels between $L_{A,}$ $L_{B,}$ which are L_{A+1}, L_{A+2}, L_{A+3}…L_{A+M}, L_{B-N} … L_{B-3},L_{B-2}, L_{B-1}. To provide minimum size estimation of un-mappable transitions, we assume that the DNA from the last mappable labels to their nearest un-mappable labels (L_A *to* $L_{A+1,}$ L_B *to* L_{B-1}) are all mappable. Therefore, the size of an un-mappable transition in a translocation with no or one un-mappable label will be calculated as zero basepairs. For translocations with at least two un-mappable lables, the minimal size of the unmappable transition will be $|L_{B-1}- L_{A+1}|$. If an un-mappable region is detected by in a translocation by both WGS and optical mapping, we defined the size of the un-mappable regions as the size defined by WGS.

Genome-wide DNA replication timing

Genome-wide replication timing was measured in A549, Caki2, G401, NCI-H460, SK-N-MC, T47D and LNCaP using the Repli-seq method [156]. Briefly, asynchronously cycling cells were pulse labeled with the nucleotide analog 5-bromo-2-deoxyuridine (BrdU). The cells were then sorted into early and late S-phase fractions on the basis of DNA content using flow cytometry. BrdU-labeled DNA from each fraction was immunoprecipitated (BrdU IP), amplified and sequenced using Illumina HiSeq 2500. Replication timing was then measured as log2 ratio of early over late reads in 5kb bins. For K562, MCF7 and SK-N-SH cell lines, raw data for 6-fraction Repli-seq was downloaded from the ENCODE portal. The data was transformed to match the early/late repli-seq by combining G1, S1 and S2 fractions to represent early S phase and S3, S4 and G2 fractions to represent the late S phase. Smoothed replication timing profiles around the breakpoints were produced by loess smoothing replication timing data separately for the upstream and the downstream segments from the breakpoints predicted by Hi-C (**Figure** 2-1b, **Figure** 2-2e).

Classification of human genome into constitutive/switching regions

48 human replication timing datasets (ENCODE, www.replicationdomain.com) were used for the annotation of the human genome into constitutive/switching regions. The datasets were windowed into 50 Kb bins. Then the following criteria were used for the annotation. A threshold of above 0.15 was used to identify an early replicating bins and below -0.15 was used to identify a late replicating bin for each dataset. If a bin was early in 2 or more cell types and late in 2 or more cell types, those bins were classified as "Switching" (S). The remaining bins were then evaluated as being either "Constitutive Early" (CE), "Constitutive Late" (CL) or left un-classified (N/A). If a bin was early in at least 46 out of 48 cell types, it was classified as CE. If a bin was late in at least 46 out of 48 cell types, it was classified as CL.

Quantifying abrupt shifts in RT

Genome-wide replication timing profiles in cancer genomes show several abrupt shifts in replication timing associated with translocations. We sought to quantify the frequency of these abrupt shifts. To this end we made a pipeline to detect abrupt shifts next to translocations identified by Hi-C. For each predicted translocation, un-smoothed RT data in 5kb bins from +/- 200kb of the breakpoint was used to scan for abrupt shifts. A span of +/- 200kb was chosen because the resolution of Hi-C translocation calls started at 100kb. Then for every 5kb bin, the difference between the median of the preceding 20 bins and succeeding 20 bins were calculated. Outliers were removed from this metric by a median filter (span=5). Then a threshold of 0.6 was used to determine the presence/absence of an abrupt shift. While the threshold was chosen empirically, the results showed the same trend across a wide range of thresholds.

Characterization of deletions

Overall disruption of genes, repeats, enhancers and insulators

We evaluated the disruption of number genes, repeats, enhancers, and insulators that were deleted by high confidence deletions. High confidence deletions are defined as those that are detected by at least two methods out of WGS, Hi-C and optical mapping from in each cell lines: A549, T47D, Caki2, K562, LnCAP, PANC-1, SK-N-MC, NCI-H460, and NA12878. The number deleted genes or repetitive elements are simply calculated by intersecting the positions of deletions with gene annotations (NCBI RefSeq) and repeat annotations (UCSC repeatMasker) in the hg38 reference genome in each cell line.

In contrast to genes and repetitive elements, enhancers and insulators can potentially have cell type specific annotations. Therefore, to identify the number of deleted enhancers in each cell line, we first match each cancer cell line with a control normal cell type from the same or similar tissue type. We use H3K27ac as a mark for enhancers and CTCF binding sites as insulators. Specifically, we use human normal mammary epithelial (HMEC) cells as a control for T47D cells, blood mononuclear cells as a control for K562 and NA12878 cells, primary pancreatic tissue as a control for PANC-1 cells, and Normal human lung fibroblasts (NHLF) as a control for NCI-H460 and A549 cells. The only exception is that we use CTCF binding sites from NA12878 to annotate insulators in K562 cells, as no CTCF is available in mononuclear cells. By intersecting high confidence deletions in cancer cell lines with enhancers or insulators in matched control cell lines or tissues, we can evaluate how many enhancers or insulators are disrupted in the cancer cell line by deletions. Further, since the overall abundance of deletions can vary in each cancer cell line, we calculate the number of lost enhancers per 100Kb of deleted genome, and then normalize this number to a constant value of 100,000 enhancers per genome.

Estimates of enhancer deletion enrichment relative to random controls

To estimate whether enhancers were preferentially deleted or retain, we performed simulation by randomly distributing the high confidence deletions each cancer genome 1000 times and then examining their overlap with enhancers. The distribution of the overlap between deletions and enhancers can then be summarized and plotted. The empirical P value is calculated based on how many times the simulated number of deleted enhancer is smaller than that number in fact observed from a given cell line.

Identifying polymorphic and novel deletions

High confidence deletions are stratified into two categories: known polymorphc deletions and novel variants. This is accomplished by intersecting deletions with variants reported in DGV SVs annotated as "deletion", "loss", and "loss and gain" using *bedtools* [75]. A detected deletion must have at least 90% reciprocal overlap between the detected deletion and deletions documented in DGV dataset to be considered as polymorphic. Some deletions reported in DGV are overlapping with each other. In such cases, if these deletions overlapped with exactly same region across the nine cell lines, these were treated as a single deletion event. Deletions that do not overlap with variants reported in DGV are defined as novel variants.

Enrichment analysis of polymorphic deletions and novel deletions

To evaluate the enrichment of various genomic featurs with polymorphic or novel deletions, we first began by sorting and merging all polymorphic and novel deletions detected by both WGS and optical mapping in K562, T47D, Caki2, and GM12878 cells. The number of polymorphic and novel deletions were then counted in each cells, and the proportion of polymorphic vs. novel deletions was then compared between cancer cell lines and NA12878 cells. The overall loss of DNA content caused by polymorphic deletions or novel deletions was also calculated by summing the length of all non-redundant deletions identified in each cell. To determine if there is an enrichment of either class of deletion with genes, polymorphic and novel deletions from the nine cell lines were intersected with

RefSeq genes. Genes were further annotated using the list of COSMIC-tumor related genes, considering only genes with clear annotations as oncogenes or tumor suppressors. The overlap of different classes of deletions with exons was evaluated by comparing polymorphic and novel deletions with non-redundant exons from refFlat records of GENCODE24. The overlap of different classes of deletions with repetitive elements was evaluate by comparing deletions with non-redundant repetitive elements obtained from the UCSC repeatMasker. For example, for polymorphic deletions in K562 cells containing *i* events, if the size of each deletion is DEL_i, and if the size of overlap with repeats from each deletion is Rep_i, the enrichment of repeats ($Enrich_{repeats}$)was calculated as:

$$Enrich_{repeats} = \frac{\sum Rep_i}{\sum DEL_i}$$

We also determined whether there was an enrichment for deletion of enhancers by polymorphic or novel enhancers. This was accomplished by randomly permuting deletions 1000 times in each cell type, and calculating the overlap with H3K27ac defined enhancers in the same control normal cell lines listed above. The empirical P-value was calculated based the random shuffling. The results from the two classes of deletions was then compared across each cell type to test whether the enhancer loss is preferentially associated with novel or polymorphic deletions.

Gene ontology analysis of deleted enhancers

To perform ontology analysis of enhancer deletions, the locations of high confidence deletions in T47D cells was intersected with H3K27ac defined enhancers in HMEC cells. After removal of duplicates, the loci of deleted enhancer were lifted-over from hg38 to hg19 and gene ontology analysis was performed by GREAT using the hg19 reference as background [76] (GREAT requires the use of the hg19 reference). Association rule was set as “Basal plus extension”, with “proximal 5.0kb upstream”, “1.0 kb downstream”, and “plus Distal: up to 1000.0kb”.

Fusion of transcripts

We downloaded paired-end RNA-seq data for 14 cell lines from the ENCODE project, European Nucleotide Archive (ENA), or Sequence Read Archive (SRA) databases (**Table** 2-S 1). We used three different pipelines (Tophat-Fusion [v2.1.0][157], Star-Fusion [v1.1.0][158], and EricScript [v0.5.5][159] to identify fusion transcripts. For Tophat-Fusion, paired-end reads were aligned to a GRCh38 reference genome (version GCA000001405.015) to identify fusion events. Tophat-Fusion was run on the following parameters: "--no-coverage-search -r 50 --mate-std-dev 80 --max-intron-length 100000 --fusion-min-dist 1000 --fusion-anchor-length 13". Tophat-Fusion outputs a list of potential fusion events, which were then processed by Tophat-fusion-post to filter out false positives by aligning sequences flanking fusion junctions against BLAST databases. Fusion events were further filtered requiring at least three split reads or three spanning read pairs. In Star-Fusion, a built-in GRCh38 reference genome with Gencode v26 annotation was used. Fusion transcripts were detected by Star-Fusion with default parameters. To reduce false positives, fusion events with a Fusion Fragment Per Million total reads (FFPM) less than 0.1 were removed. EricScript detects fusion transcripts by aligning the reads to a pre-built reference transcriptome (Ensembl Version 84) provided by the authors. Candidate fusions are further required to be supported by at least three spanning read pairs and three split-reads. We also included a fourth set of fusion transcripts from Kljin et al[160]. The final set of fusion transcripts was obtained by considering the union of fusion calls from the three pipelines and the fourth set of fusion events identified by Kljin et al.

Identification of allelic imbalance in expression: To evaluate the effects of TAD fusion events on altered gene expression in *cis*, we tested whether TADs containing rearrangements showed different patterns of allele specific gene expression compared to TADs that lack rearrangements. For each cell

line where we had WGS (T47D, Caki2, K562, A549, NCI-H460, PANC-1, LNCaP, SK-N-MC), we aligned RNA-seq data to the genome using STAR. We then implemented the WASP pipeline (PMID: 26366987) for filtering and re-aligning reads to identify reads that show inherent allelic mapping biases. We then computed the number of reads that aligned to each allele at each single nucleotide variant within an exon of any GENCODE gene using samtools mpileup. The number of reads aligning to each allele was normalized by the total number of reads (RPM), to account for sequencing depth differences between cell lines. To compute the degree of bias in expression between alleles, we used a simple chi-squared statistic. To account for potential differences in copy number between alleles, the expected value of the chi-squared statistic for each SNV was derived from the observed ratio of coverage between alleles from WGS. Specifically, the expected value for each allele was calculated as the fraction of reads from WGS aligning to that allele multiplied by the sum of the RNA-seq RPM values across both alleles.

Re-prediction of gap sizes

To gain a list of candidate unresolved gap regions, recurrent deletions detected by optical mapping at least twice in cancer cells lines and at least once in normal cells were collected from 12 samples, including 8 cancer cell lines (T47D, Caki2, K562, A549, NCI-H460, PANC-1, LNCaP, SK-N-MC) and 4 normal cells (GM12878, 3078entB, 3045entB, and 3391entB). Recurrent deletions were then intersected with hg19 gaps using *bedtools*. Only gaps where at least 80% of the gap overlap with a deletion and the gap accounts for at least 30% of the deletion are retained for gap size re-estimation. When using hg19 as the reference genome, the gap size was predicted by subtracting the deletion size from gap size in hg19. To evaluate the predictions, the gap regions were lifted over to GRCh38, and the sizes of the same regions in GRCh38 were compared with our prediction and the size in hg19. Some

gaps will ultimately have a negative value, meaning that the size of the deletion is shorter than annotated gap in the reference genome, potentially due to the variation across populations.

To predict the size of unresolved gaps in GRCh38, we repeated our analysis of deletions overlapping gap regions using GRCh38 as the reference genome as described above. In some cases, the re-estimated size of the same gap could vary among different cell lines, and the degree of variation is relatively small with respect to the overall change of perceived scale of gap size. Therefore, we report the median, the maximum, and minimal gap size of each gap from our estimation, as this variation can represent polymorphisms of gap sizes in the population. We then annotate what genes are spanned by those adjusted gaps and could be affected by intersecting re-estimated gaps with gene list in GRCh38. We further compare our gap size predictions in GRCh38 with results from previous publications [40, 103].

Profiling of gene copies using optical mapping

In order to identify genes that had undergone copy number alterations, we compared copy number profiles from optical mapping in the 4 primary normal tissues and 8 cancer cell lines with gene lists from RefSeq Gene annotation. The longest isoform was used for characterization of copy number changes. For each gene, the average copy number profiles of each 50kb bin spanned by the gene was considered as the copy number of that gene. The CNV of genes were also profiled by WGS normalized coverage (Control-FREEC) in T47D and Caki2 for differential gene expression analysis.

Differential gene expression from gene dosage or enhancer deletion

To evaluate the effects of gene dosage and enhancer deletions on gene expression, we evaluated the expression of genes in T47D or Caki2 cell lines where we detected copy number alterations of the gene itself or of linked enhancers. For T47D, we used RNA-seq data from HMEC cells as a normal

control, and for Caki2, we used RNA-seq data from primary kidney tissue as a normal control. We downloaded FASTQ files of paired-end RNA-seq data from T47D, HMEC, Caki2, and primary kidney from the SRA database or ENCODE. Each sample contained two replicates. The raw reads were aligned, and differential expression analysis was performed using *Tophat* and *cufflinks* [161]. For analyzing the impact of gene dosage on expression, we grouped genes into 4 classes: homozygous deletions (0 copy), genes with LOH (1 copy), normal genes (2 copies), and amplified genes (≥3 copies) according to CNV profiles from WGS. We calculated the expression (FPKM) fold change of all genes in each category relative to the control sample.

For analyzing the impact of enhancer deletion of gene expression, we first filtered genes and removed those with deletions of exons or entire genes to control for deletions that the impact of gene dosage on expression. We further filter genes and focus only on the 9672 genes with evidence of expression in HMEC cells (FPKM ≥ 1). Enhancers were annotated enhancers as homozygous deletion or LOH based on WGS coverage, and were examined for linkage to filtered genes from significant interactions identified by capture Hi-C in GM12878 cells. The expression fold change in expression between T47D and HMEC cells was then computed for the 530 genes with a copy number loss of linked enhancers was compared with 9142 unaffected genes using the Wilcoxon test.

TAD fusions

To evaluate the effects of SVs on TAD structure, we analyzed breakpoint crossing Hi-C signal. Our initial observations identified cases where the nearest TAD boundaries to the breakpoint were being "fused" together to create a new TAD. To evaluate whether such TAD fusion events were generally the case, we analyzed whether the breakpoint crossing Hi-C signal between the nearest TAD boundaries showed a local enrichment, which is characteristic of "normal" TADs.

We begin this analysis with a list of breakpoints within each cell type. For each breakpoint, we identified the nearest breakpoint proximal TAD boundary based on TAD calls from H1 hESCs. We chose TAD calls from H1 hESCs as we wanted to use TAD calls from an independent, non-rearranged cell type, in case the rearrangement was altering TAD calls within the rearranged cell line. We should note that TAD calls are highly stable between cell types, such that these results are similar regardless of the source of the TAD calls. We then identified the predicted "peak" of the TAD "triangle" by identifying the bin representing the interaction between each of the nearest breakpoint proximal TAD boundaries. The bin representing the interaction between each of the breakpoint proximal TAD boundaries was then considered as the center of a sub-matrix. We calculated the average interaction frequency of all bins within the 41x41 bin sub-matrix centered on the TAD boundary interacting bin. Each bin was then normalized to this average interaction frequency, such that the new sub-matrix would represent a fold change above the average value in the sub-matrix. This was then log-transformed (with a pseudocount of 1 added to avoid taking the log of zero and to minimize the effects of noisy low frequency interactions). The reason for normalizing to mean of the submatrix is to account for the differences in interaction frequencies that would be expected due to genomic distance alone. In other words, without normalizing to the central bin, the aggregated Hi-C data would be dominated by short distance interactions. The log-fold change sub-matrix was then averaged for all breakpoints in all cell types, yielding a single aggregate log fold-change sub-matrix. For display purposes, this was then exponentiated to represent these values again as a fold change. This process was also applied to a random set of TAD boundaries. Random TAD boundaries analysis was performed by first randomly permuting the TAD boundaries from H1 hESCs, using the following the following approach: for TADs on chromosomes affected by SVs, we generate a random number between 1 and the size of the chromosome where it is located. This number is then added to the start and end coordinates of the every

TAD on the chromosome. If the randomly generated TAD is larger than the size of the chromosome, the size of the chromosome in base-pairs is then subtracted. This is done to preserve to observed size and spacing of TADs in the random dataset to limit any artifacts or bias of randomization. This set of permuted TADs was then used for the input into the same process as described to evaluate the chromatin interactions across the breakpoints. We want to point out that the only data that is being randomized are the positions of TADs, and the SVs and chromatin interaction maps used for the plot are both from the true cancer cell lines in this study. This randomization was repeated 1,000 times.

Chapter 3

Whole genome optical mapping reveals previously unrecognizable structural variants in leukemia patients' samples

Abstract

While genomic analysis of tumors has stimulated major advances in cancer diagnosis, prognosis and treatment, current methods fail to identify a large fraction of somatic structural variants in tumors. We have applied optical genome mapping in conjunction with whole genome sequencing to twelve adult and pediatric leukemia samples, which revealed on average over five thousand structural variants per sample. Our computational methods determined that 5-10% of the variants, including insertions, deletions, translocations and inversions, likely arose as somatic mutations. These somatic structural variants affected 37 leukemia associated genes as well as 209 cancer driver genes not previously associated with leukemia and at least 109 recurrently disrupted genes not previously associated with cancer. Fifteen of the genes not previously associated with AML but mutated in multiple patients' samples significantly affects survival of AML patients. In addition, many variants (42%) resided exclusively in intergenic regions and a significant fraction of these caused cis-acting alterations in expression of neighboring cancer-associated genes. Our results suggest that current genomic analysis methods fail to identify a majority of structural variants in leukemia samples and this shortcoming may hamper diagnostic and prognostic efforts.

Introduction

Genomic analysis of tumors has stimulated major advances in cancer diagnosis, prognosis and treatment, shifting the focus from morphological and histochemical characterization to consideration of the landscape of driver mutations in the tumor [15, 162, 163]. This has been particularly true for leukemia, and especially so for acute myeloid leukemia (AML), in which the spectrum of driver mutations provides a much more rigorous classification of disease subtypes, with a correspondingly more robust prognostic power, than previous histological characterization [164, 165].

Somatic driver events in a tumor – point mutations, small indels copy number changes and structural variants (SVs) including insertions, deletions, inversions, translocation and copy number change – are currently identified by some combination of karyotyping, comparative genome hybridization, fluorescence in situ hybridization (FISH), RNA sequencing and genome sequencing of either targeted gene panels, whole exomes or whole genomes [14-16, 21, 162]. However, our recent study interrogating a variety of cancer cell lines using an integrative framework for detecting SVs, consisting essentially of whole genome sequencing, optical genome mapping and chromosome conformation capture, identified a large number of variants that were undetectable by the standard tools for cancer genome analysis [166]. Moreover, some of these previously undetected SVs affected cancer relevant genes through their gain or loss or through alteration in expression. In the latter case, gene expression could be reduced by deletion of an associated regulatory domain or activated by fusion of topologically associated domains, bringing an otherwise inactive oncogene in functional proximity to an active enhancer region. This study strongly suggested that non-coding SVs are underappreciated drivers in cancer genomes. However, since this study investigated only cell lines, it could not differentiate between cancer promoting variants versus variants that arose during establishment and propagation of the cell line itself nor could it identify somatic versus germ line variants.

Today, our understanding of the SV landscape in AML remains a limited group of well-known translocations and inversions, such as t(8;21) and inv(16), beyond which the rest majority of patients' SVs are less uncovered and non-interpretable. Here we present this pilot study that adopts long-molecule strategy to profile and interrogate novel SVs genome-wide on leukemia clinical settings, and for the first time an exploration to assess the genome-wide impact of non-coding SVs. We have applied optical mapping in conjunction with whole genome sequencing (WGS) to obtain a significantly enhanced view of somatic SVs in a dozen different adult and pediatric leukemia samples. To distinguish somatic variants from the much larger fraction of germline variants we developed a computational pipeline, which, in contrast to our previous study with cell lines, we were able to test by access to patients' germline samples. In almost all cases, our analysis identified all the structural rearrangements previously determined by standard karyotype analysis. However, our analysis also revealed hundreds of additional SVs, particularly insertions and deletions but also inversions and translocations, which were not evident from standard genomic analyses. A number (304) of these variants affected tumor associated genes, whose role in prognosis and treatment in the individual cases could not otherwise have been considered. Our work further confirms that the extent of somatic SVs have not been fully recognized nor effectively integrated into disease assessment. The methods described here may offer a remedy for that shortcoming.

Results

Identification of somatic structural variants in leukemia samples.

We used optical mapping in conjunction with whole genome sequencing to identify structural variants in blood samples from leukemia patients. Patients included seven adult AML cases, two pediatric AML cases, one pediatric T-cell ALL case, one pediatric B-cell ALL case and one adult B-cell

lymphoma (**Table 3-S**1). We performed whole genome sequencing on all samples at an average depth of 50X and optical mapping at 100X coverage on a Bionano Genomics Irys or Saphyr optical mapping instrument. For optical mapping, large genomic fragments (>250 kb) are extracted from cells, fluorescently labeled with a site-specific DNA binding protein and then passed through nanochannels of an Irys or Saphyr chip that force the molecules into a strictly linear conformation. After they are linearized and migrate through the nanochannels, DNA molecules are imaged, with the fluorescent tags providing a bar code that allows subsequent assembly of individual molecules into larger contiguous maps, which are compared to a reference genome to identify insertions, deletions and rearrangements.

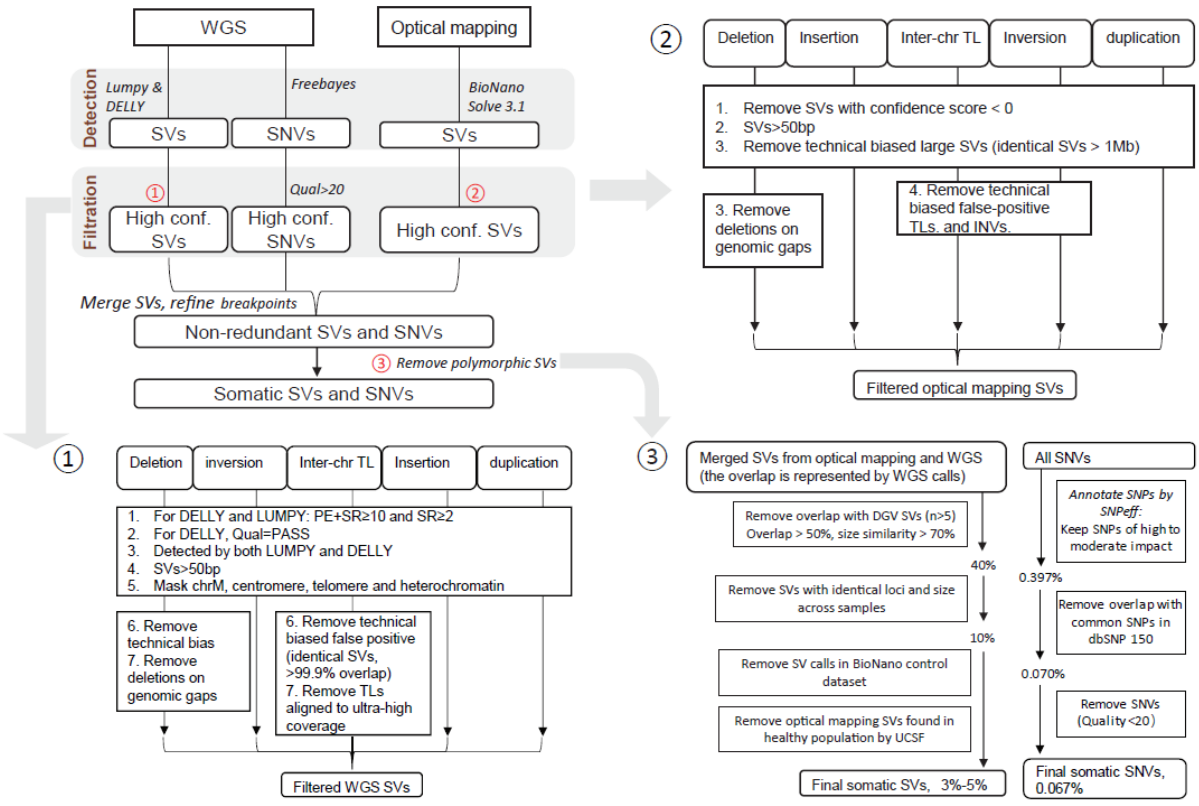


Figure 3- 1. Computational Workflow for Detection of Structural Variants.

The computational workflow for extracting structural variants from a combination of whole genome sequencing and optical mapping is diagrammed in the upper left hand figure with details of each of the subroutines provided in the numbered figures. See Materials and Methods for a detailed explanation. Note that step 3 removes likely germ line polymorphisms from the SV calls, reducing the number of original SVs by 95-97% on average.

Data processing to identify structural variants in individual samples is outlined in **Figure 3-**1. Whole genome sequence data was mapped to human genome reference hg38 using BWA and then filtered for structural variants by two independent software pipelines, LUMPY and DELLY. Those variants identified by both programs were retained and sorted into subtypes: deletions, insertions, duplications, inversions and intra- and inter-chromosome translocations. Copy number variants were determined by Control FREEC. Structural variants were extracted from optical mapping data using Bionano Genomics Access software. The non-redundant union of the variants determined by each of the methods yielded in each sample 1500-3000 deletions, more than 2000 insertions, hundreds of inversions and copy number variants and tens of translocations (**Table 3-S**2).

Determining which of the structural variants arose as somatic mutations versus those that were preexistent in the patient's germ line would require comparing those present in the leukemia sample to those in the patient's normal genome. However, since normal tissue is not readily available from most leukemia patients, we developed a computational pipeline to distinguish somatic mutations from germline polymorphisms by filtering the list of variants against various databases of known genomic polymorphisms. We first compared the position and extent of each variant against the Database of Genomic Variants [167] and removed any variant that significantly overlapped a previously identified variant. We then removed any variant whose start and end point were identical in two or more of our patient samples. Finally, since many of the variants identified by optical mapping could not have been previously revealed by other technologies, we compared our remaining variant list against that obtained from optical mapping of 154 normal individuals in a study recently conducted by Kwok and colleagues [168], as well as that in Bionano Genomics' dataset of variants found in normal individuals. As noted in **Figure 3-**1 and **Table 3-S**2, this filtering process significantly reduced the number of variants such that on average only 13% of the initially identified copy number gains and only 5% of initially identified

deletions, insertions and inversions were retained as likely somatic variants. In contrast, all of the interchromosomal translocations initially identified were retained as likely somatic events.

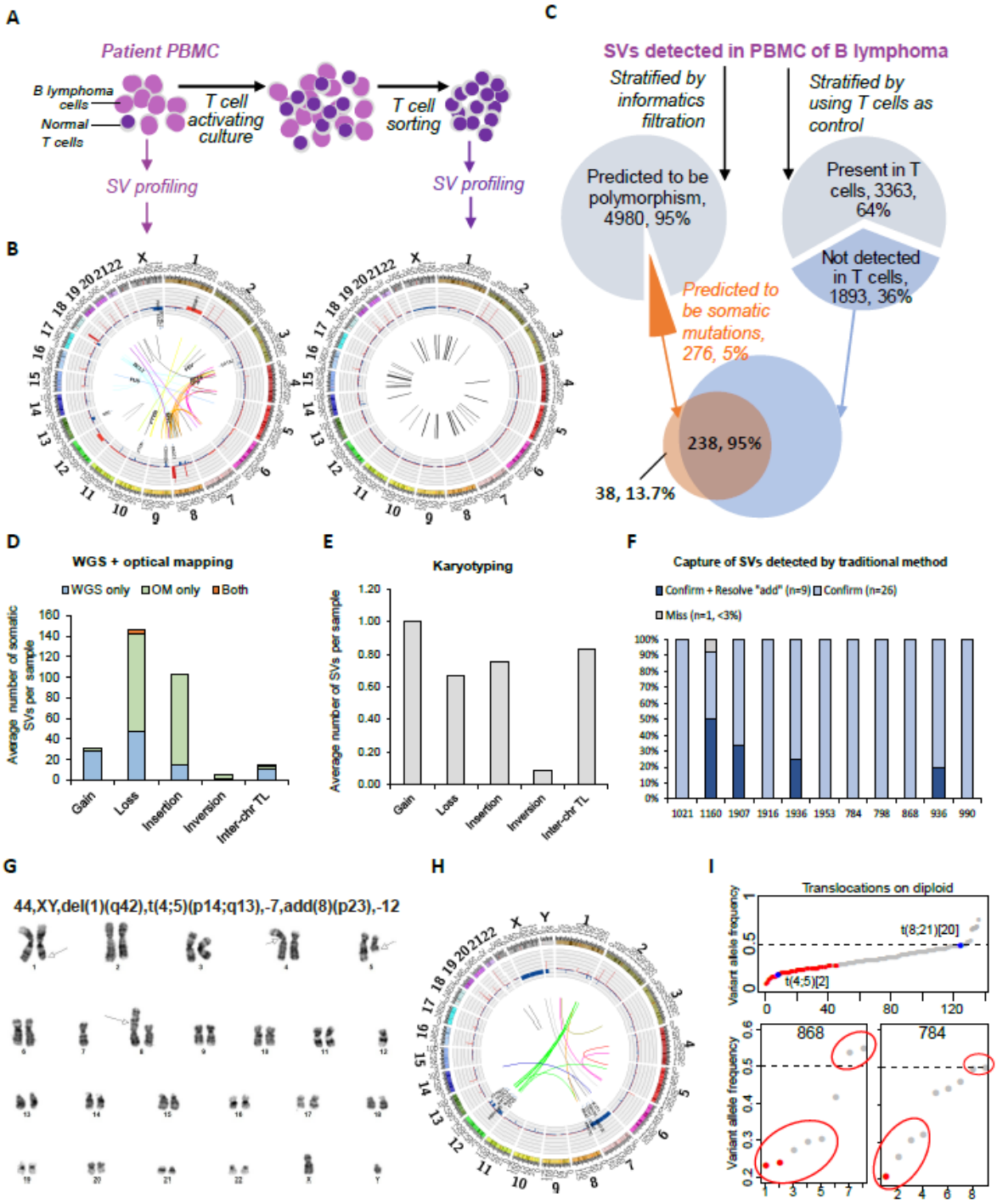


Figure 3- 2. Figure 2. Detection of Structural Variants by WGS+OM versus Karyotyping

(A) Workflow for isolating normal T cells from a B lymphoma sample for validating somatic SVs predicted by our bioinformatics pipeline. (B) Circos plots for B lymphoma sample 1160 and the corresponding T cell demonstrate the euploid genome and the absence of interchromosomal translocations in the T cells population. Chromosomes are arrayed in clockwise order from 1 to X, with inversions and translocation shown in the center and copy number variations arising from deletion (blue) or duplication (red) shown on the inner ring. (C) Workflow for determining the false discovery rate for somatic SV prediction by comparison to the T cell control. (D) The average number of different types of somatic SVs over twelve patient samples detected by whole genome sequencing plus optical mapping. Gains include any duplication of genomic sequences greater than 50 bp and losses refer to local or extended elimination of genomic sequences. (E) The average number of SVs detected by karyotype analysis of the twelve patient samples. (F) For each patient sample, the percent of SVs previously identified by karyotyping that were confirmed by WGS+OM, subdivided into those that were only confirmed (light blue) and those that were confirmed and the source of added material was resolved (dark blue). (G) The karyotype image of sample 936. Arrows indicate translocation fusion points. (H) The circos plot of the structural variants derived from optical mapping and whole genome sequencing of sample 936. (I) Determination of variant allele fraction (VAF) from WGS data. VAF was calculated as the total number of read spanning the translocation breakpoint divided by the total number of reads spanning the breakpoint plus the total number of read mapping to the intact chromosome at the same site of either one of the participating chromatids. Two translocations identified by karyotyping are indicated by blue dots, corresponding to t(8;21) in sample 784, which was observed in 20/20 karyotype images, and t(4;5) in sample 936, which was observed in 2/20 karyotype images. Lower panels show that VAF separates translocations in each sample into homogenous mutations (upper circle) and sub-clonal mutations (lower circle).

We tested the validity of our filtering algorithm in identifying somatic variants in one case in which we were able to obtain normal tissue for the patients. We amplified the small subset of normal T cells from the leukemic blood sample by selective application of growth factors as described in Materials and Methods (**Figure 3-**2A-C). We performed optical mapping and whole genome sequencing on these germ line samples and compared those profiles to those of the corresponding leukemia sample to identify somatic variants. We then compared the collection of somatic variants identified by direct comparison to germ line sequences to that obtained by the computational filtering process described above. As evident from **Figure 3-**2C, >86% of the somatic variants identified by our filtering process were not observed in the T cell genome, providing a false discovery rate of <0.14. The majority of the false positives were short private SVs –germline but individual specific— which are impossible to eliminate computationally by using public databased but comprise only a small fraction of the total

variants. These results support our filtering pipeline as a convenient, cost effective and accurate method for pinpointing somatic variants in leukemia genomes.

Comparison of karyotyping, optical mapping and whole genome sequencing.

All of the leukemia samples we examined had been previously analyzed by cytological karyotyping as part of the patients' standard clinical evaluation. **Figure 3-**2D and E and **Table 3-S**3 presents a comparison of the somatic SVs identified by each of these methods alone or in combination. As evident, karyotyping revealed only a small fraction of the SVs present in the sample. Optical mapping identified essentially all the variants noted by karyotyping, missing only those in low-abundance subclones, and was effective in identifying inversions (84%), insertions (86%) and larger deletions (67%). On the other hand, whole genome sequencing was adequate for identifying most copy number gains (91%) and interchromosomal translocations (91%) but failed to identify the majority of insertions and deletions.

As evident from **Figure 3-**2D, **Table 3-**S3 and from previous work [166], WGS and optical mapping provide synergistic data on SVs. For insertions and deletions, optical mapping picked up larger variants while WGS identified smaller events. In 39% of cases in which WGS failed to flag a variant detected by optical mapping, one or both endpoints lie in a low mappability region of the genome. On the other hand, for 57% of the translocations detected by WGS but unreported by optical mapping, one side of the variant was too short to encompass at least nine labeling sites, which is the minimum for the mapping software to provide statistically reliable calls (**Figure 3-**S1). This was particularly evident in cases of chromothripsis. Nonetheless, for many of the variants identified by only one of the two methods, the second method does provide confirmation of the validity of the call (**Figure 3-**S1), either

by confirming one half of the variant or its reciprocal event. In sum, WGS confirmed 75% of the translocations identified by optical mapping while optical mapping confirmed 67% of the translocations identified by WGS or involved fragments too small to be detected. Finally, we verified by PCR amplification ten out of ten tested of the translocations uniquely identified by WGS. Accordingly, we are confident that our integrated method reveals a large fraction of SVs previously unrecognizable.

The combination of optical mapping and whole genome sequencing identified 157 interchromosomal translocations in twelve samples, the majority of which were missed by karyotyping. These included all but one of 36 genome rearrangement reported by karyotyping (**Figure 3-**2F), the missing event being present in only a very small fraction of the patient's sample. Optical mapping plus sequencing provided a more detailed characterization of translocations than was available from karyotyping. **Figure 3-**2G and 2H shows the karyotype image from sample 936 and the corresponding circular genome structure (circos) plot derived from optical mapping and whole genome sequencing. As evident from the circos plot, chromosome 12 had undergone chromothrypsis in this patient's sample with the majority of the residual fragmented chromosome transposed to chromosome 1. This was not apparent from the karyotype analysis. In a second case, 1021, our analysis documented a three-way reciprocal rearrangement among three separate chromosomes (**Figure 3-**S2), suggesting that the triple rearrangement occurred as a concerted event. In nine other cases, karyotyping reported that unidentified genetic material had been added to a chromosome without specifying the source of that additional material. In all such cases, our methodology was able not only to identify the source of the exogenous DNA but also to pinpoint the precise junction of the added material (**Table 3-S**4 and **Figure 3-**3). For example, karyotyping indicated additional material on chromosome 12 in patient 1160. Optical mapping identified the extra sequence as arising from chromosome 12 itself, involving an internal inverted duplication of ca. 50 Mb in the middle of the chromosome (**Figure 3-**3). This analysis also

accounted for the duplicated segment of chromosome 12 identified from our copy number determinations. Circos plots for all the samples analyzed in this study are shown in **Figure** S3.

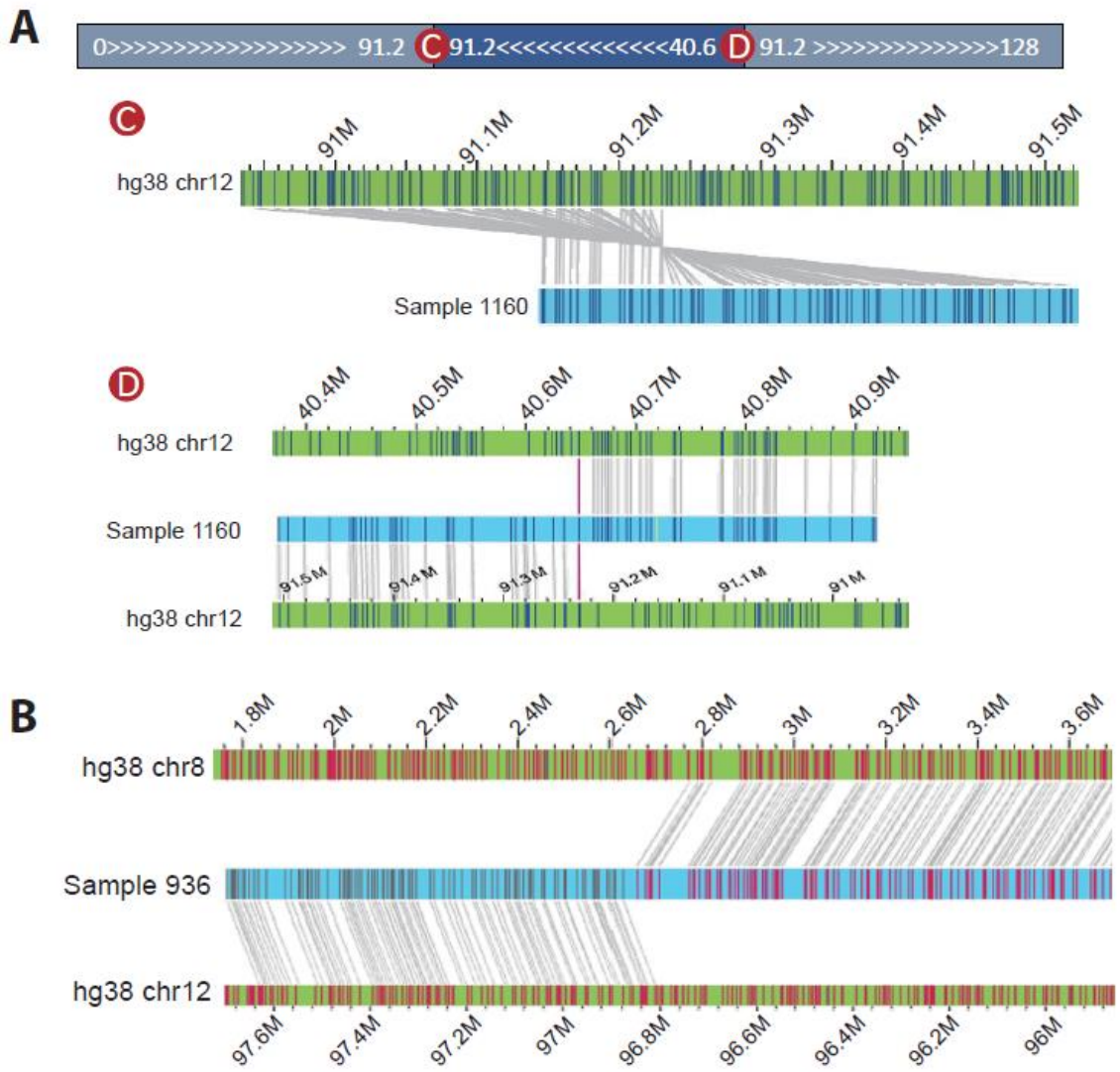


Figure 3- 3. Identification of previously undetermined added chromosomal sequences.

(A) Diagram of chromosome 12 derived by optical mapping of patient sample 1160, which had been identified by karyotyping only as chromosome 12 with additional material. Optical mapping indicates that the additional material is a 50 Mb inverted duplication starting at position 91.23 Mb and reconnecting to the remainder of the chromosome at position 40.65 Mb. Below are genome maps showing the hg38 reference chromosome in green with in silico determined labeling sites indicated as dark blue vertical lines, aligned with the chromosomal assemblies around the inversion initiation site at 91.2 Mb (C) and the inversion termination (D) obtained from optical mapping. (B) Genome assembly map in light blue of a chromosome 8/12 fusion obtained by optical mapping of patient sample 936. Detailed mapping of the sample indicates that the chromosome 12 fragment derived from chromothrypsis of one of the chromosome 12 chromatids.

Our method provides information on the relative abundance of a SVs in a leukemia sample. For instance, calculating the number of reads from WGS that span a translocation breakpoint in relation to the number of reads spanning the intact chromosome at the same site retrieves the allele fraction of that translocation in that sample. For a heterozygous translocation present in 100% of the cells, the allele fraction would be 0.5 for a diploid local region. As shown in **Figure 3**-2I, that is the case for the t(8;21) in patient 784, consistent with the karyotype data indicating that the translocation is present in 20 of 20 images. However, in line with the fact that most somatic SVs and single nucleotide variations (SNVs) are present in subclonal populations in AML cases, we find that the allele fraction of the translocation in most cases is less than 0.5. For instance, we calculated that the allele fraction of t(4;5) in patient 936 is 0.18, consistent with the karyotype report identifying the translocation in 2 of 20 images. Further, in applying this analysis to individual samples, we observe evidence of distinct subclones within the population. As evident in **Figure 3**-2I for samples 868 and 784, several translocation cluster at an allele fraction of 0.2-0.3 while other translocations cluster around an allele fraction of 0.5. In sum, our method provides not only the identification of SVs in patient samples, but also the relative abundance of each variant and evidence of subclones within a sample.

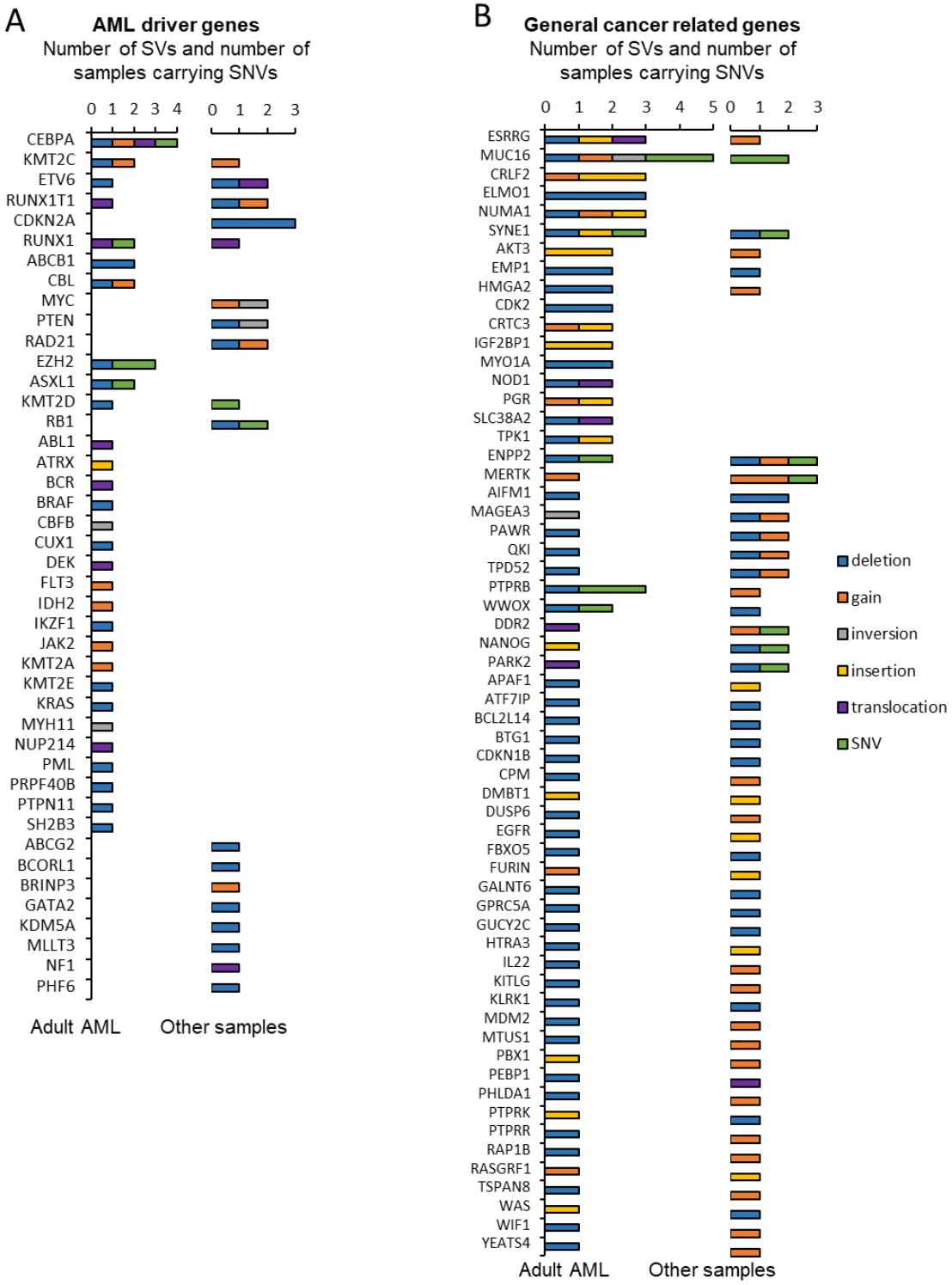
Functional significance of somatic structural variants.

The somatic SVs identified in our leukemia patient samples overlap those previously associated with leukemia but also affect additional cancer genes. Whole genome sequence analysis revealed SNVs or small insertions or deletions in each sample that were previously recognized as driver mutations. Our combined analysis further identified in many samples structural rearrangements in genes previously linked to leukemia. In sum, over all twelve patients we identified SVs in thirty-six genes and SNVs in an

overlapping set of fifteen genes that were previously implicated as genetic drivers in leukemia, twenty-two of which were mutated in two or more patients (**Figure 3**-4A). Only one patient carried an FLT^{ITD} mutation (FLT-internal tandem duplication), one of the most common mutations in all the AML cohorts analyzed to date. This was noted in their clinical reports and confirmed in our hands by WGS and targeted PCR.

In several cases, we were able to identify loss of tumor suppressor genes that could not be readily detected by conventional methods. In one example shown in **Figure 3-**5A and B, a somatic inversion disrupted the *PTEN* gene on chromosome 10 and a somatic deletion removed the terminal exon of *PTEN* on its homolog. Neither of these SVs were present in the patient's clinical report nor identifiable with whole genome sequencing alone. As a second example in **Figure 3-**5C and D, *BCL6* is disrupted by an inversion on chromosome 3 while its homolog is disrupted by a deletion. As above, neither of these were reported for the patient nor readily evident in the absence of optical mapping.

We also identified SVs associated with genes previously identified as cancer-associated but not frequently with leukemia (**Figure 3**-4B). We found *CRLF2* altered in eight patients, twice by insertion, once by amplification and five times by point mutation. *CRLF2* encodes a type I cytokine receptor, which along with the IL7 receptor activates the JAK2-STAT pathway, and has been found rearranged in B-cell ALL but not previously in AML [169-171]. We also observed alteration in three patients of *RSPO2*, a gene encoding a member of the R-spondin family of proteins that activate WNT signaling. Mutations in RSP02 has been seen in a gastric, liver and colorectal cancer and neuroblastoma but not previously reported in leukemia [172-174]. As a final example, *NUMA1*, an essential component in the formation and organization of the mitotic spindle, is altered variously by point mutations, insertion, deletion and amplification. A chromosomal translocation of this gene has been associated with acute promyelocytic leukemia [175, 176].


A
AML driver genes
Number of SVs and number of samples carrying SNVs
0 1 2 3 4
0 1 2 3
CEBPA
KMT2C
ETV6
RUNX1T1
CDKN2A
RUNX1
ABCB1
CBL
MYC
PTEN
RAD21
EZH2
ASXL1
KMT2D
RB1
ABL1
ATRX
BCR
BRAF
CBFB
CUX1
DEK
FLT3
IDH2
IKZF1
JAK2
KMT2A
KMT2E
KRAS
MYH11
NUP214
PML
PRPF40B
PTPN11
SH2B3
ABCG2
BCORL1
BRINP3
GATA2
KDM5A
MLLT3
NF1
PHF6
Adult AML
Other samples
B
General cancer related genes
Number of SVs and number of samples carrying SNVs
0 1 2 3 4 5
0 1 2 3
ESRRG
MUC16
CRLF2
ELMO1
NUMA1
SYNE1
AKT3
EMP1
HMGA2
CDK2
CRTC3
IGF2BP1
MYO1A
NOD1
PGR
SLC38A2
TPK1
ENPP2
MERTK
AIFM1
MAGEA3
PAWR
QKI
TPD52
PTPRB
WWOX
DDR2
NANOG
PARK2
APAF1
ATF7IP
BCL2L14
BTG1
CDKN1B
CPM
DMBT1
DUSP6
EGFR
FBXO5
FURIN
GALNT6
GPRC5A
GUCY2C
HTRA3
IL22
KITLG
KLRK1
MDM2
MTUS1
PBX1
PEBP1
PHLDA1
PTPRK
PTPRR
RAP1B
RASGRF1
TSPAN8
WAS
WIF1
YEATS4
Adult AML
Other samples
deletion
gain
inversion
insertion
translocation
SNV

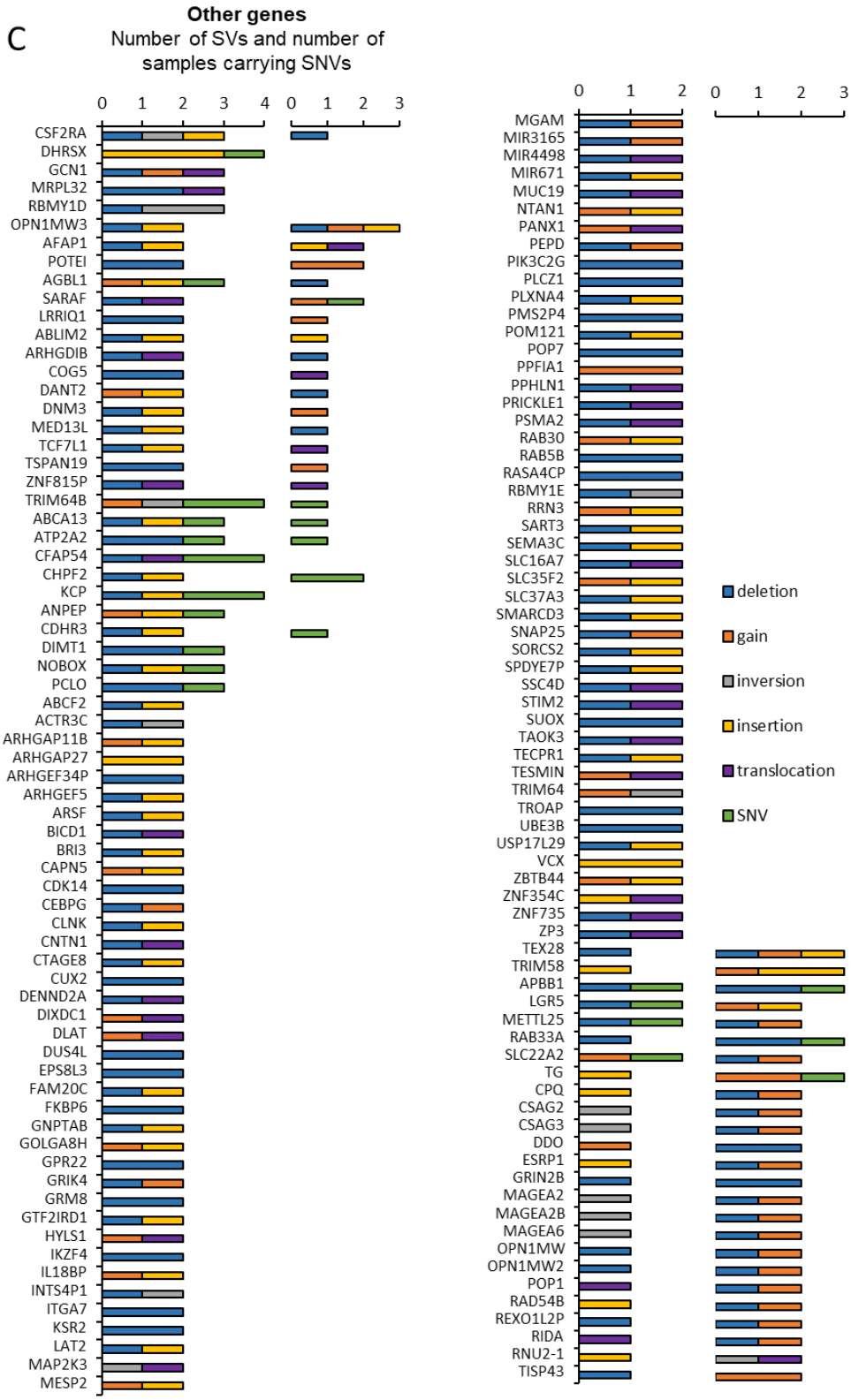


Figure 3- 4. Genes disrupted by structural variants in our cohort.

The number and type of structural variants identified in our cohort of leukemia patients affecting genes (A) previously associated with AML, (B) previously associated with cancer but not AML, and (C) not previously associated with cancer. The left column of each panel represents adult AML patients whereas the right column represents pediatric AML patients, ALL or lymphoma.

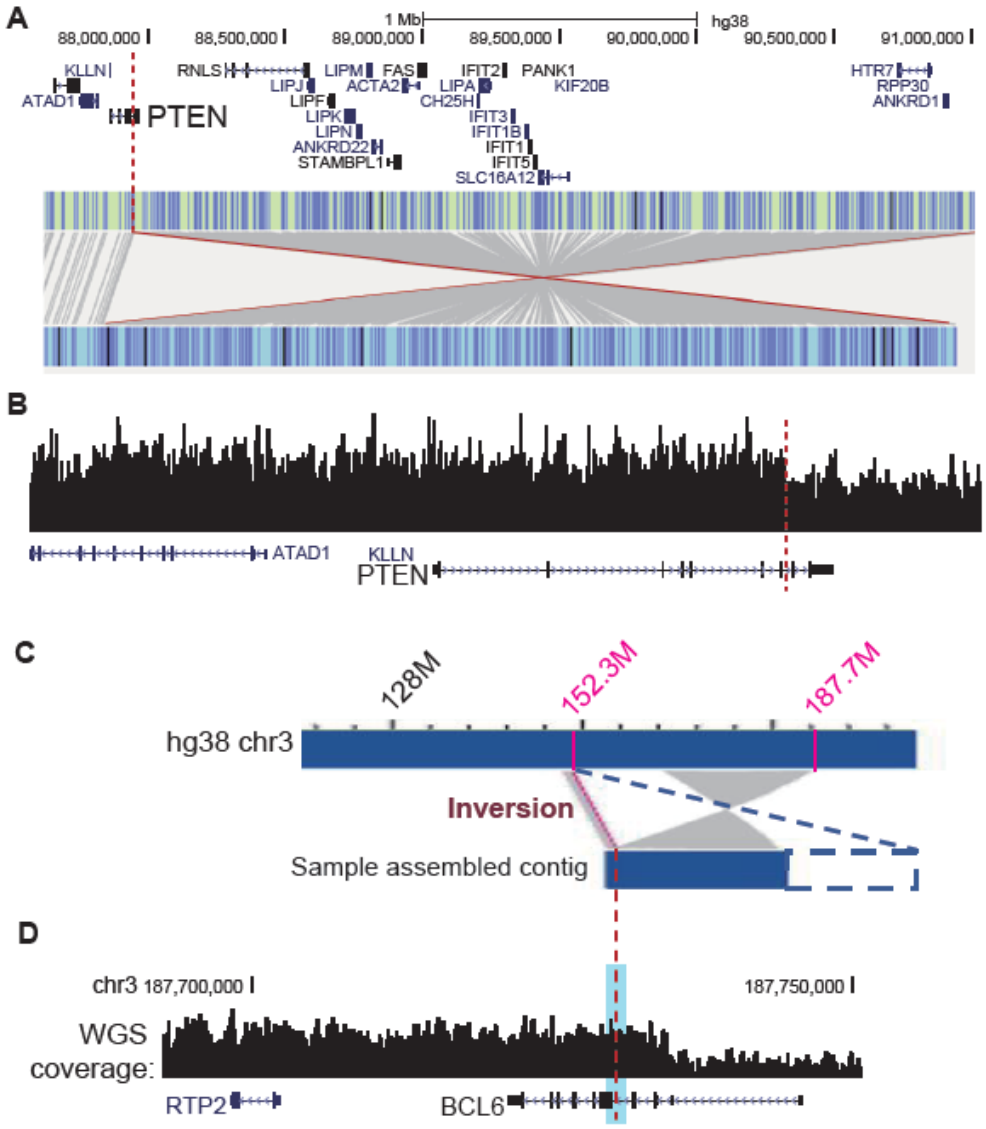


Figure 3- 5. Biallelic disruption of tumor suppressor genes by distinct structural variants

(A) Gene map of the region around PTEN scaled and aligned to the in silico generated optical hg38 reference map (light green with blue tic marks indicating the sites of labeling for optical mapping) under which is shown the optical map of patient sample 1160, indicating the position of a 3 Mb inversion, one endpoint of which lies in the PTEN gene. (B) Whole genome sequence read depth over the PTEN region from patient sample 1160 over the scaled and aligned gene map of the region. Dashed red line indicates the start site of a deletion on the chromosome homolog of that in A. (C) Optical map and whole genome sequence coverage (D) from patient sample 1160 positioned over a gene map of the BCL6 locus on which are indicated (dashed red lines) the 35 Mb inversion break point on one homolog and the deletion breakpoint on the other. The reference genome is shown in the top bar, which, due to compression of the tic marks representing labeling sites is solid blue. WGS coverage level of the BCL6 gene indicates deletion of the last two exons of BCL6 in the second homolog. The current contig contains only one breakpoint of the inversion but does not cover the other breakpoint due to the limited length of the contig. Blue dashed lines represent virtual extension of the contig that is likely the extent of the actual inversion.

Finally, we observed that 109 genes previously unassociated with cancer were affected by SVs in three or more patient samples and 1040 genes affected in two or more patients (**Figure 3-**4C). The *AFAP1* gene is repeatedly mutated in our patient samples. The protein encoded by this gene is a Src binding partner that may function as an adaptor protein by linking Src family members and/or other signaling proteins to actin filaments and by mediating Src activation of TGF-β [177-179]. By extracting copy number values from SNP array data in a TCGA AML cohort of 142 samples, we observed that the *AFAP1* coding region is specifically amplified in the AML cohort, while the immediate surrounding region is unamplified (**Figure 3-**6B). Moreover, using the TCGA data we found that stratifying the patient population on the basis of AFAP1 expression level provides a statistically significant indicator of patient outcome (**Figure 3**-6A). We observed three different types of SVs affecting *AFAP1* in our cohort, so we cannot determine whether loss or gain of function of the gene is driving oncogenesis. However, the TCGA data corroborates that alterations in *AFAP1* have an impact on AML onset and/or progression.

As a second example, the *ENPP2* gene, which encodes the phospholipase autotaxin that catalyzes production of lysophosphatidic acid [180], was altered in three patients. Autotaxin is overexpressed in breast and ovarian cancers but has not been associated with clinicopathologic parameters in those or any other cancers [181]. In examining the TCGA AML database, we observed that the ENPP2 gene but not the surrounding region is amplified in the cohort and that increased gene expression is significantly associated with worse outcomes (**Figure 3-**6A).

As a further example, the zinc finger protein multitype 2 (*ZFPM2*) gene, also known as friend of *GATA2* (FOG2), encodes a transcriptional cofactor of members of the GATA-binding family that regulates expression of key genes essential for the development of multiple organs [182]. By interacting with GATA factors, *ZFPM2* modulates this regulatory activity, and is known to play important roles in

cardiac, gonadal, and pulmonary development. We find that *ZFPM2* was affected variously by deletion, duplication and point mutation in four different patients. As above, we interrogated the TCGA AML database, using the genome wide SNP data to determine copy number levels over and around the *ZFPM2* gene. We found that the coding region but not the surrounding genome was specifically amplified in patient samples and that high expression of the gene was associated with poor outcomes (**Figure 3-**6A). Since *ZFPM2* is a transcriptional cofactor, we extracted from TCGA AML data those genes whose expression is correlated with that of *ZFPM2* (**Figure 3-**6C-D) and showed that those genes significantly overlapped with those bound by *GATA2* and were enriched in proto-oncogenes and those associated with transcriptional misregulation in cancer (**Figure 3-**6E). As shown in **Figure 3-**6F, *GATA2* binds to the promoter of one such gene, *TAL1*, an erythroid differentiation factor [183], which suggests that TAL1 expression may be regulated by *ZFPM2*.

Finally, 15 genes frequently altered in our cohort but not previously associated with cancer, such as *CPQ, COG5, TPD52, AIFM1, RAB33A, ZNF275, TBP* and others, provide prognostic information in the TCGA cohort on the basis of their expression levels (**Figure 3-S**4). In sum, this study has revealed ca. 300 previously unrecognized SVs affecting leukemia associated genes and other cancer associated genes as well as 1040 genes not previously associated with cancer. Outcomes data suggest that some of these newly identified genes could have significant prognostic value.

Structural variants in non-coding regions affect expression of cancer associated genes.

In addition to structural variants that affect the coding region of suspect genes, we observed SVs residing within 1 Mb of genes but not affecting their coding region. Such SVs could alter the expression of the adjacent gene by deleting a cis-acting regulatory element such as an enhancer, by duplicating an

enhancer element or by fusing the gene to a novel enhancer [99, 152]. The cancer genes lying within 1 Mb of an SV in each patient sample are listed in **Table 3-**1.

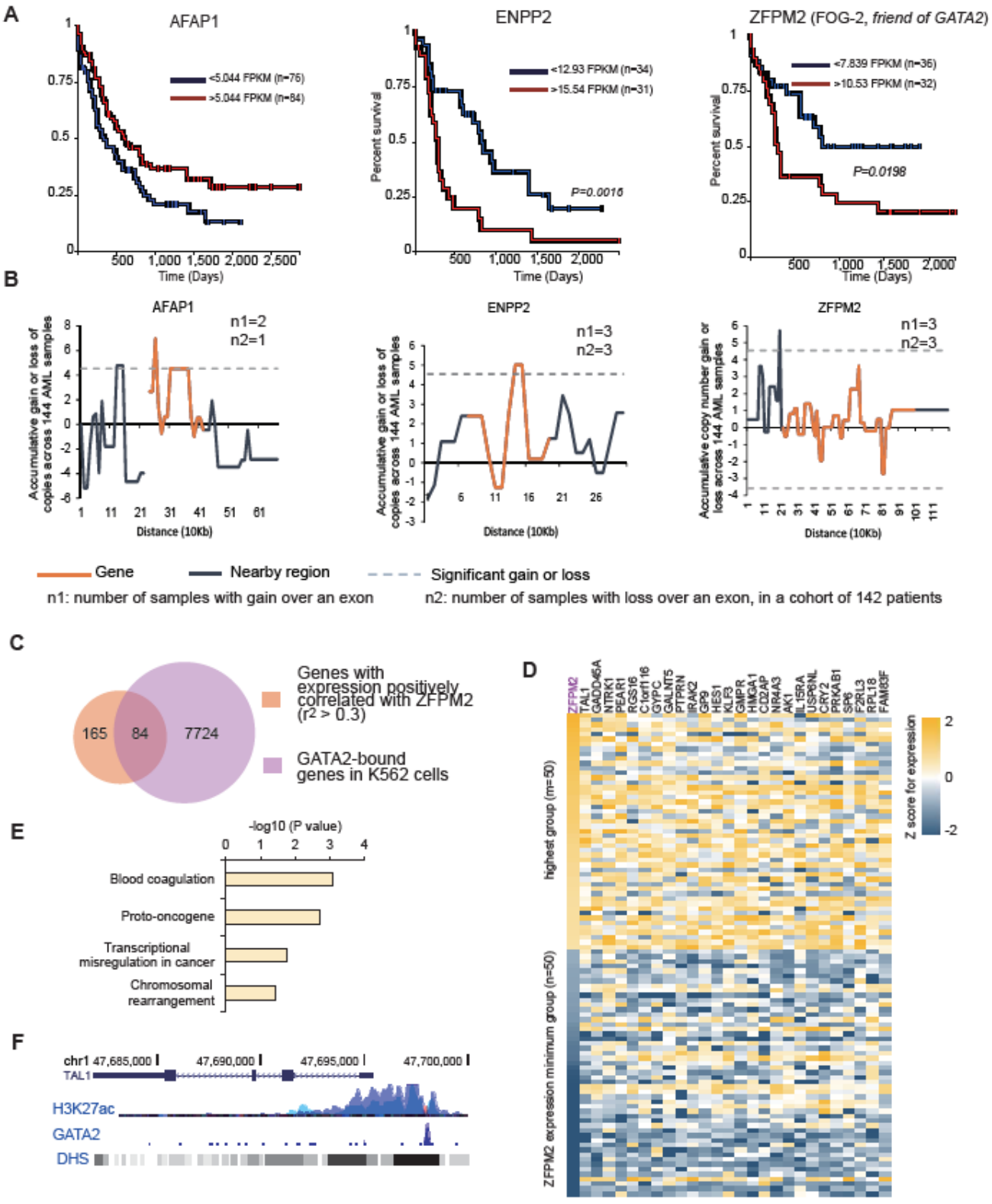


Figure 3- 6. Some genes frequently altered by somatic structural variants affect AML outcomes.

(A) Kaplan Meier survival plots of patients in the TCGA cohort stratified on the basis of gene expression for the indicated gene at the thresholds listed for each gene. (B) Genomic copy number alterations of TCGA AML cohorts. Plotted is the z-score for the variation in average copy number in the AML cohort over each 10 kb bin across the gene of interest (orange line) and the adjacent genomic regions (black line). Dotted line indicated the $p<0.05$ significance threshold. n1 and n2 indicate the number of samples in the cohort in which copy number over the gene was increased or decreased, respectively. (C) ZFPM2 regulated genes overlap those bound by GATA2. Expression of ZFPM2 exhibits positive correlation with that of 249 genes, 84 of which display binding by GATA2 within the gene body or within 10 kb of the gene in the K562 leukemia cell line. (D) Expression heatmap of a subset of the genes whose expression is correlated and bound by GATA2 for the 50 patients in the AML cohort with the highest ZFPM2 and the 50 with the lowest, sorted by ZFPM2 expression levels. (E) The top four David GO term categories of the 84 genes highlighted in (C). (F) GATA2 binds to the TAL1 promoter. Shown are the genome map of TAL1, the H3K27ac and GATA2 abundance and the DNase hypersensitive sites (DHS) over that region in the K562 leukemia cell line. H3K27ac marks promoter domains.

To test whether SVs alter cancer gene expression in our cohort, we performed whole transcriptome analysis of our leukemia samples by RNA sequencing. We then merged our raw expression data with that from the TCGA study, quantile normalized the merged data set and then determined the average gene expression of all genes. We then assessed whether the expression of a cancer related gene lying within 1 Mb of an SV endpoint in our cohort differed significantly from the average expression of that gene over all samples. As evident from the data in **Table 3-**1, 34% of cancer genes lying near somatic SVs exhibited significantly altered expression relative to the combined TCGA cohort or were the highest or lowest expressed sample in our cohort. We predominantly observed overexpression of the cancer gene suggesting that the SV relocated the gene to a new, stronger enhancer or duplicated a preexisting enhancer. In a few cases, we observed reduced expression of the target gene, an unexpected outcome given the expected heterozygosity of the SVs. However, in only one of these cases was the structural genes altered in copy number (**Figure 3-**S5), indicating that for all other genes, altered expression was a consequence of perturbation to a cis or trans-acting regulatory element.

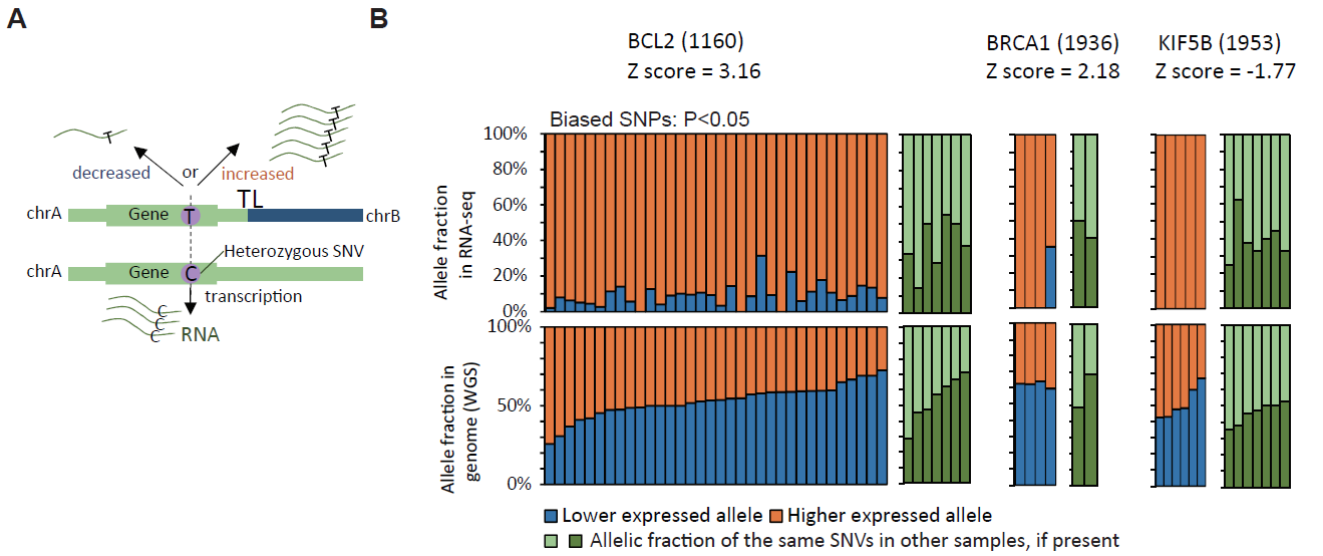


Figure 3- 7. Intergenic SVs affect expression of genes in cis.

(A) Schema for determining allelic imbalance of expression of genes lying near intergenic structural variants. (B) Plotted for the indicated genes in the indicated samples are the allelic fractions for heterozygous SNPs across the gene in that sample as determined by whole genome sequencing (lower panel, ordered by increasing ratios) and by RNA sequencing (upper panel). Each column in orange and blue represents a SNV with significant biased expression, by Chi-square test of RNA read counts and WGS read counts of two alleles. Panels in green are the same data for a patient in which a structural variant does not lie near the gene.

To test more directly whether the altered expression of these cancer genes was a consequence of altered cis-regulatory elements, we calculated the allelic bias of the transcripts from that gene in the affected patient. For each gene in question, we identified from whole genome sequencing single nucleotide polymorphisms in the transcribed region that were heterozygous in the relevant patient's sample and then determined the allelic ratio of those polymorphisms in the RNA transcript sequences in the affected patient sample. If the altered expression were a consequence of the intergenic SVs acting in cis, then we would expect to observe a significant bias in the RNA transcripts, since the variant should affect only one of the two alleles (**Figure 3-**7A). Representative results of that analysis are shown in **Figure 3-**7D and summarized in **Table 3-**1 and **Table 3-S**5. For some of the genes allelic expression could not be determined but for 43% of the genes one allele was predominantly expressed. None of these genes showed biased expression in those samples in which they were not adjacent to an intergenic

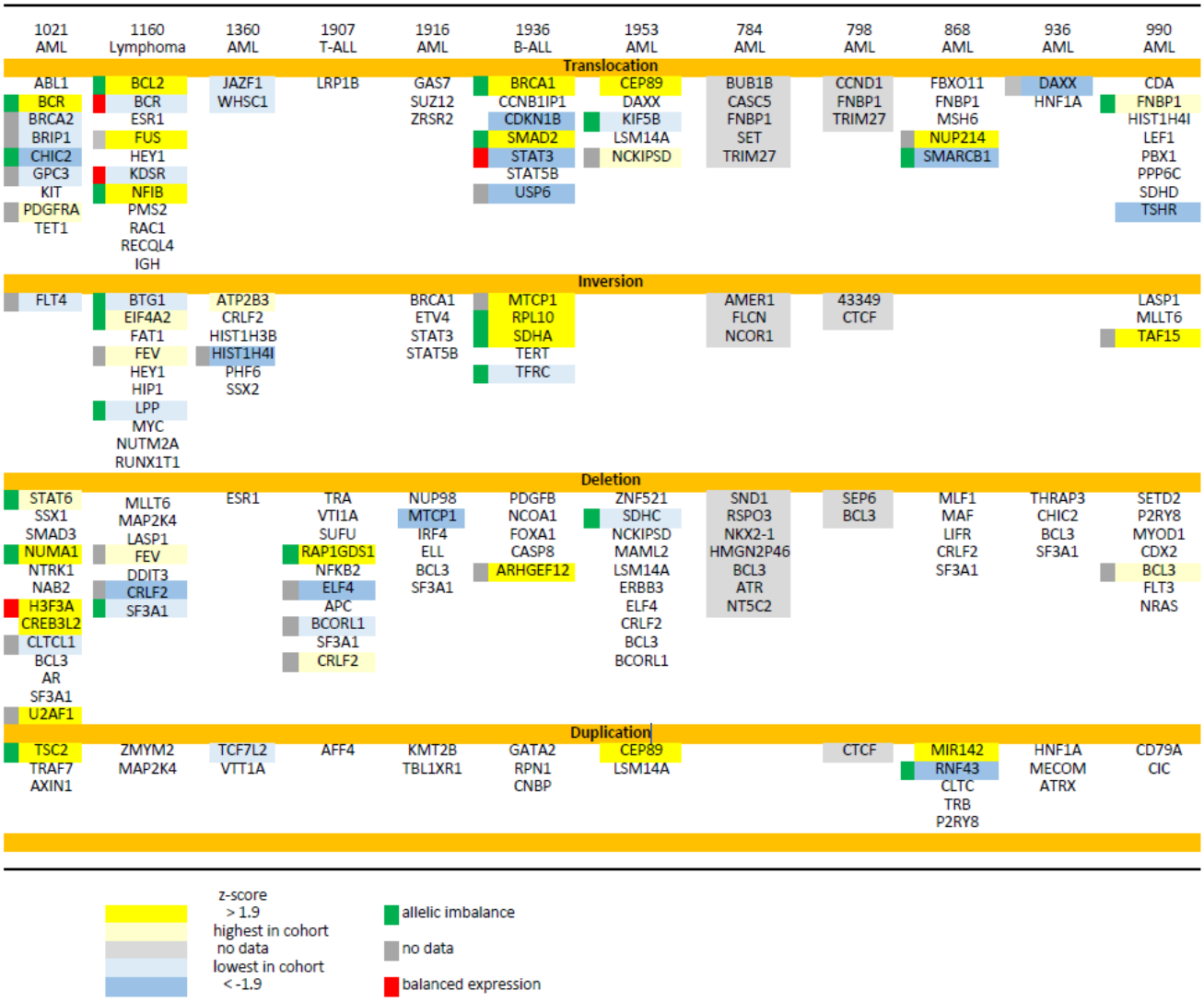
SV (**Figure 3-**7B). For instance, the increased expression of the *BRCA1* gene in patient 1936 comes almost exclusively (3:1 allelic bias) from one allele but the gene exhibits equal expression from both alleles in other patient samples. Similarly, the reduced expression of *KIF5B* in patient 1953 results from attenuation of expression from only one allele. In sum, more than one-third of the cancer genes adjacent to intergenic SVs exhibited significantly altered expression and of those for which allelic bias could be assessed, 85% were expressed predominantly from one allele (**Table 3-S**5). None of these gene with imbalanced expression are naturally imprinted genes. These data demonstrate that intergenic SVs, which are not captured by gene panel or exome sequencing, could play a role in cancer gene expression and their associated role in cancer onset or progression.

Discussion

This report examining leukemia patient samples documents that application of optical mapping in conjunction with whole genome sequencing reveals a large number of SVs unrecognized, and essentially unrecognizable, by conventional genomic analysis, including whole genome sequencing alone. While whole genome sequencing has been extensively applied to cancer genomics and optical mapping has been sporadically applied to a few individual samples or cell lines [184, 185], our study suggests that the combination of the two methods recovers twice as many SVs as revealed by whole genome sequencing alone and our evaluation of the previous cell line study suggest that this combination is adequate to recover the vast majority of SVs [166]. By comparison to datasets of known polymorphic SVs, we could pinpoint those variants that likely arose as somatic alterations. In one case, we were able to confirm the validity of this computational approach by comparing variants identified in the leukemia sample with those of in the patients' normal genomes. Thus, our procedure provides a

facile means to identify somatic SVs in leukemia samples. Moreover, our application of this method to a cohort of leukemia patients revealed reoccurring alterations whose relevance would not be evident from evaluation of single samples.

Table 3- 1. Cancer Genes Adjacent to Structural Variants



Listed for each patient sample are the cancer genes located near each class of structural variants. Those in colored boxes are statistical outliers in the context of the TCGA cohort or the highest or lowest value in our patient cohort. Gray boxes indicate samples for which RNA sequencing could not be performed. The box adjacent to the gene name indicates whether gene expression exhibited allelic imbalance (green), balanced expression (red) or could not be determined (gray).

Our study identified somatic SVs in 37 genes, mutations of which have been previously associated with leukemia. The study also revealed hundreds of genes each affected in multiple patient

samples, some of which had been implicated in cancers other than leukemia and some of which had not been previously associated with any cancers. The role variants in these genes play in leukemia onset and progression certainly warrants further investigation. In particular, we are quite interested in determining the therapeutic value of targeting those genes altered in various leukemia samples. For instance, *ENPP2* overexpression is associated with poor outcomes of AML patients, suggesting that inhibition of the autotaxin phospholipase activity might improve outcomes in a subset of patients.

The previous study on SVs in cancer cell lines documented that deletions led to elimination of enhancers or topologically associating domain boundaries, resulting in altered transcription of associated genes [99, 152]. We have observed similar loss of cis-acting elements in the primary leukemia samples from our patients and have determined that these variants can alter expression of the associated gene. Clearly, determining whether down regulation of expression of these genes attenuates proliferative capacity of the associated leukemia cells would be warranted. For instance, we find that *SMAD2*, an intermediary in TGF-β signaling [186], is upregulated by an intragenic translocation in one of our leukemia samples. *SMAD2* has been shown to be upregulated and over activated in $CD34^+$ BM progenitors from MDS patients. Moreover, pharmacologic inhibition of the TGF-β pathway in vivo, using a small-molecule inhibitor of the TGF-β receptor, *ALK5*, alleviates anemia in a mouse model of MDS [187, 188]. Accordingly, determining whether such pharmacologic inhibitors alter the proliferative behavior of those cells could suggest a novel therapeutic approach for select patients with the disease.

Recent studies characterizing the genomic alterations in AML have generated relative consistent classification systems based on the particular spectrum of driver mutations in a sample [164, 165]. These classifications provide fairly robust prognostic power in predicting the likely outcome of individual patients. Our documentation of SVs provides additional information on the genetic alterations in patients and can refine their classification. Whether this additional information enhances the prognostic

capability of the existing classification schemes will require additional correlation of our SV data with clinical outcomes. However, we reported here that stratification of patients on the basis of expression or copy number of fifteen genes we found repeatedly mutated in our cohort provided a statistically significant difference in outcomes. This suggests that additional studies of previously underappreciated SVs may identify additional useful prognostic markers. Furthermore, these studies offer the potential for providing novel targets for therapeutic intervention.

Supplementary Figures and Tables

Table 3-S 1. Patient Moleucular Diagnosis and Outcome

ID	Patient	Cytogenetics	Other Analysis (FISH, PCR, etc)	*Outcome/ days
784	55yo M with AML	46,XY,t(8;21)(q22;q22)[20]		L/2207
798	63yoF with AML	46,XX, der(7)t(7;11)(q35;q12),inv(16)	+11q23(MLLx3) inv(16); MYH11/CBFB fusion	L/1784
868	29yo M with AML	46,XY,t(9;22)(q34;q11.2)[20]	BCR-ABL p210 mRNA transcript level elevated	L/1636
936	55yo M with AML	44,XY,del(1)(q42), -7, add(8)(p23), -12[18]/44,idem, t(4;5)(p14;q13)[2]		D/246
990	65yo M with AML	46,XY,del(13)(q12q14)[2]/46, XY[18]		L/1417
1021	58yo M with AML	46,XY,t(6;9)(p23;q34), t(8;13;11)(q22;q12;q23)[20]	DEK/NUP214 fusion	L/1374
1160	71yo F with B-cell lymphoma	49,X,der(X)t(X;1)(q13;q11), del(3)(q21),add(4)(q31.3), add(6)(q23),-8, add(9)(p11),add(12)(q22), del(13)(q12q14),add(14)(q32), add(15)(p11.1), -18, +5mar[11]/46,XX[9] Mayo: 13q14/-13, t(11;14), t(14;18), del (17p), MYC (8q24) and t(8;14) MYC/IGH/CEN8	MYC rearrangement, BCL2/IGH fusion	D/39
1361	58yo F with AML	46,XX[20]	FLT3-ITD ASXL1 WT1 mutations	L/437
1907	2yo F with T-cell ALL	46,XX, add(9)(p13)[7]/46,XX[13]	STIL gene deletion with retention of TAL1 (1p32); CDKN2A deletion 9(p21)	L/1342
1916	10yo F with AML	47,XX,+8[16]/46,XX[4]	Trisomy 8	D/397

1936	5yo M with B-cell ALL	46,XY,del(6)(q13q23), add(12)(p11.2)[17]/46,XY[3]	ETV6-RUNX1-positive [fusion]; CNKN2a (9p21) [deletion]	L/976
1953	19yo M with AML	46,XY,t(4;19)(q12;q13.3)[20]	-	L/800

*Outcome is represented by L(alive) or D (deceased) / survival days post diagnosis

Table 3-S 2. Number of polymorphic and somatic SVs by combination of OM and WGS

	Gain/Dup.		**Deletion/loss**		**Insertion**		**Inversion**		**Inter-chr TLs**	
	Somatic	polymor.	Somatic	polymor.	Somatic	polymor.	Somatic	polymor.	Somatic	polymor.
1021	36	202	188	2591	94	2535	4	116	11	0
1160-L	58	187	101	2522	57	2102	11	97	49	0
1360	30	136	82	1768	175	2055	14	138	2	0
1907	11	201	150	2365	100	2367	1	102	5	0
1916	37	174	266	2574	164	2450	2	91	10	0
1936	36	211	235	2632	150	2287	3	114	11	0
1953	27	213	205	2563	105	2259	6	85	10	0
784	26	196	88	2404	78	2340	4	117	9	0
798	10	182	109	2519	90	2205	8	89	7	0
868	29	194	58	2426	48	2203	2	110	8	0
936	30	221	152	2719	87	2298	4	130	24	0
990	30	268	120	2915	73	2195	5	135	21	0
Average	30	199	146	2500	102	2275	5	110	14	0
Percent	13%	87%	6%	94%	4%	96%	5%	95%	100%	0%

Table 3-S 3. Effectiveness of Different Methods to Detect Different Classes of Somatic Structure Variations.

SAMPLES		1021	1160-L	1361	1907	1916	1936	1953	784	798	868	936	990	Average	Percent
Copy gain and tandem duplication	OM+WGS	36	59	30	11	37	36	27	26	10	29	30	30	30.1	100.0%
	WGS only	36	55	29	5	32	30	25	26	9	22	22	28	26.6	88.4%
	OM only	0	2	0	5	4	6	2	0	1	5	5	1	2.6	8.6%
	both	0	1	1	1	1	0	0	0	0	2	3	1	0.8	2.8%
	KA+	0	4	0	2	1	2	0	0	0	0	3	0	1.0	3.3%
Deletion and copy loss	OM+WGS	191	103	82	150	266	235	205	88	109	56	152	120	146.4	100.0%
	WGS only	35	71	9	26	58	48	59	35	39	23	85	77	47.1	32.2%
	OM only	147	24	70	117	201	182	143	48	68	34	62	37	94.4	64.5%
	both	6	6	3	7	7	5	3	5	2	1	5	6	4.7	3.2%
	KA+	0	2	0	2	0	2	0	0	0	0	1	1	0.7	0.5%
Insertion	OM+WGS	97	59	176	101	168	152	104	79	90	50	90	74	103.3	100.0%
	WGS only	10	13	14	13	15	9	13	14	13	15	18	22	14.1	13.6%
	OM only	84	44	161	87	149	141	92	64	77	33	69	51	87.7	84.8%
	both	0	0	0	0	0	0	0	0	0	0	0	0	0.0	0.0%
	KA+	0	6	0	1	0	1	0	0	0	0	1	0	0.8	0.7%
Inversion	OM+WGS	4	11	14	1	2	3	6	4	8	2	4	5	5.3	100.0%
	WGS only	0	2	0	1	0	0	1	0	2	1	0	1	0.7	12.5%
	OM only	4	8	14	0	2	3	5	4	5	1	4	4	4.5	84.4%
	both	0	1	0	0	0	0	0	0	1	0	0	0	0.2	3.1%
	KA+	0	0	0	0	0	0	0	0	1	0	0	0	0.1	1.6%
Inter-chr TL	OM+WGS	11	49	2	5	10	11	10	9	7	8	24	21	13.9	99.4%
	WGS only	7	34	2	4	7	8	5	7	6	6	17	20	10.3	73.2%
	OM only	1	9	0	1	3	2	4	0	1	0	7	1	2.4	17.3%
	both	3	6	0	0	0	1	1	2	0	2	0	0	1.3	8.9%
	KA+	4	1	0	0	0	1	0	1	1	1	1	0	0.8	6.0%

Table 3-S 4. Resolution of "Added" Sequences

Sample	Cytogenetics	OP+WGS Results
BN936	add(8)(p23),	TL chr8-chr12-2,769,428-96,778,624
BN1160	add(4)(q31.3)	TL chr3-chr4-148,259,617-156,883,828
	add(6)(q23)	chromothripsis chr6-chr8-chr18
	add(9)(p11),	chromothripsis chr9-chr6-chr8
	add(12)(q22),	Inv chr12-chr12-40,647,772-91,231,824
	add(14)(q32),	TL chr14-chr18-105,864,250-63,104,165
	add(15)(p11.1),	TL chr15-chr18-28,135,947-2,823,424
Peds 1907	add(9)(p13)[7]	TL chr6-chr9-43,687,796-33,130,539
Peds 1936	add(12)(p13.2)[17]	TL chr12-chr21-11,885,827-34,910,564

Karyotyping of the listed samples documented that extra material was added to a chromosome, specified as "add(chr)(location)," without determining the source of the added material. The mapping results obtained in this study by OP + WGS clarified the source of the added material and the specific junction site on each of the participating chromosomes.

Table 3-S 5. Allelic Imbalanced Expression of Cancer Genes Adjacent to Structure Variations

Gene	Patient	SV	z-score across TCGA samples	Z-score across our samples	RNA fraction of the predominant allele (number of significant SNPs, p<0.05)
Significantly higher or lower across TCGA samples and same trend in our cohorts					
BCR	1021	TL	2.18	1.51	0.79 (6)
CHIC2	1021	TL	-2.10	-0.34	0.93 (2)
NUP214	868	TL	3.43	1.40	NA
SMARCB1	868	TL	-3.53	-2.10	1 (7)
DAXX	936	TL	-1.92	-1.45	NA
BCL2	1160	TL	3.16	2.06	0.89 (30)
FUS	1160	TL	2.91	0.86	0.61
NFIB	1160	TL	2.38	1.92	1(8)*
BRCA1	1936	TL	2.18	1.56	0.81 (4)
SMAD2	1936	TL	4.92	1.20	0.76 (3)
STAT3	1936	TL	-2.34	-2.49	0.51
USP6	1936	TL	-1.95	-1.35	NA
NUMA1	1021	DEL	3.31	1.76	0.82 (1)
CRLF2	1160	DEL	-2.50	-1.87	NA
RAP1GDS1	1907	DEL	3.43	1.91	0.84 (3)
ELF4	1907	DEL	-4.16	-2.08	NA
ARHGEF12	1936	DEL	2.57	1.93	NA
TSC2	1021	DUP	2.60	2.03	1 (1)
RNF43	868	DUP	-3.38	-2.13	1 (1)
MTCP1	1936	INV	2.47	0.76	NA
RPL10	1936	INV	2.15	1.20	0.75 (2)
SDHA	1936	INV	3.52	0.84	1 (2)
HIST1H4I	1361	INV	-2.32	-1.10	NA
Highest or lowest expression across our samples					
BRCA2	1021	TL	-0.99	-1.81	NA
BRIP1	1021	TL	-1.39	-1.61	NA
GPC3	1021	TL	-0.91	-1.96	NA
PDGFRA	1021	TL	0.96	1.00	NA
BCR	1160	TL	-0.27	-1.30	0.54
KDSR	1160	TL	0.06	-1.10	0.53
KIF5B	1953	TL	-1.77	-0.88	1 (5)
NCKIPSD	1953	TL	-0.09	1.24	NA
FNBP1	990	TL	1.56	1.34	0.77 (3)
JAZF1	1361	TL	-1.54	-1.31	NA
WHSC1	1361	TL	0.51	-1.33	NA
U2AF1	1021	DEL	2.07	-2.12	NA
SDHC	1953	DEL	1.49	-2.30	1 (3)
STAT6	1021	DEL	1.02	1.52	0.82 (2)
H3F3A	1021	DEL	2.14	-1.39	0.52
CLTCL1	1021	DEL	-0.70	-1.32	NA

SF3A1	1160	DEL	-1.03	1.32	0.90 (3)
FEV	1160	DEL	0.80	1.41	NA
BCORL1	1907	DEL	-1.15	-1.31	NA
CRLF2	1907	DEL	0.99	1.35	NA
BCL3	990	DEL	-0.66	1.04	NA
TCF7L2	1361	DUP	-0.56	-1.20	1 (2)
FLT4	1021	INV	-0.58	-1.12	NA
BTG1	1160	INV	1.18	2.00	0.72 (1)
EIF4A2	1160	INV	1.69	1.72	0.721533258
FEV	1160	INV	0.81	1.42	NA
LPP	1160	INV	0.78	-1.62	1 (1)
TFRC	1936	INV	-0.55	-1.69	1 (1)
TAF15	990	INV	2.21	1.32	NA
ATP2B3	1361	INV	1.57	1.55	NA

NA	no heterozygous genomic SNPs available or too few RNA reads
X	Balanced expression
*	Remarkable imbalanced expression but statistically not significant due to limited RNA reads

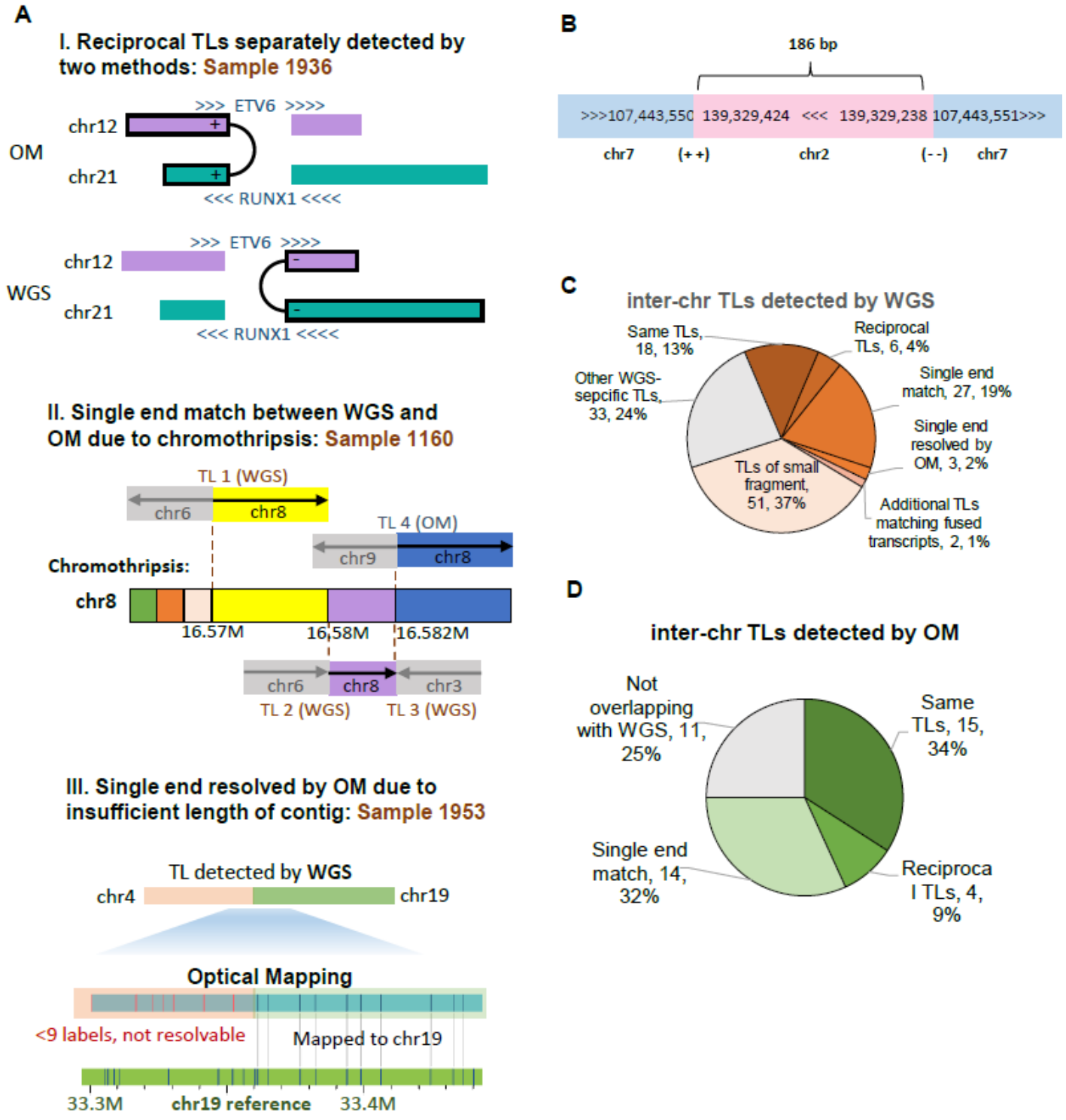


Figure 3-S 1. . Overlap of inter-chr translocation calls between WGS and OM at different levels.

A. Three models that WGS- and OM-specific calls provide evidence for mutual confirmation. I. WGS and OM detects TLs with similar positions but inconsistent orientations, representing the possible existence of reciprocal

TLs, each component detected by one method, exemplified by sample 1953. Fused RNA transcripts of both orientations are detected. II. WGS and OM detects TLs with one end similar and the other end distinct, also with varied orientations. This is prevalent in two samples 936 and 1160 that harbor chromothripsis, where many small DNA fragments shattered from a local region join into various chromosomes and regions across the whole genome. III. One end of the WGS-specific TL is consistent with the breakpoint identified by OM, but OM is not able to resolve the other part of TL, because of insufficient length of the contig. OM requires at least nine labels (>80Kb) to uniquely map one arm of a TL. **B.** An example of WGS detects TLs of small DNA fragment (<100Kb) that are undetectable to optical mapping, indicated by a pair of TL breakpoints. **C** and **D**. Overlap of inter-chr TL calls between the two methods. Same TLs refer two SVs with consistent loci of both ends and concordant orientations. Reciprocal TLs, single end match and single end resolved by OM refer to models I, II and III in A. Fusion of transcripts from RNA-seq data are also used to validate additional TLs that are not confirmed by any other situations. TLs of smaller fragment refers to the situation described in B.

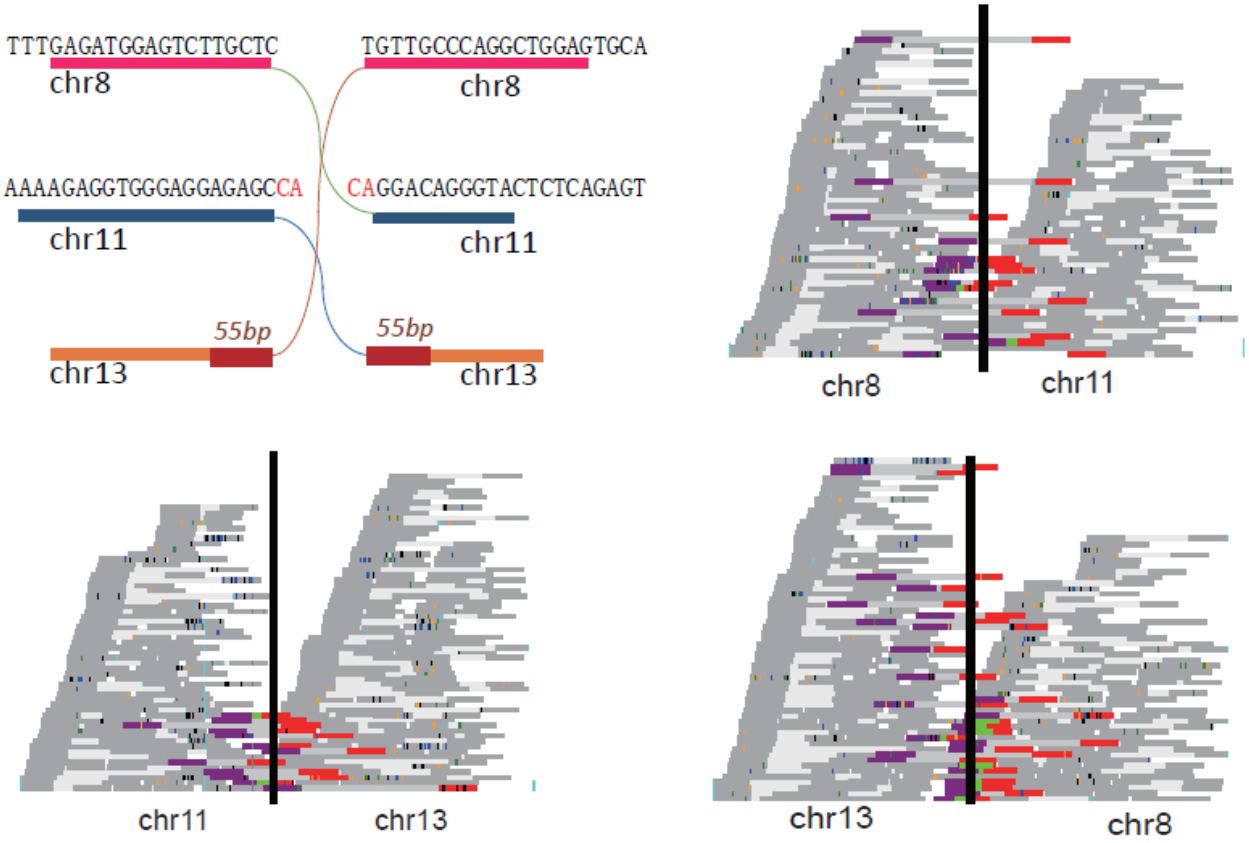


Figure 3-S 2. Three-way translocation identified by OM/WGS

Diagram of the organization of an unusual three way reciprocal translocation identified in sample 1021 (upper left). The breakpoints of all three junctions are precise: one junction (chr8-chr11) exactly joins two chromosomes without loss or gain of sequences, one with a 2 bp duplication (chr11-chr13), and the other with a 55 bp duplication (chr13-chr8). Pileups of paired-end short read sequences around the breakpoints are shown the other three panels, with each horizontal line representing a fragment on which the dark grey lines represent the sequenced ends of the fragment and the intervening light grey line the inferred unsequenced segment separating them. Reads crossing the boundary are indicated either by dual red/purple colored bars in which the paired ends of a single fragment map to different chromosomes or by tricolored bars in which the green region represents a sequenced segment spanning the junction.

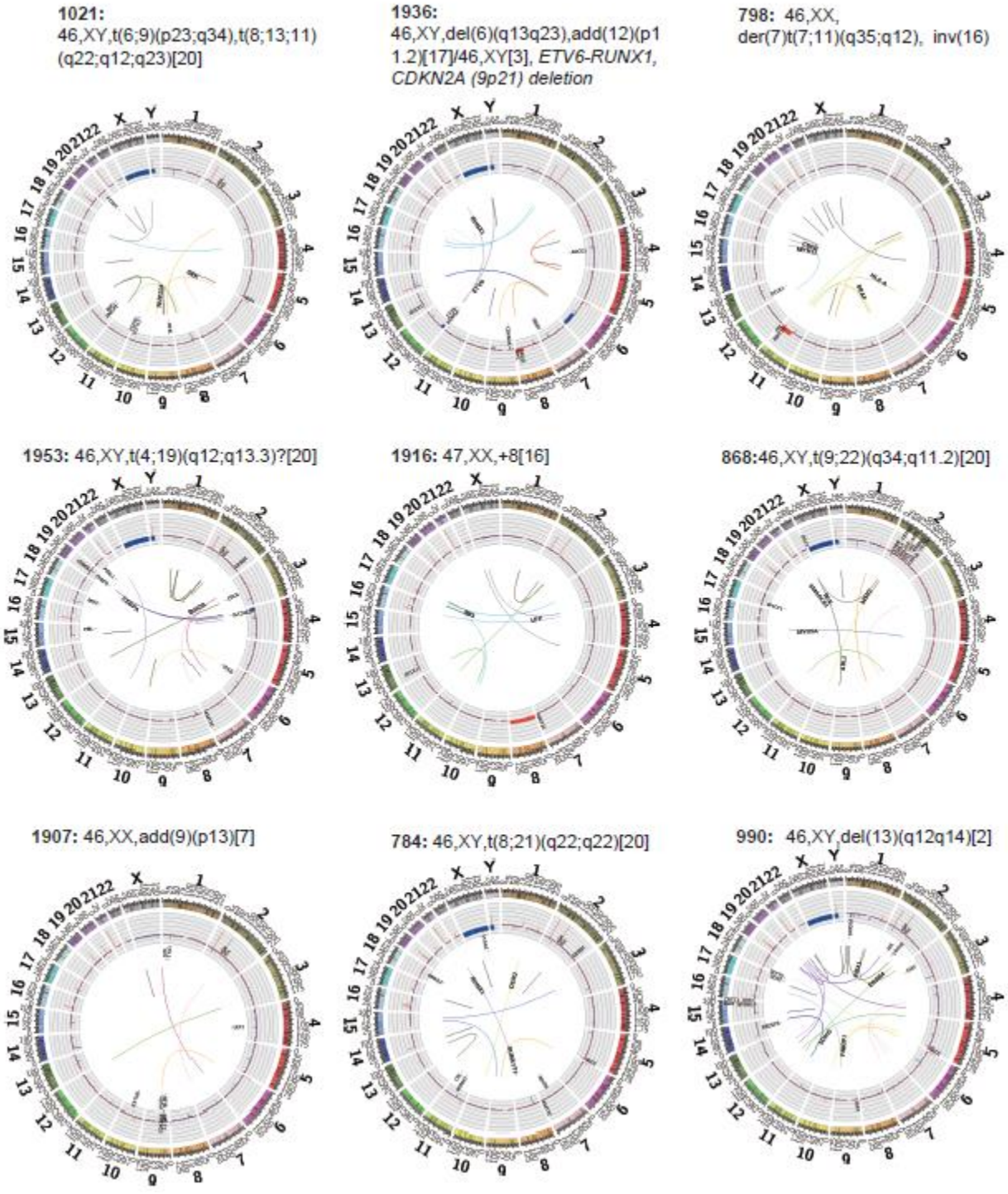


Figure 3-S 3. Comparison of karyotyping and structure variation determination in the samples of this study.

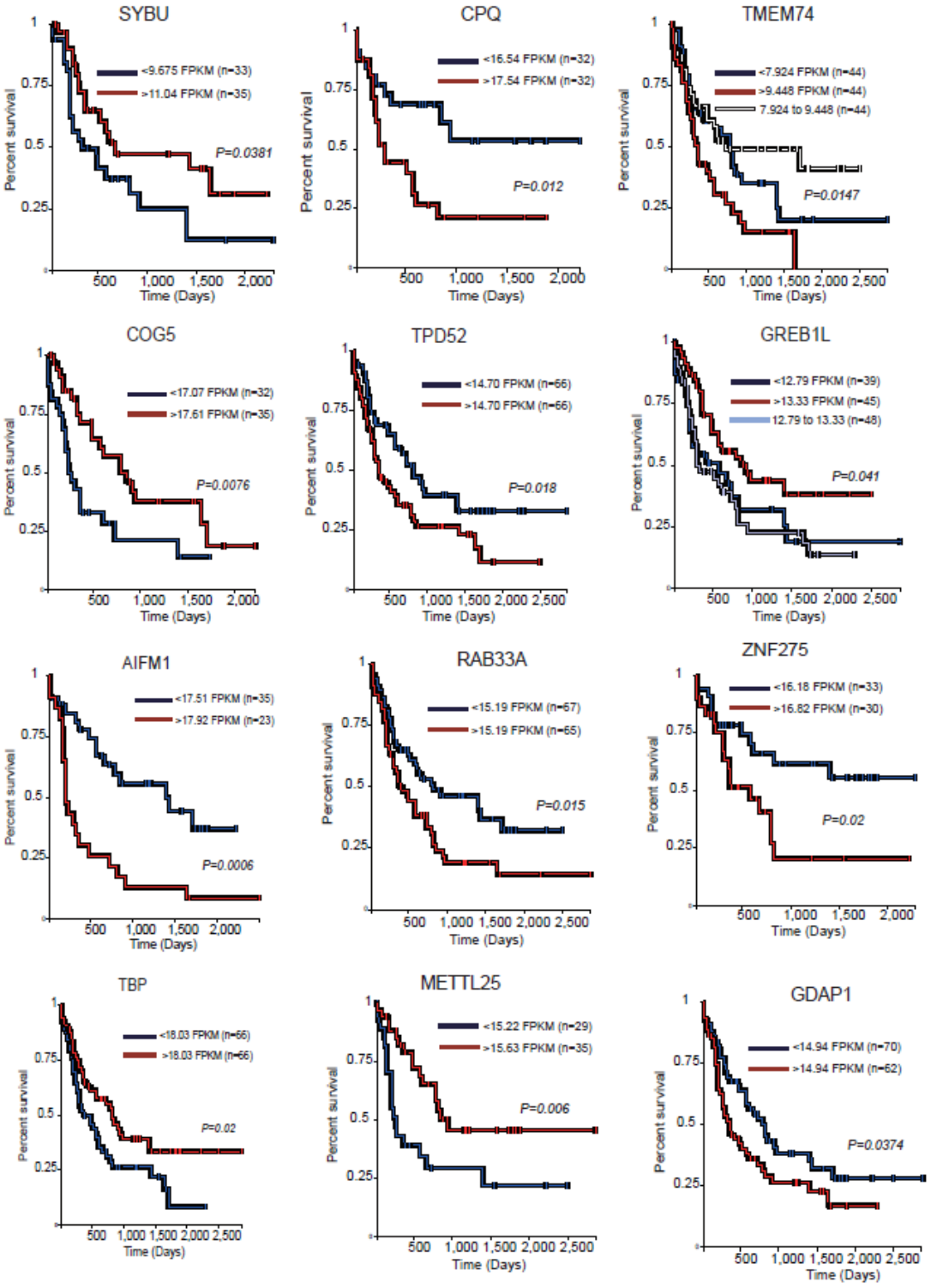


Figure 3-S 4. Survival data stratified by expression levels of genes identified in our cohort.

Figure 3-S 5. Copy number of cancer related genes within 1Mb to SVs in this study.

Materials and Methods

Materials and Experiments

Patient Samples

Bone Marrow (BM) aspirates or Peripheral Blood (PB) samples were obtained from AML patients, after informed consent using protocols under the Penn State Hershey IRB-approved protocol PRAMS Y00-186 or protocol PRAMS 40532. Mononuclear cells (MNCs) were isolated by density gradient separation (Ficol-Paque, GE Healthcare Life Sciences, Pittsburgh, PA) and frozen for later use. Anonymized adult leukemia samples were obtained from the Penn State Hematology/Oncology Biobank. Anonymized pediatric leukemia samples were obtained from the Pediatric Hematology/Oncology Biobank. Patient clinical and demographic data are summarized in **Table 3-S**1.

Cell culturing

T-cell Expansion from Patient PBMCs. T-cell expansion was performed using the Miltenyi T Cell/Activation Expansion Kit (130-091-441) according to the manufacturer's instructions. T-cell activation beads were prepared prior to thawing patient cells and stored at 4°C until use. 100µl of CD2-Biotin, CD3-Biotin, and CD28-Biotin were added to 500µl of anti-Biotin MACSiBead particles. MACSiBead buffer (0.5% human serum albumin and 2mM EDTA in PBS, PH 7.2, 200µl) was added to the mixture and the beads rotated slowly at 4°C for two hours. PBMCs from the patient were thawed in TEXMACs media with 1% Pen/Strep and counted using an automated cell counter. T-cell activation beads were added to the PBMCs at 25µl/5x10^6 cells. IL-2 was added 96 hours later at 0.8µl IL-2/ml of media. Cell cultures were monitored and IL-2 media was added as needed. On day 14 post-thaw, additional T-cell activation beads were added to the culture.

T-cell Sorting. Cultured cells $3.0x10^7$ were harvested by centrifugation and resuspended in 2ml PBS to which was added 5µl of CD3-APC (BD Biosciences - 340661) and 20µl of CD33-PerCPCy5.5 (BD Biosciences - 341640) followed by incubation in the dark for 15 minutes at room temperature. CD3 positive/CD33 negative cells were recovered by sorting on a BD FACS Aria Sorter II.

Optical Mapping

DNA extraction and nickase labeling were performed as described previously [166]. For direct labeling of genomic DNA, 750ng of gDNA was mixed with DLE-1 Buffer (Bionano Genomics, Part#20350), DL-Green (Bionano Genomics, Part#20352), and DLE-1 Enzyme (Bionano Genomics, Part#20351) and incubated for 2 hours at 37°C in a thermocycler. Proteinase K solution was then added to the reaction and incubated for 30 minutes at 50°C. Finally, a DLS-membrane (Bionano Genomics, Part#20358) was placed upon 60□L of DLE-1 buffer in one well of a DLS-microplate (Bionano Genomics, Part#20357). DNA was transferred onto this membrane, incubated at room temperature for one hour, transferred onto another membrane with DLE-1 buffer and incubated for 30 minutes at room temperature. DTT (Bionano Genomics, Part#20354), Flow Buffer (Bionano Genomics, Part#20353), and DNA stain (Bionano Genomics, Part#20356) were added to the DNA in an amber tube and the tube was mixed at 5rpm for 1 hour at room temperature and then stored in the dark overnight at room temperature. Each labeled sample was added to a BioNano Saphyr Chip (Bionano Genomics, Part#20319) and run on the Bionano Saphyr instrument, targeting 100X human genome coverage.

Library Preparation and DNA Sequencing

DNA libraries were prepared from DNA extracted from PBMC or blood using according to the KAPA HyperPrep PCR-free Kit (Roche). Illumina NovaSeq S2 150 bp paired-end sequencing was performed to achieve 40X genome coverage. RNA libraries were prepared from total RNA following

rRNA depletion with KAPA RNA HyperPrep Kit RiboErase according to manufacturer's instructions (Roche). Illumina NovaSeq 50 bp paired-end sequencing was performed to obtain 50 million raw reads per library.

Data Analysis

Variant detection and filtration from WGS results

SV and SNV detection: We used two pipelines to independently identify SVs. The first pipeline uses BWA-MEM (v0.7.15-r1140) [189] to align the paired-end reads to human reference genome GRCh38 (version GCA000001405.015). Duplicated reads were removed by Sambamba (v0.6.6) [190]. Reads with mapping quality ≥ 20 were retained for downstream SV calling by Delly (v0.7.7) [116], which reports SVs as deletion, inversion, insertion, tandem duplication or inter-chromosomal translocation. SVs were also independently detected by the Speedseq pipeline (v0.1.2) [191], in which paired-end reads were aligned to the same GRCh38 reference genome with BWA-MEM. Duplicated reads were removed by SAMBLASTER (v0.1.24) [192]. SAMBLASTER then extracted discordant read pairs and split reads for downstream SV detection, which was accomplished by Lumpy (v0.2.13) [115] with default parameters. During the SV detection, Delly and Lumpy exclude a list of telomeric, centromeric, and 12 heterochromatic regions provided by the Delly software (https://raw.githubusercontent.com/dellytools/delly/master/excludeTemplates/human.hg38.excl.tsv).

Copy number variants (CNV) detection: were detected by Control-FREEC (v11.0) [117] with the following parameters "breakPointThreshold = 0.8, coefficientOfVariation = 0.062, ploidy = 2". Control-FREEC normalizes copy numbers for genome GC contents, mappability, and ploidy. Copy number profile for each 50 kb bin of the genome was used for making Circos plots.

SNV detection: SNVs were detected using FreeBayes (version: v0.9.21-19-gc003c1e, included in SpeedSeq pipeline (version: 0.1.2)) (https://arxiv.org/abs/1207.3907) with the following parameters "—min-repeat-entropy 1". Low-quality SNVs (QUAL field < 20) were removed for downstream analysis. SNVs were annotated with SnpEff (version 4.3) [193] using default parameters and filtered for potential protein altering variants (annotated as high/moderate putative impact). This filtered SNV set was then compared against common SNPs (dbSNP150 with allele > 0.01) to keep only potential somatic SNVs.

SV filtration and classification: We employed the following criteria to filter SVs detected by WGS: SVs had to be 50 bp or greater, could not map to chromosome Y or the mitochondrial genome and had to be supported by at least 10 reads combining spanning paired-end reads (PE) and split reads (SR) and an additional 2 split reads. SVs calls from Delly and Lumpy were merged to form a consensus call. Merging criteria differed depending on the type of SVs. A deletion or duplication was merged if the overlap of the size and location of the SV as determined by the two methods was greater than 50% of that as determined by either method alone. Deletion coordinates determined by Lumpy were used for the merged call set. An inversion was merged if the overlap of the size and location of the SV as determined by the two methods was greater than 90% that as determined by either method alone and Lumpy coordinates were used. Translocations were merged if both break-point ends mapped within 50bp of each other and if the strand of the break-point ends matched. Final translocation coordinates were based on Lumpy calls. Coordinates for insertions were obtained from Delly since Lumpy does not detect insertions.

We merged deletions detected by LUMPY/DELLY and loss of copies detected by Control-FREEC to form a non-redundant list of "deletions". Similarly, we merged duplications detected by LUMPY/DELLY and gain of copies detected by Control-FREEC, removed redundant ones, and defined

the overlapped ones as “duplications.” For SVs detected by both LUMPY/DELLY and Control-FREEC, we use the breakpoints provided LUMPY/DELLY.

We excluded inter-chromosomal translocations that were also found in a human normal cell line (GM12878) in order to remove likely polymorphisms. WGS data for GM12878 were downloaded from European Nucleotide Archive (Accession number: ERR194147) and analyzed from SVs by the same aforementioned pipelines. We also removed inversions in each patient sample that share a RO ≥ 99.9% with inversions detected in GM12878. We removed inter-chromosomal translocations whose both break-point ends are within 50 bp in any two individuals, since intra-chromosomal translocations that are shared between two individuals at the nearly same location are likely to be polymorphisms or false positive.

SV detection and filtration from optical mapping results.

De novo assembly, SV detection and SV classification. We performed de novo assembly of cancer genomes using long optical mapping molecules, from which we identified SVs by comparing the generated cancer genome to the reference genome GRCh38, using software BioNano solve 3.1.1 with RefAligner and pipeline 7196/7224. DNA molecules used for assembly met the following criteria: length >150Kb and spanning at least nine labels, with a signal to noise ratio higher than 2.75 and backbone intensity lower than 0.6. Parameters used for de novo assembly and SV detection are the same as described in the method section in our previous work [166].

Raw SV output comprises deletions, insertions, inversion, duplications and translocations, which include interchromosomal translocations and any intra-chromosomal translocations that are larger than 5Mb. We ran software smap2vcf to convert SV output to VCF format to determine orientation and then

separated intra-chromosomal translocations into deletions (5'->3') and inversions (5'->5' and/or 3'->3') according to their orientation.

Filtration of detected SVs. We removed all SVs that were smaller than 50bp and all intra-chromosomal SVs with confidence score smaller than 0. We further removed false positive SVs generated due to technical bias such as similar labeling pattern of distinct genomic regions, which results in misalignment and misidentifications of SVs. We also removed large identical SVs (defined as >99.99% overlap) that were found in more than one sample, since identical somatic SVs are unlikely to repeatedly occur in a small collection of samples. We removed deletions overlapping genomic gaps, which represent correction of gap size of the reference genome rather than true deletions. Finally, we generated a list of false-positive translocations and inversions from our previous work and we removed SVs whose breakpoints that are within 500Kb to these previously identified SVs.

Integration of SVs from optical mapping and WGS.

We integrated SV calls to combine SVs independently identified by both methods into a single call and to represent each SV with breakpoints of highest resolution available. WGS provides SVs breakpoints with base pair resolution, while optical mapping provides only the nearest labeling site to the left and right of the SV (SV interval) instead of its start and end. We therefore set the following criteria for determining whether SVs independently detected by optical mapping and WGS refer to the same event: 1) Deletions, insertions and duplications detected by WGS must overlap at least 50% with the SV interval demarcated by optical mapping and the difference in size predicted by the two methods must be less than 30%. 2) For translocations and inversions, the breakpoint detected by WGS must lie within 500Kb to that detected by optical mapping and the orientation of the SV determined by the two methods should be consistent. If a copy number gain matches duplications found by optical mapping, we

specified the SV to be a duplication. 3) A duplication detected by WGS can also be an insertion detected by optical mapping if the duplication is completely within the insertion range and both share similar sizes, with a difference of size smaller than 30%. All SVs detected by both methods are represented by the breakpoints obtained by WGS. Finally, we combine the SVs detected by both methods and those only detected by one method, to generate a union of SV calls without redundancy.

Determination of somatic SV mutations

We used several filtering strategies to distinguish between polymorphic SVs and somatic mutations. First, we compared our deletions, duplications, copy gain and inversion with corresponding SV type in the database of genomic variations (DGV, hg38 updated on 09-06-2016) [167] with the stipulation that an SV that appeared in at least five individuals in the DGV is a polymorphism. We removed from our somatic mutation list any SV that overlapped at least 50% with a DGV polymorphic SV with less than 30% difference in size. Second, we removed SVs with identical sizes and positions in any two or more of our samples or in the NA12878 cell line [166]. Third, we removed SV calls matching any identified in the UCSF optical mapping dataset of polymorphisms [168]. We interpreted an SV detected by optical mapping in three of the 150 normal individuals in the study to be a polymorphism. Fourth, we removed SV calls that match any observed in the BioNano Genomics control dataset, which is publically available within the Bionano Access software download (https://bionanogenomics.com/support/software-downloads/) at /pipeline/Solve3.3_10252018/VariantAnnotation/10252018/config.

Circos profiling of leukemia genome

Leukemia genome profiles of each samples were generated using Circos [194], which includes three tracks: copy number variation genome-wide, deletions and duplications, and inversions and

translocations. We used copy number at 50Kb bin size measured by Control-FREEC. The SVs we plotted were the integrated union from WGS and optical mapping. We display genes that are directly overlapping with deletions and copy gains in the outer track. For inversions and translocations, we set a buffer zone of 50Kb to represent to possible position of SV breakpoint detected by optical mapping. We display genes that are overlapping with the possible position of breakpoints of translocations and inversions in the inner track.

Comparing SVs to Karyotype

We defined an SV detected by our method as identical to that identified by karyotyping if 1) the position of the SV detected by optical mapping or WGS corresponds to that provided by karyotyping, demarcated by chromosome and the band on the p or q arm; 2) the SV detected by our method is larger than 1Mb, which would be of sufficient size to be detected by karyotyping; and 3) the type of SV is consistent between methods: deletion or copy loss in our method corresponds to “del” or “-” in cytogenetics; inversions correspond to “inv()”; translocations or insertions correspond to “t()”, “der()” or “ins()”; gain of copies or polyploidy correspond to “+”. Complex forms of copy gain such as fragment duplication, inverted duplication or translocated duplications are generally identified as “add” in karyotyping.

Identification of frequently disrupted genes

We intersected RefSeq gene exons (GRCh38) with somatic SVs we detected and considered a gene disrupted if 1) part or all of one or more exons overlaps any part of a deletion, loss or gain of copies, or duplications; 2) the breakpoint of an inversion or inter-chromosome translocations lies within the gene; 3) the coding region carries an indel or SNV resulting in nonsense, frameshift or missense mutation or a splicing sequence alteration. Genes inside of an inversion but not interrupted by the

breakpoint are not considered disrupted. We divided genes into three exclusive groups based on data from COSMIC (https://cancer.sanger.ac.uk/census): 86 AML driver genes, 534 other general cancer-related genes, and 23631 other genes without clear evidence for association to cancer.

Outcomes analysis

For each novel gene frequently disrupted by somatic mutations, we examined whether its copy number or gene expression correlated with disease outcome. Kaplan Meier survival plots were constructed from clinical outcomes data from GDC AML patient cohorts (https://xenabrowser.net/datapages/; https://portal.gdc.cancer.gov/projects/TCGA-LAML). Patients were stratified on the basis of gene expression or gene copy number evenly into two groups, one containing half of the cohorts with above the average expression/copy number of the gene and the other group containing the half of cohorts with below average expression/copy number of gene.

Copy number linkage analysis

We obtained copy number variation of TCGA AML cohort (n=142) profiled by SNP array from GDC (https://portal.gdc.cancer.gov/projects/TCGA-LAML) , segmented the genome into 10Kb bins and use an in-house pipeline to calculate the average CNV from all patient for each 10Kb bin. We calculated the Z score and the corresponding P value for each bin genome-wide and used that to set thresholds above or below which represents significant gain or loss across the AML cohort.

Simulating distances between SVs and cancer-related genes

Using 86 previously defined AML-driver gene [164, 165] and 535 additional cancer-related genes from COSMIC (https://cancer.sanger.ac.uk/census), we calculated the number of such genes within a specific distance interval to the nearest SV for each SV subtype in patient samples as well as

gene density (genes per Mb) . Genes that directly overlap an SV were excluded. We then permuted the distance distribution by fixing the positions of the SVs and randomly distributing the positions of the list of genes and then calculated the number of genes within specific distance intervals to the nearest SV after each individual permutation. In detail, the entire gene body is moved in the permutation, so the gene size affects the result. We measure the shortest distance between a closest SV to the nearer end of the gene. The simulations were run one thousand times to generate a distribution of expected number of genes and gene density for each distance interval. We then calculated the Z-score and P value of the actual gene density by comparing to the distribution of expected gene density for each interval.

RNA-seq data processing

RNA-seq reads were processed using the ENCODE standard RNA-seq processing pipeline (https://github.com/ENCODE-DCC/long-rna-seq-pipeline). Briefly, raw RNA-seq reads were mapped to human genome reference GRCh38 (version: GRCh38_no_alt_GCA_000001405.15) with STAR (v2.5.3a_modified) [195]. Mapped reads were quantified and aggregated at gene level by RSEM (v1.2.31) [196]. FPKM values for each gene were used for downstream analysis. To investigate the level of gene expression of our patient samples in general AML populations, we downloaded gene expression for two AML cohorts from TCGA (https://portal.gdc.cancer.gov/projects/TCGA-LAML, https://portal.gdc.cancer.gov/projects/TARGET-AML). We then performed quantile normalization for FPKM values across patient sample in this study and TCGA cohorts to eliminate batch effects. To quantify the level of gene expression in this study in general AML population, we calculated the Z-score for each gene on the log-transformed FPKM values relative to the average FPKM value for that gene in all samples both across only our leukemia and across our samples plus the TCGA cohort.

Allelic gene expression analysis

We processed the bam files generated from RNA-seq with the WASP pipeline [197] to correct bias towards certain SNV alleles, which can be introduced during mapping. In running WASP, we input SNVs detected from WGS, and WASP outputs a new bam file with bias removed. We then identify SNV on the newly generated bam file using samtools mpileup.

We then pick all the heterozygous SNVs that appeared in WGS data, by the criteria that the allele ratio between reference and alternative alleles should be between 0.333 and 3. We examine the allele fraction of the same loci in RNA-seq data and performed chi-square test on the basis of the allele fraction from WGS and RNA-seq. For each gene, we counted the number of significant SNVs, and we calculated the expression percentage contributed by the dominant allele, normalized by the allele fraction in WGS.

Data Access

Raw and aligned next generation sequencing files have been submitted to the European Genome-Phenome Archive (EGA; https://www.ebi.ac.uk/ega) within study accession EGAS00001003431. Bionano variant calls and mapped reads for our samples can be downloaded from https://research.med.psu.edu/departments/personalized-medicine/publications/.

Chapter 4

Subtype-specific and structure variation-induced chromatin spatial reorganization in acute myeloid leukemia

Abstract

Acute myeloid leukemia (AML) is a set of heterogeneous myeloid malignancies hallmarked by mutations in epigenetic modifiers, transcription factors and kinases that can cause epigenetic reshaping. It is unclear whether those mutations drive chromatin 3D structure alteration and contribute to oncogenic dysregulation in AML. By performing Hi-C and whole genome sequencing in 21 primary AML and healthy donors' samples, we identified recurrent AML- or subtype-specific alteration of compartments, TADs, and chromatin loops. To study the impact on gene regulation, we performed RNA-Seq, ATAC-Seq and CUT&TAG for CTCF, H3K27ac, and H3K27me3 in the same samples. We observed dysregulation of many AML-related genes, represented by *MYCN, MEIS1, WT1, ERG, MYC GATA3, BCL11B* and *IKZF2,* intimately linked to the recurrent gain of loops and switch of compartment or TAD, alongside acquisition of AML-specific enhancer or repressor. Further, we profiled structure variations using WGS and Hi-C data to reconstruct the cancer 3D genome, by which we identified structure variation-induced neo-loops and enhancer-hijacking events. Furthermore, through conducting whole genome bisulfite sequencing in patient samples, we found altered methylation correlated with A/B compartment switch, and loss of CTCF insulation due to hypermethylation, leading to extensive gain of loops in AML. By treating the AML cells with DNA hypomethylation agent 5-azacytidine, the altered chromatin structure and gene expression can be restored, with switched compartment reverted and

gained loops dissociated, alongside compromised AML cell proliferation, overall providing insights into AML treatment through therapeutic restoration of chromatin structure.

Introduction

Acute myeloid leukemia (AML) is a heterogeneous and complex set of myeloid malignancies characterized by differentiation blockade and clonal proliferation of abnormal myoblasts in the bone marrow, at the expense of normal hematopoiesis. The National Cancer Institute of the National Institutes of Health estimates that 21,450 new cases of AML occurred in United States during 2019, taking 10,920 lives and leaving a five-year survival rate around 28.3%. Based on how mature the cancer cells are at diagnosis and the key genomic mutations, the different subtypes of AML have distinct prognosis and strategy of treatment[198, 199]. The World Health Organization (WHO) classifies AML into several groups according to some highly recurrent and outcome-associated genetic abnormalities, including but not limited to *NPM1* mutations, biallelic *CEBPA* mutations, inversion 16, *RUNX1* mutations including t(8;21) and other mutations, t(9;22) (*BCR-ABL1* fusion), t(15;17) (*PML-RARA* fusion), t(1;22), inversion 3 and *MLL* mutation[7]. AML patients also harbor many additional mutations that tend to be co-occurred with certain driver mutations, such as *FLT3*-ITD (internal tandem duplication), *TET2* with *NPM1* mutation, aneuploidy with *TP53* mutation, and *GATA2* with *CEBPA*[165]. Driver mutations disrupt functions of genes that encode histone or DNA modifiers such as *EZH2*, *MLL* family, *DNMT2A*, *TET2* and *IDH1/2*, transcription factors (TF) like *MYC, GATA1/2, RUNX1 and CEBPA*, chromatin structure proteins like *STAG2* and *RAD21,* or kinases like *KRAS, NRAS, FLT3*, *KIT* and *ABL1*, driving the epigenetic reshaping of chromatin[7, 164, 200]. Therefore, AML subtypes adopt unique chromatin landscapes of DNase hypersensitive sites and binding of TFs, forming subtype-specific regulatory

networks[201]. However, little is known about the physical structural basis that accommodates the AML regulation network. It has been appreciated that chromatin folding and spatial organization plays a critical role in cis-regulation in normal development and diseases, and aberration in epigenetic modification is also known to disrupt chromatin 3D structure[55, 57]. Therefore, we ask whether AML manifests recurrent and subtype-unique change of chromatin conformation, and whether and how they contribute to the oncogenic transcriptional misregulation.

A hallmark of AML is large structural variations (SVs), including inversions, deletions, duplications, and translocations. Large SVs have been linked to alterations in chromatin architecture and domains, including the formation of "neo-TADs"[32, 90]. SVs also bring originally distant genes and cis-regulatory elements in proximity by interactions to form the so-called "neo-loops"[100], opening the avenue for "enhancer hijacking" to take place when an ectopic enhancer activates the transcription of the gene[91]. While previous studies revealed specific incidents of such events in driving glioma, T-ALL and developmental diseases[97-100, 202, 203], it is unclear whether "enhancer hijacking" rarely or frequently occur in cancer or AML. Genome-wide assessment of its frequency is challenging, because it requires thorough interrogation of chromatin interactions formed across SVs and assignment of an ectopic enhancer to a gene that have not been paired before in reference genome. However, current analytic tools tackling Hi-C data are not adapted to rearranged genome. To resolve the whole picture of AML chromatin reconfiguration and not to miss those that rise from SVs, a tool that can map Hi-C matrix to the rearranged cancer genome is needed to identify neo-loops genome-wide.

Given extensive chromatin structure change in AML that underlies the transcriptional misregulation, an important question is whether there are means of AML treatment through therapeutically restoring the chromatin structure. Previous studies showed that cancer genomes are frequently hypermethylated at CpG island and CTCF binding sites, leading to CTCF displacement and

disruption of genomic insulation[97, 98, 204]. Also, altered methylation can reconfigure the chromatin structures by recruiting proteins[205]. Therefore it is crucial to understand the association between DNA methylation and chromatin 3D structure in AML, and to test whether DNA hypomethylation agents like 5-azacytidine (5-AZA) are able to restore chromatin structures, as a potential therapeutic pathway [206-208].

In this study we performed in-situ Hi-C and RNA-seq in 21 primary samples to identify change of chromatin structure and transcription dysregulation, including peripheral blood mononuclear cells (PBMC) from 3 healthy donors and 18 leukemia patients of various WHO subtypes,. We adopted PCR-free whole genome sequencing (WGS) on all leukemia samples to more accurately profile mutations and SVs (**Figure 4**-1A). We found subtype-specific alteration of 3D genome structure, including compartment, TAD and loops (**Figure 4-**1B). By conducting ATAC-seq and CUT&TAG of H3K27ac and H3K27me3, we identified AML-specific enhancers and repressors that co-occurred with gain of chromatin loops (**Figure 4**-1C). We further applied our newly developed computational method *Neo-loop Finder* to detect SV-induced neo-loops and enhancer hijacking events genome-wide. Next, we associated DNA methylation with change of chromatin structure, through performing whole genome bisulfite sequencing (WGBS) on 2 controls and 10 AML samples at 30X coverage (**Figure 4-1D**). We integrated our CUT&TAG of CTCF to show disruption of CTCF insulation by hypermethylation as a cause of gain of loops. We proceeded to verify that methylation inhibitor 5-AZA can restore the altered chromatin structure and gene dysregulation, by performing Hi-C, WGBS, CUT&TAG of CTCF and RNA-seq on two AML cell lines. We showed that DNA methylation can be targeted as a pathway to the restoration of chromatin structure and transcription cis-regulation in leukemia treatment.

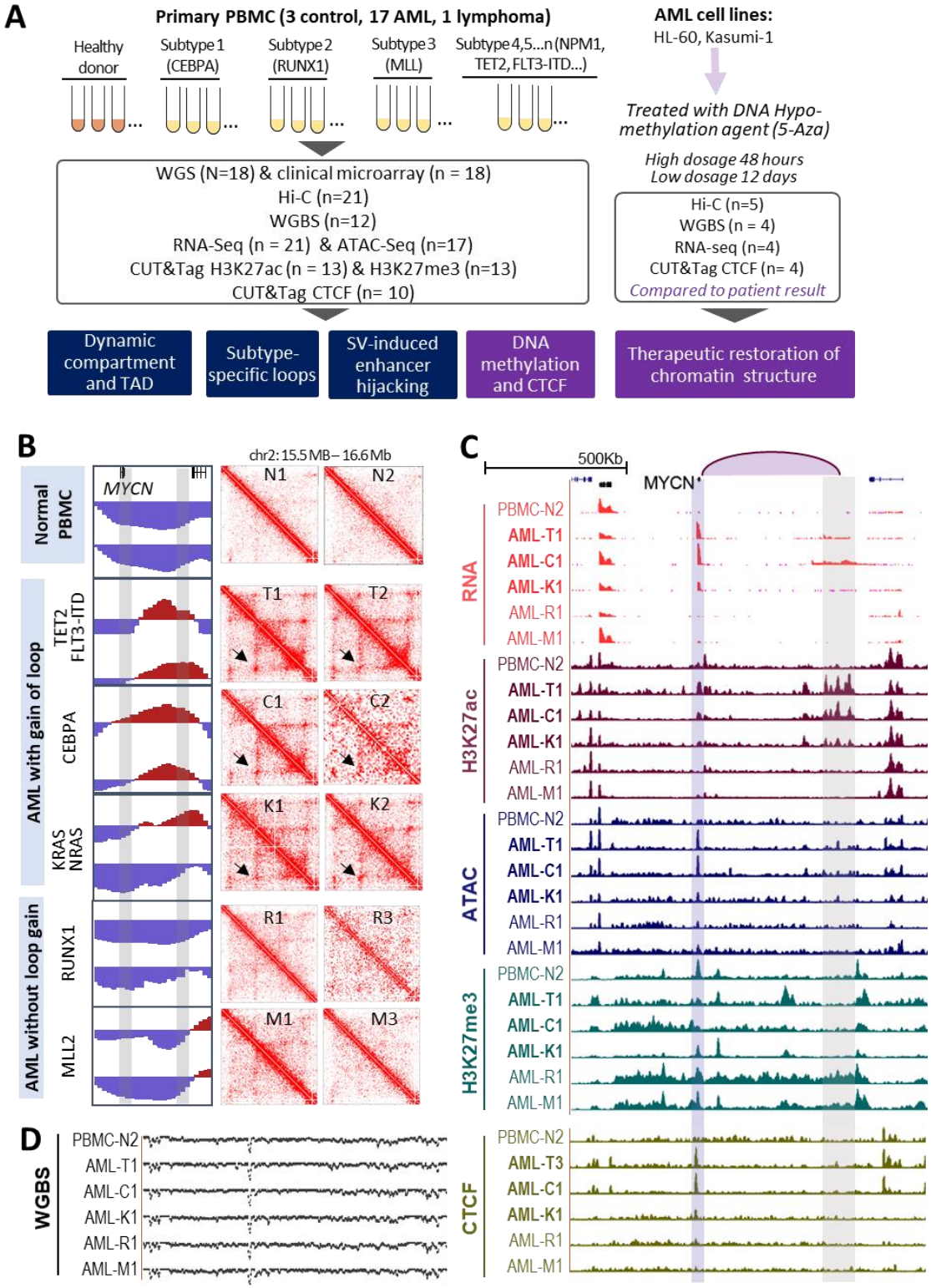


Figure 4- 1. Figure 1. Identification of chromatin reorganization and related cis-regulatory dysregulation in primary AML samples.

A. Overall design and workflow of this study. 18 Primary leukemia samples and 3 control PBMC from healthy donors are included. **B.** Representative identification of subtype-specific differential compartment (left) switch and gain of chromatin loops (right) at gene MYCN and its interacting regions. The arrow points to the stripe of interaction hotspots that demarcate a cluster of loops gained for the MYCN. N1, N2, T1, T2, etc. are the ID of each patient with the subtype-defining mutations stated in the left box. **C.** Activated gene transcription, active promoter and enhancer co-occur with gain of chromatin loop for MYCN. Shown are the representative RNA-seq, ATAC-seq, and CUT&TAG data of H3K27ac, H3K27me3, and CTCF from one sample of each subtype and control. Sample ID with bold font are the ones with gain of loops. The purple arc links the loop anchors, with gene promoter highlighted in light purple and distal regions labeled in gray. **D.** Representative DNA methylation profiled by WGBS.

Results

AML of same subtypes share similar alteration of chromatin compartmentalization

In-situ Hi-C was performed in three normal PBMC, 17 AML samples known for mutations of t(6;9), t(8;21); t(9;22), inv(16), NPM1, DNMT3A, TET2, FLT3-ITD, CEBPA, RUNX1 and MLL family point mutation, etc. (**Figure 4-2A).** Each sample generated in average around 600 million paired-end raw Hi-C reads. We also included one lymphoma sample to test if Hi-C distinguishes between different blood malignancies. PCR-free WGS was conducted with in average 40× coverage to facilitate mutation subtyping, in addition to the results acquired from clinical diagnosis.

First, we tested whether AML subtype is related to chromatin 3D structures. The unsupervised hierarchical clustering based on first principle component of the Hi-C matrix is able to distinguish between AML samples of different WHO subtypes with highly consensus mutation and separates them apart from the lymphoma (**Figure 4-2A**). Noticeably a group of samples are clustered together by commonly sharing mutations in KMT2B gene (MLL2), which has long been known as a hotspot of AML but has not been used as a dominant subtyping classifier. Compartment analysis shows that A-to-B

compartment switch is more prevalent than B-to-A switch (6.1% VS 3.2%) in AML, defined by a region consistently A or B in three controls but turning into B or A in AML samples **(Figure 4-2B**). Intriguingly we found certain regions show recurrent or subtype-specific compartment alteration (**Figure 4-2C**), in which reside many cancer related gene (**Figure 4-S1**). The A-to-B compartment switch at the promoter of all genes or cancer-related genes is highly correlated with decreased transcription, while B-to-A with increased transcription (**Figure 4-2D and Sup Figure 4-2A**). As exemplified in **Figure 4-2E** and Sup **Figure 4-2A**, *GATA3, BIRC3, BCL11B, ATM* and *RAD21* that are known for being frequently mutated in various blood malignancies are found recurrently turning their promoter from A to B compartment[164, 209], with the expression of the gene either repressed or kept silenced. Meanwhile, B-to-A switch affects genes related to promoting tumor growth, such as *WT1*, *FGF13*, *POU2AF1* and *IGF1R*, which are also associated with transcription activation (**Figure 4-2D, Figure 4-S1C**). WT1 specifically show B-to-A switch in samples with *TET2* or *FLT3*-ITD mutations. The compartment switch of WT1 is also associated with gain of ATAC-seq and H3K27ac peaks and loss of H3K27me3 at the gene promoter in the same samples, as shown by **Figure 4-S2B**. This finding is consistent to previous studies showing the potential association between *WT1* activation and *FLT3*-ITD mutation[210].

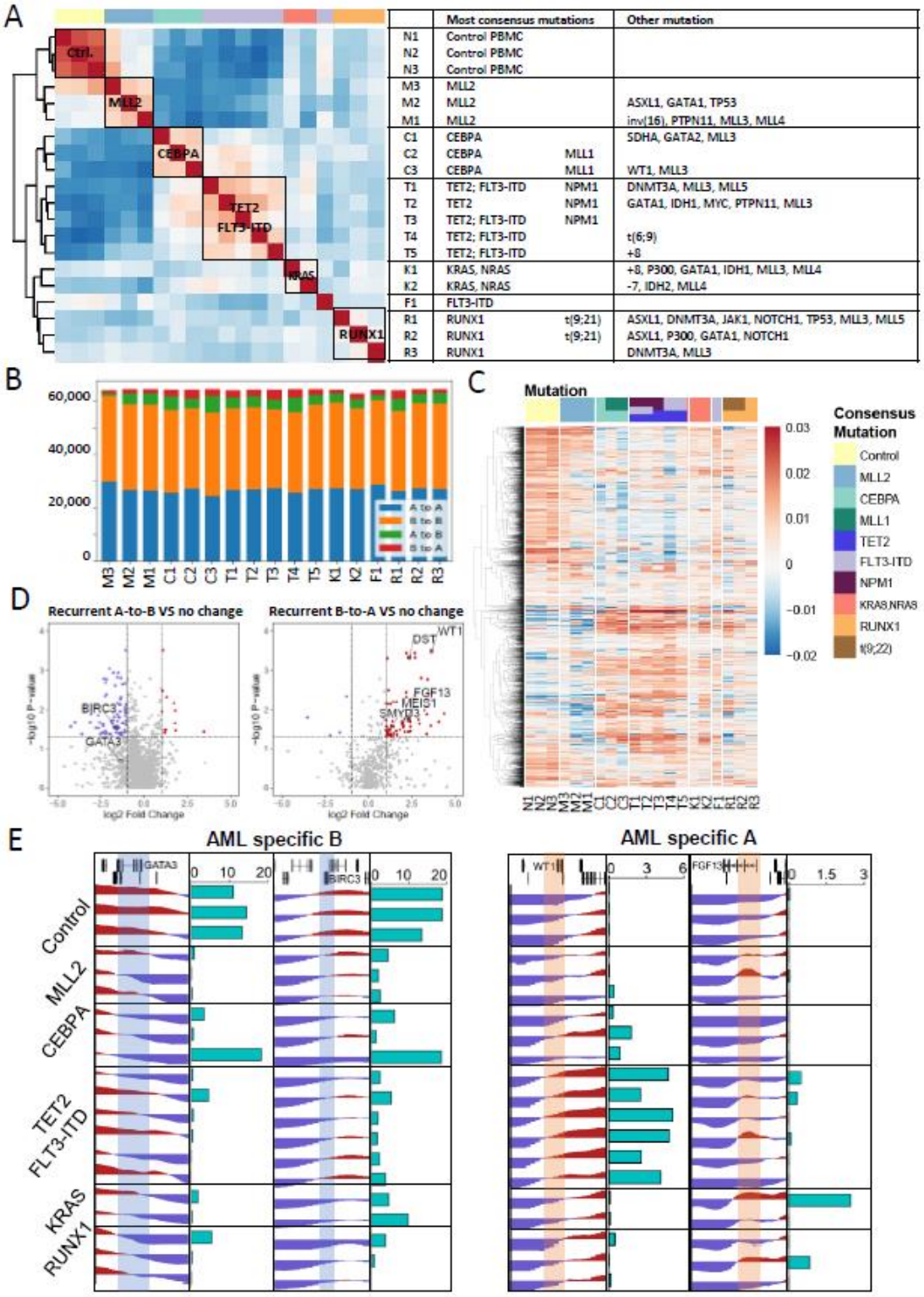


Figure 4- 2. AML of same subtypes share similar alteration of chromatin compartmentalization.

A. Unsupervised hierarchical clustering of leukemia samples and controls based on the top 10% most variable first principle component of Hi-C matrix (left). The right panel shows the patient ID with profiled mutations of the AML-relevant genes. The consensus mutations are not pre-selected but summarized from clustering result. **B.** Number and proportion of compartment switch in each AML samples. **C.** Hierarchical clustering of the top 10% most variable regions in A based on first principle component. **D.** Differential gene expression in samples with compartment switch compared to the expression of the same genes in the samples without compartment switch. P value is calculated from Wilcoxon rank sum test. Selected AML-related genes are demarcated. **E.** Correlation between compartment switch and gene transcription (marked by green bars). A compartments are marked red and B compartments marked blue. The most differential regions between the controls and altered samples that overlap with the gene promoters are highlighted in blue or orange.

Recurrent TAD disruption

To test the structural variation of AML chromatin domains, topological associating domain (TAD) was called at 40Kb resolution using *DomainCaller* in all the samples[61]. We depict the alteration of TAD by comparing the span of each TAD in AML samples to that in the three controls, and inferred three forms of TAD alteration: shrink, expand and shift while the three controls have to be consistent, illustrated in **Figure 4-S3A**. Surprisingly, we found extensive TAD alteration for samples with similar sequencing depth, with many cancer-related genes involved in altered TADs (**Figure 4-S3B**). However, unlike compartment switches, most TAD alteration does not seem to perturb the expression of genes inside. This is consistent with recent finding from analyzing the consequences of mutation at TAD boundaries in hundreds of tumor samples covering various tumor types [55, 211]. We found a few exceptions of genes *ERG*, *MYC* and *GATA3* with TAD alteration correlated to expression (**Figure 4-3C**), featured by co-occurring with gaining or losing cross-TAD interactions averaged from all AML samples (**Figure 4-3A, B**). For example, a TAD is left-expanded in some samples and it incorporates *ERG* into it. *ERG* gains increased interaction with a distal region 0.88Mb apart coming from the expanded TAD (**Figure 4-3C**). *MYC* resides in a highly self-constrained TAD in control PBMC, but in some AML samples the TAD is right-expanded and incorporates a cluster of previously

discovered super enhancers 1.7Mb downstream of *MYC* [212]. For these two genes, TAD expansion either engages the gene or engages a super enhancer, both associated with transcription upregulation. GATA3, in contrary, have reduced interaction frequency with nearby regions while the TAD is expanded/shifted in AML samples.

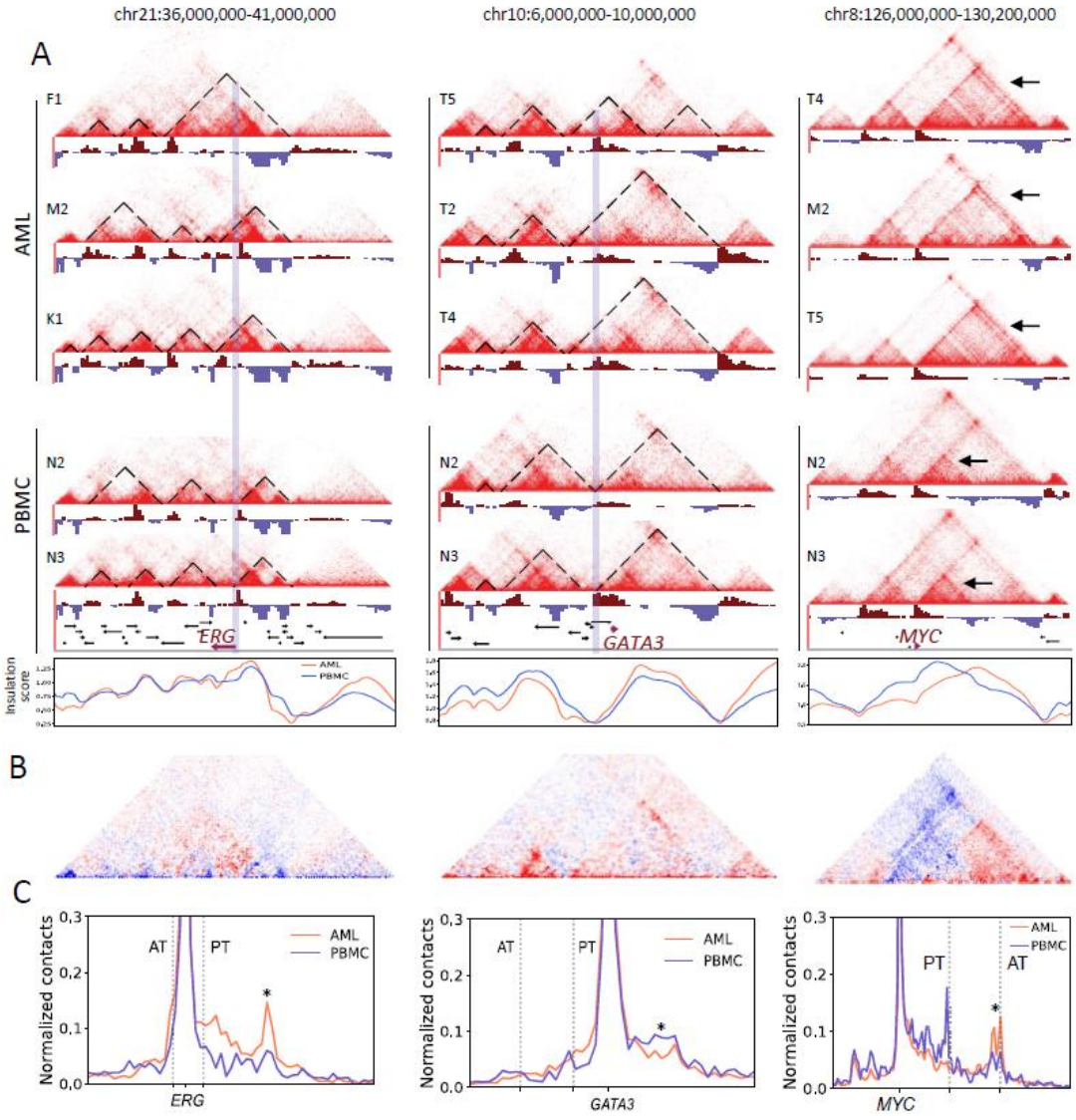


Figure 4- 3. Recurrent TAD disruption associated with cross-boundary interactions.

A. TAD shift in regions containing gene ERG (left), GATA3 (middle), and MYC (right). The TAD boundaries are marked in dashed black lines, or pointed out by the black arrow. The red and blue track below each heat map is directionality index. The shift of TAD boundary is highlighted by the vertical grey line. The track below the bottom Hi-C heatmap marks the location of all coding genes, with the affected gene highlighted in red. The box in the bottom shows the insulation score for the same region. **B.** Differential

chromatin contacts of the same region from the above to show change of interaction associated with TAD boundary shift, calculated by average Hi-C contacts in AML minus that in controls. **C.** Virtual 4C showing the averaged differential interaction of gene ERG, GATA3 and MYC between AML samples and PBMC controls, marked by asterisk. AT is short for the shifted TAD boundary in altered AML, and PT stands for the original TAD boundary in the PBMC samples. Boundary movement from PT to AT are for showing the incorporation of affected gene promoter or a distal region in the altered TAD in AML.

Gain of loops and stripes links genes to co-occurred AML-specific enhancer or repressor

We call loops at 10kb resolution using our machine-learning based loop caller *Peakachu*[213], which outputs probability score of each loop and we use it to identify differential loops between AML samples and controls. For each AML sample, we define gain of loops as loops specifically exist in this AML sample but absent in all three controls, and loss of loops as loops exist in all three controls but absent in this sample (**Figure 4**-4A). At comparable sequencing depth of each sample, we found hundreds of gain of loops and tens of loss of loops each samples. In average, gain of loops are five to ten times more prevalent than loss in AML samples (**Figure 4-4B**). To distinguish chromatin looping or an artifact of copy number alteration, which is frequent seen in cancer, we examined the copy number of all loop anchors. We ensured that overall less than 1.3% of gain of loops are confounded by gain of copy number, and less than 3% of loss of loops are confounded by loss of copy number, except one sample K2 that has extensive loss of copy genome-wide (**Figure 4-S4 A-C**). Comparing between AML samples we see a strong pattern of subtype-specific shared gain of loops (**Figure 4-4C**), with generally cancer-related genes involved, and globally correlated with transcription upregulation (**Figure 4-4D**). For recurrent loss of loops, however, we see comparable number of genes with upregulated or downregulated transcription (**Figure 4-S4D**). Gene set enrichment analysis of genes that recurrently gain loops demonstrate enrichment for cellular status including hematopoietic stem cell, chronic meylogeneous leukemia, and acute promyelocytic leukemia, as well as an enrichment in genes that are regulated by *c-Myc* (**Figure 4-S4E**).

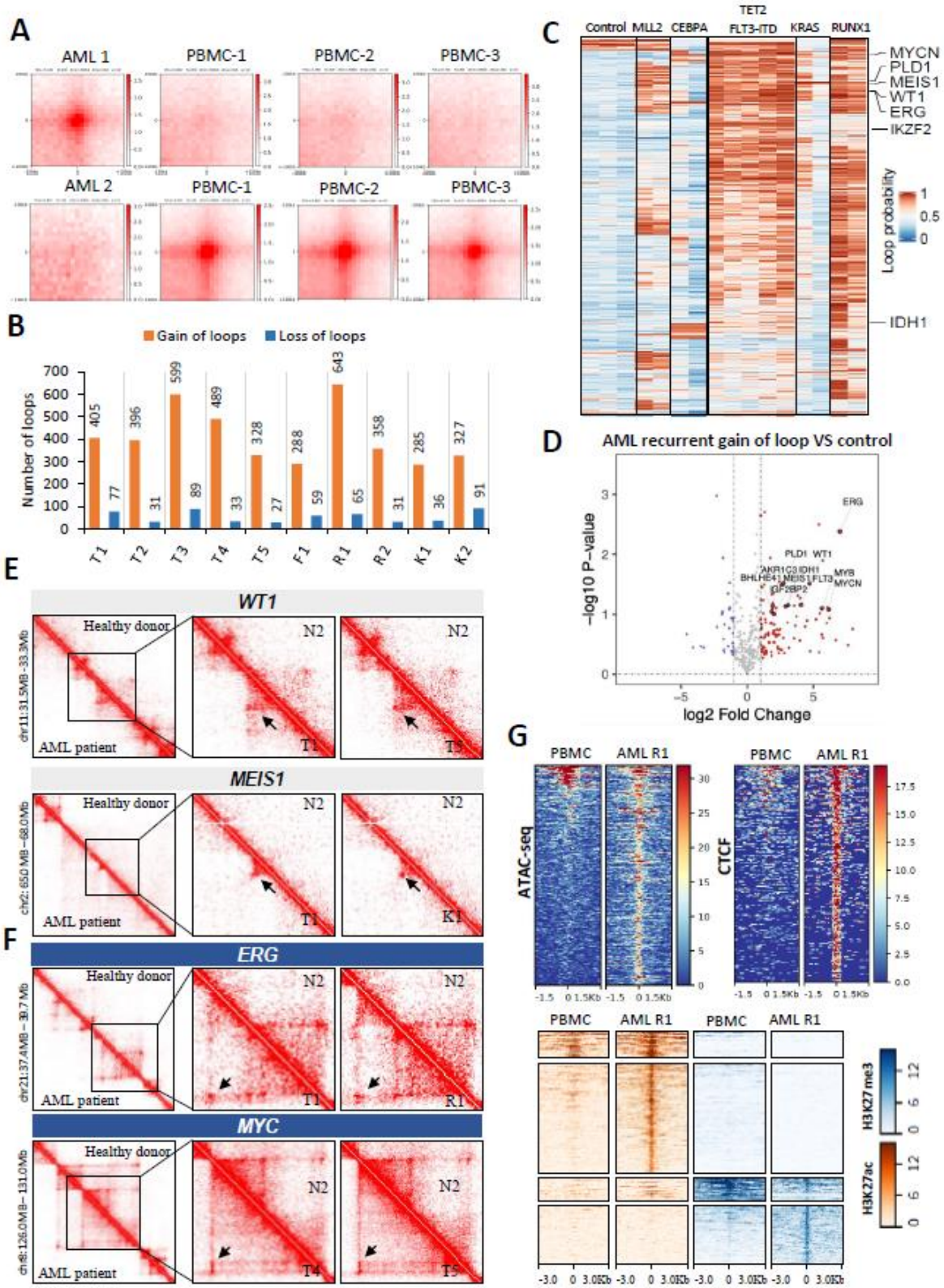


Figure 4- 4. Gain of loops and stripes link genes to co-occurred AML-specific enhancer or repressor

A. APA plot illustrating the definition of differential gain of loops (upper panel) or loss of loops (lower panel) in AML, exemplified by AML samples T3 and C3, respectively. Differential loops are calculated based on the Gaussian mixture model of the fold change of Peakachu probability with FDR lower than 5%. **B.** Number of gain of loops and loss of loops in each AML samples. **C.** Subtype-specific loops, presented by hierarchical clustering of all loops merged from 17 AML samples, based on Peakachu probability score. Each row is a loop and each column is a patient sample. Selected genes that are involved in gained loops are labeled in the right. **D.** Differential transcription of genes in samples with gain of loops compared to transcription of the same gene in controls. Selected AML-relevant genes are demarcated. P value is calculated by t test. **E and F.** Heatmap showing the gain of loops in the form of FIRE (E) and stripe (F) in AML samples (left lower panels) in contrast to that in PBMC controls (upper right panel). Zoomed-out heat map in the left for showing equivalent presentation of heat maps from AML and controls. **G.** Heatmap of ATAC-seq peaks and CTCF peaks separately (upper panel), and H3K27ac and H3K27me3 peaks together (lower panel) at gained chromatin loops in AML sample R1.

As exemplified in **Figure 4-1B** and **Figure 4-4D-F**, AML proto-oncogenes including *MYCN*, *WT1*, *ERG*, *MEIS1*, *MYC*, MYB, FLT3 and IGF recurrently gain loops, correlated with transcription activation (**Figure 4-4D**). Those genes show subtype-specific pattern (**Figure 4-4C**). Specifically, *MYCN* forms interaction with a region around 650Kb downstream, in the samples with *TET2/FLT3-ITD, CEBPA* or *KRAS/NRAS* mutations, but not in samples with *RUNX1* or *MLL2* mutations (**Figure 4-1B**). The acquisition of loop for gene *MYCN, MEIS1 and WT1* are also associated with B-to-A compartment switch on the gene promoter or the distal end of the loop (**Figure 4-1B, Figure 4-2E and Figure 4-S2B**). For *WT1* and *MEIS1*, multiple gained loops reside in some AML-specific domains with heavy intra contacts (**Figure 4-4E**), which represents the frequent interacting regions (FIRE) identified in previous studies[86]. Intriguingly, as we performed CUT & TAG of H3K27ac and H3K27me3 in AML samples, we found that the distal loop anchor or the entire FIRE of all the exemplified genes show simultaneous gain of enhancers, and loss of repressive marks (**Figure 4-S5 A-C**). Inversely, we also observed loss of loop for tumor suppressor *BCL11B* in five samples, all co-occurring with loss of enhancer and silencing of gene transcription, as shown in **Figure 4-S5D**. The above observations let us ask whether loops are always gained alongside formation of AML-specific novel enhancers. By mapping our ATAC-seq and CUT & TAG data, we saw indeed two types of loops acquisition (**Figure**

4-4G): The majority co-occur with acquirement of open chromatin status and novel enhancers, whereas a small proportion of loops build structural basis for linking pre-established enhancer to genes.

In addition to gain of enhancers, we identified many loops absent of enhancer marks but gain of repressive marks (**Figure 4-4G**), which we suspect are "repressive loops" that mediates gene downregulation[214]. As we analyzed the composition of gained loops, we found 36.8% of gained loops involves a gene promoter with a cis-regulatory element, 5.9% are between two promoters (P-P), 24.6% between promoter and enhancers (P-E), and 9.6% between promoters and repressors (P-R), with very few overlap between enhancer and repressors (**Figure 4-S6 A-C**). We then stratified the transcription analysis separately for gain of P-E loops and P-R loops. As shown in **Figure 4-S6D**, genes with gain of P-E loops show greatly increased expression in AML, and genes that gain P-R loops show significant decreased expression. *IKZF2*, a known tumor suppressor in lymphoblastic leukemia, simultaneously gains loops and novel repressors at the distal end of those loops, in our samples specifically with *TET2* and *FLT3*-ITD mutation (**Figure 4-S6E and Figure 4-4C**). The acquirement of these loops correlates with dramatically reduced to near-silenced transcription of *IKZF2* (**Figure 4-S6F**), which is also associated with significantly poorer prognosis in TCGA's AML cohort (**Figure 4-S6G**). Overall, we showed that both enhancer loops and repressor loops can be acquired in AML.

Noticeably, genes like *MYCN*, *ERG* and *MYC* gain a cluster of interactions between its promoter and a continuous trunk of distal regions, forming the stripe structure previously described as a result of loop extrusion (**Figure 4-1B and Figure 4-4F**)[215]. To more comprehensively identify gain of stripes, we developed a stripe caller by implementing a previously proposed conceptualized method[215]. As shown in **Figure 4-S7**, we were able to find in average 587 gained or stronger stripes (>200kb) in AML samples, which in APA plot manifests an anchor interacting with a sliding zone (**Figure 4-S**7A). Previous work demonstrated that stripes are enriched of super enhancers. While we have consistent

findings for promoter-super enhancer stripe involving genes like *MYC*, we also found a considerable number of promoter-repressor stripe recurrent in our samples, as exemplified by *HOXD* family and *KLF4* (**Figure 4-S7 B-C**). In fact, when comparing to loops (**Figure 4-S6A**), stripes contains more abundant cis-regulation, especially P-R interactions that increased to 23.7% (VS 9.6% in loops), whereas P-E stripes also increased to 32.4% (VS 24.6% in loops) (**Figure 4-S7D**). Genes engaged in the gained P-E stripes have globally higher expression than those engaged in the gained P-R loops (**Figure 4-S7E**). Our result indicate that gain of stripes in AML represents hubs for chromatin structure change that are more functionally relevant to cis-regulation.

SV-mediated neo-loop with enhancer or repressor hijacking

In addition to regular gain of loops, our previous work suggested that SVs can induce formation of neo-TAD and neo-loop[32]. SV profiling by WGS and *Hi-C Breakfinder* show that our AML samples harbor abundant large SVs. In average each samples has 6 del, 2 inversion and 21 translocations that are larger than 1Mb, which in Hi-C map results in aberrant contacts and interactions between originally distant regions (**Figure 4-S8A**). Interestingly, we found more SVs than expectation within 1Mb to AML-related genes but are not overlapping with the gene body (**Figure 4-S8B**). We hypothesized that those SVs might be able to affect transcription of genes across the SVs through the "enhancer-hijacking" mechanism, so we applied our newly developed tool *Neo-loop finder* to comprehensive identify SV-induced neo-loops. It first reconstructs the Hi-C map along the SV breakpoints (**Figure 4-5A**), guided by locus, orientation, copy number, and signature of complex SVs, and then detects loops adaptive to reconstructed Hi-C map. We performed this analysis on all of our samples and AML cell lines HL60, Kasumi-1, THP-1 and CML cell line K562[216]. As a result, we found 328 cancer related genes wired

into neo-loops mediated by SVs, among which 56 recurrently form neo-loops, with the other end of the loop vary in different regions caused by different SVs (**Figure 4-5B**). These include AML-related proto-oncogenes: *CDK5, CBL, MYC, ETS1, LMO3, and FLT3* (**Figure 4-5 C-E**), all of which are highly expressed in the affected samples and interacting with distal enhnacers[217] (**Figure 4-5 F-G**). Also, all of those exemplified genes have no copy number change that could lead to high expression. Therefore, the activated transcription might be a reflection of the "enhancer-hijacking" events. To check if this is widely happening in AML than a few samples, we annotated all the loop anchors for gene, enhancers and suppressors, using our CUT&TAG data. As shown by **Figure 4-5H**, we found that 33.7% of neo-loops involve one promoter, 20.5% links a promoter with an enhancer and 8.3% links a gene with repressor. Genes that loop with enhancer show a significant increase of gene expression. (**Figure 4-5I**).

Aberrant DNA methylation associated with alteration of chromatin structure in AML

DNA methylation is known for extensive change in cancer and leukemia, which can alter the chromatin status and interfere with binding of CTCF[218, 219]. To understand its role in driving AML-specific chromatin structure change, we performed WGBS in 10 AML samples and 2 controls. As a result, we saw overall higher global methylation at AML samples, except samples with MLL2 mutation (**Figure 4-S9A**), consistent with previous findings[220]. One AML sample C1 exhibits extremely high methylation, because it has an *SDH* mutation, which was known to cause demethylation defects and typically high methylation in cancers[98]. Specifically, while globally both CpG island and CTCF binding sites show depletion of DNA methylation in the center (**Figure 4-S9B**), we see by comparison higher methylation in CpG island and CTCF binding sites in AML samples, at the binding sites profiled from normal controls, with exception of samples carrying MLL2 mutation (**Figure 4**-6A). Our CUT &

TAG of CTCF shows that the hypermethylation of CTCF motifs in AML substantially displaces their CTCF binding (**Figure 4-6B**). Comparing between AML samples, we found a strong subtype-specific pattern for the distribution of the DNA hypermethylation (**Figure 4-6C**).

We then tested the association between DNA methylation and the chromatin structure, by first looking at the A/B compartment. We found lower methylation at transcription starting site (TSS) but higher methylation at gene bodies for genes in A compartment, comparing to genes in the B compartment (**Figure 4-S9C**). This is consistent with previous knowledge about the distinct roles of methylations at different parts of genes: It prevents transcription firing in the TSS but it also stabilizes transcription elongation in the gene body by preventing spurious transcription initiation[221]. Then we tested whether A-to-B or B-to-A compartment switch, as we observed in AML samples, is associated with differential methylation between AML samples and controls. As shown in **Figure 4-6D**, we see a strong correlation showing that the compartment switch accompanies concordant change of methylation. Next, we ask if topological change is related to DNA methylation. We examine the methylation levels between two types TAD boundaries: the conserved ones between the control and the AML samples, and the variant ones. To our surprise, while there is clear a depletion of methylation at the conserved TAD boundary, the variant TAD boundaries do not seem to exhibit much depletion pattern, both in AML and control samples (**Figure 4-S10**). From what we observed, it seems that DNA hypo-methylation protects the steadiness of TAD boundaries, and those without hypo-methylation are more easily altered. Further, at loop levels, we observed that subtype-specific loop anchors are also enriched for differential methylation (**Figure 4-S9D**).

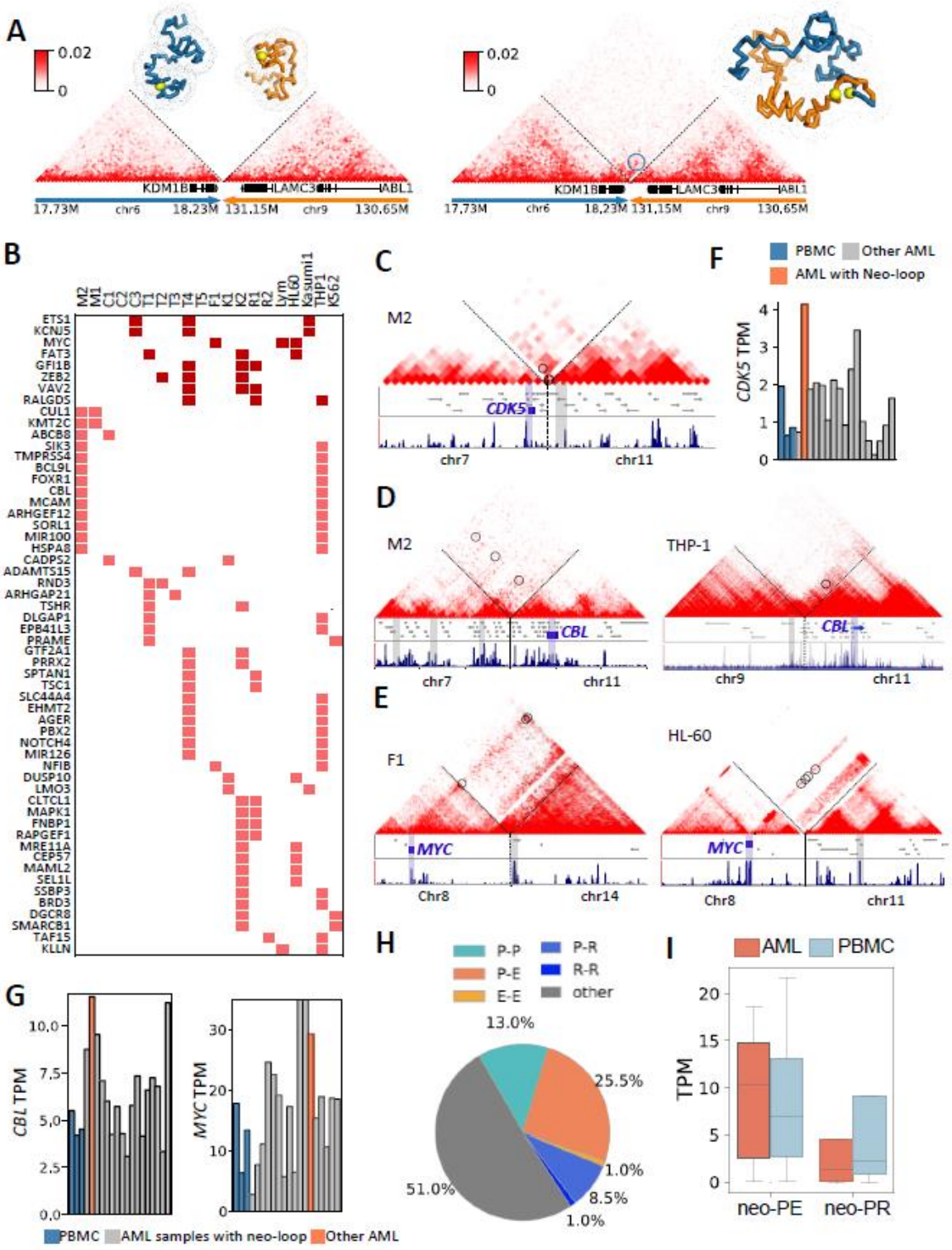


Figure 4- 5. SV-mediated neo-loop with enhancer or repressor hijacking.

A. Model of neo-loop detection. Hi-C map of chr6 and chr9 are concatenated along an SV detected in the sample in the right panel. The identified neo-loop is circled. The same regions for normal controls are shown in the left panel. **B.** A list of cancer-related genes that are recurrently involved in SV-induced neo-loops across AML samples and AML/CML cell

lines. Genes with 3 recurrences are labeled dark red. **C-E.** Examples of gain of neo-loops (marked by the circle) induced by SVs in different samples. The anchor of each neo-loop is shaded, involving gene CDK5 in sample M2, CBL in sample M2 and cell line THP-1, and MYC in sample F1 and cell line HL-60. The bottom track below each heatmap shows the H3K27ac signals from the same samples. The regions used for concatenating the Hi-C map from C to E are as below: chr7:150990919-151740919(-) and chr11:117734283-118484283(+) for CDK5, chr7:148490916-150990916(+) and chr11:118484298-120984298(-) for CBL in M2, chr9:13460000-15460000(+) and chr11:118480000-120480000(-) for CBL in THP-1, chr8:127000000-129700000(+) and chr14:96630000-99330000(+) for MYC in 1360, and chr8:127000000-129000000(-) and chr11:94600000-96600000(+) for MYC in HL-60. (+) and (-) indicate the orientation of SVs. The map is reversely placed with 3'-to-5' direction under (-) mark in the left or (+) in the right. **F-G.** The mRNA expression of the corresponding gene in all samples. **H.** Percentage of neo-loops formed between the indicated genomic elements. P: promoter, E: enhancer marked by H3K27ac excluding promoters, R: repressor marked by H3K27me3 excluding promoters. **I.** mRNA expression of genes in SV-induced neo-loops between promoter and enhancers, or promoter and repressors.

Next, as we see hyper-methylation displaces CTCF binding in AML, we ask whether it leads to loss of insulation and gain of interactions across the lost CTCF binding site. To test that, we picked the 100 loss-of-CTCF sites due to most significant hypermethylation, and for control we randomly collected 100 normal CTCF sites, both from AML samples. Aggregating the AML Hi-C heatmaps that put the CTCF sites in the middle, we saw a clear loss of insulation and increased interactions across the lost CTCF, as shown by the upper panel of **Figure 4-6F**. The aggregated Hi-C map of PBMC using the same CTCF sites, in contrast, did not show this pattern, confirming that the loss of insulation and gain of interaction is caused by hypermethylation-induced loss of CTCF (**Figure 4-6F**). We then ask whether any observed gain of loops are caused by this mechanism. For such cases, the frequency of chromatin loops should be inversely correlated with the intensity of CTCF that bind in the between of the two loop anchors, across our samples. As shown by **Figure 4-6G**, we found many aforementioned gain of loops that contain genes show this pattern, including AML master regulators and oncogenes *MEIS1*, *MYCN*, *MLLT3, RARB*, and *WDR66*[222]. For example, the highlighted CTCF in **Figure 4**-6H, is lost in a few AML samples due to hypermethylation. This is correlated with gain of many loops across this site, including a loop involving *WDR66* promoter [**Figure 4-6H**]. Overall, DNA hypermethylation-induced loss of CTCF binding can be a cause for extensive gain of chromatin loops in AML.

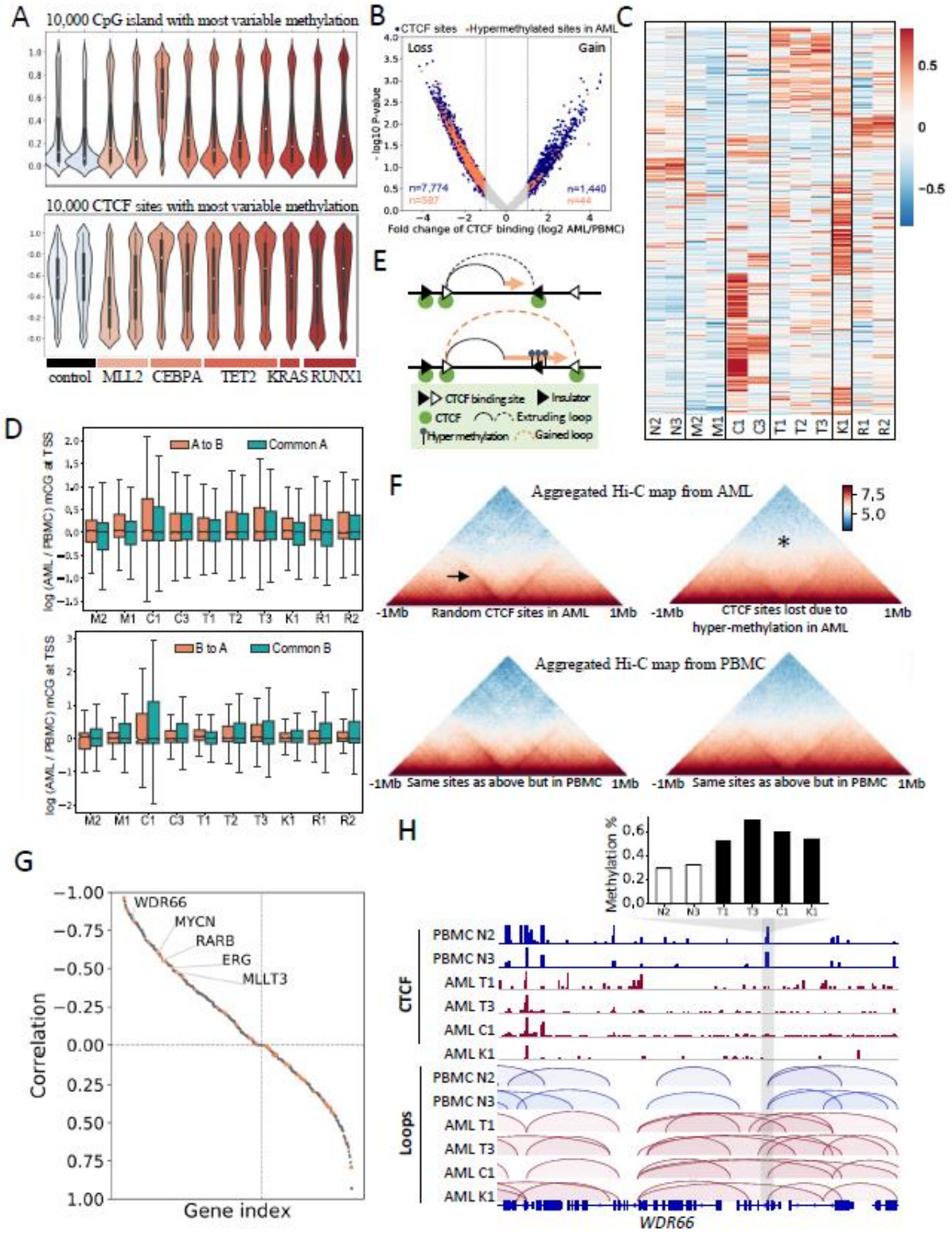


Figure 4- 6. Aberrant DNA methylation associated with alteration of chromatin structure in AML.

A. Distribution of CG methylation levels for the top 10,000 most-variably methylated CpG islands (upper) and the top 10,000 CTCF sites with the most variable methylation (lower) across AML samples and controls. **B.** Hypermethylation displaces CTCF binding. Differential CTCF binding sites between AML samples and controls (blue dots). Fold change and P-value of CTCF binding are derived from DiffBind. CTCF sites with hypermethylation in AML are marked in orange. Hypermethylation is defined as at least 1.5 fold increase of methylation in AML samples and the basal methylation in PBMC greater than 0.1. **C.** Hierarchical clustering of differential methylated regions (DMRs) across subtypes of AML samples and controls based on normalized methylation levels. **D.** Fold change of CG methylation levels at TSS regions within altered compartments or conserved compartments between AML and controls. A to B: regions that are A compartment in controls but B in AML samples. Common A: regions that are A compartment in both controls and AML samples. B to A: regions that are B compartment in controls but A in AML samples. Common B: regions that are B compartment in both controls and AML samples. **E.** A model illustrating gain of chromatin loops induced by loss of CTCF binding as a result of hypermethylation at CTCF binding sites. **F.** Aggregated Hi-C maps centered at 100 random CTCF sites (left) or 100 loss-of-CTCF-binding sites (right) due to hypermethylation in AML. The upper panel aggregates the Hi-C map of AML samples, whereas the lower panels aggregates the Hi-C of the control samples centered at the same sites. The arrow points to the insulation boundary or stripes that are lost in the right panel. The asterisk demarcates the region with increased contacts, which are interactions across the lost CTCF sites. **G.** Correlation between CTCF binding intensity and the probability score of the gained loops formed across the CTCF site. Only loops involving genes are plotted, and the orange dots mark all the cancer-related gene. Selected AML-related genes are named. **H.** Gain of *WDR66* loop correlated to loss of CTCF binding due to hypermethylation.

Inhibition of DNA methylation restores chromatin structure and gene expression

We then ask whether it is possible to restore the normal chromatin structure and gene regulation through modulating DNA methylation using hypomethylation agent like 5-AZA. As expected, 5-AZA significantly suppresses the growth of two AML cell lines Kasumi-1 and HL60 in a dosage-dependent manner, eventually completely blocks the cell proliferation and resulted in cell death (**Figure 4-S11A**). With 48-hour treatment of 1uM 5-AZA in HL-60 cells and 2uM in Kasumi-1 cells, global DNA methylation quickly decreased by 14% (**Figure 4-S11B**), similar to the observation from previous clinical trials with 2 to 4 days' treatment[223]. CpG island and CTCF binding sites profiled from control PBMC also showed demethylation (**Figure 4-S11C**). We also applied longer treatment of 5-AZA with lower dosage to Kasumi-1 cells (0.5uM for 12 days), to mimic a drug delivery that is more physiologically tolerable. We then tested the impact of this treatment on chromatin conformation by

performing in-situ Hi-C, using cells treated with DMSO as control. The treatment rapidly switched over 1000 bins of B compartment to A at 48 hours, with fewer A compartment regions turning to B (**Figure 4-7 A-B**). With 12-day treatment, compartment swap is further increased, especially in A-to-B manner, which we hypothesized are related to removal of active chromatin loops. Transcription of genes that reside in the B-to-A switch regions are significantly upregulated (**Figure 4-7C**). Those altered regions are highly representative of the patient samples as shown by **Figure 4-S11D**: 6% of genomic regions in Kasumi-1 are B compartment but A in PBMC, 87% of which are also B compartment in our AML samples. A short-term 48-hour exposure to 5-AZA restored 11.8% to A compartment, which was increased to 49.8% with 12-day treatment. Specifically with 12 days' treatment, compartment was reverted for genes *GATA3* and *BCL11B* that recurrently turn to B compartment in patient samples, and *WT1* that recurrently switched to A compartment (**Figure 4-7D**).

Since we identified global gain of loops correlated to hypermethylations in AML, we investigated whether it is possible to dissociate gained loops by using 5-AZA. As shown in **Figure 4-7E**, we found 107 differentially gained loops in Kasumi-1 comparing to PBMC. Intriguingly, a significant part was dissociated upon 12-day treatment of 5-AZA, and further, the 12-day treatment even restores a small portions of lost loops in Kasumi-1. We hence ask whether this is due to reduced methylation through rebuilding the CTCF insulation. We therefore picked the most differentially hypermethylated CTCF motifs in Kasumi-1 and HL60 cells, and checked their aggregated Hi-C maps with DMSO or 5-AZA treatment. As shown by **Figure 4-7F**, we saw a stronger pattern of insulation and decreased interactions across those CTCF sites upon 5-AZA treatment. Intriguingly, 12-day treatment completely erased the gained stripe anchored at *MYCN*, which was recurrently identified in both patients and in Kasumi-1 cells as correlated to CTCF loss (**Figure 4-7G**). We also observed hollow in the FIRE corresponding to the position of *WT1* interactions after 12 days. RNA-seq at 48 hour already show

significant decrease of *MYCN* and WT1 expression (**Figure 4-7H**), suggesting that controlling the DNA methylation level is able to diminish the activation of some oncogenes through dissociating chromatin loops.

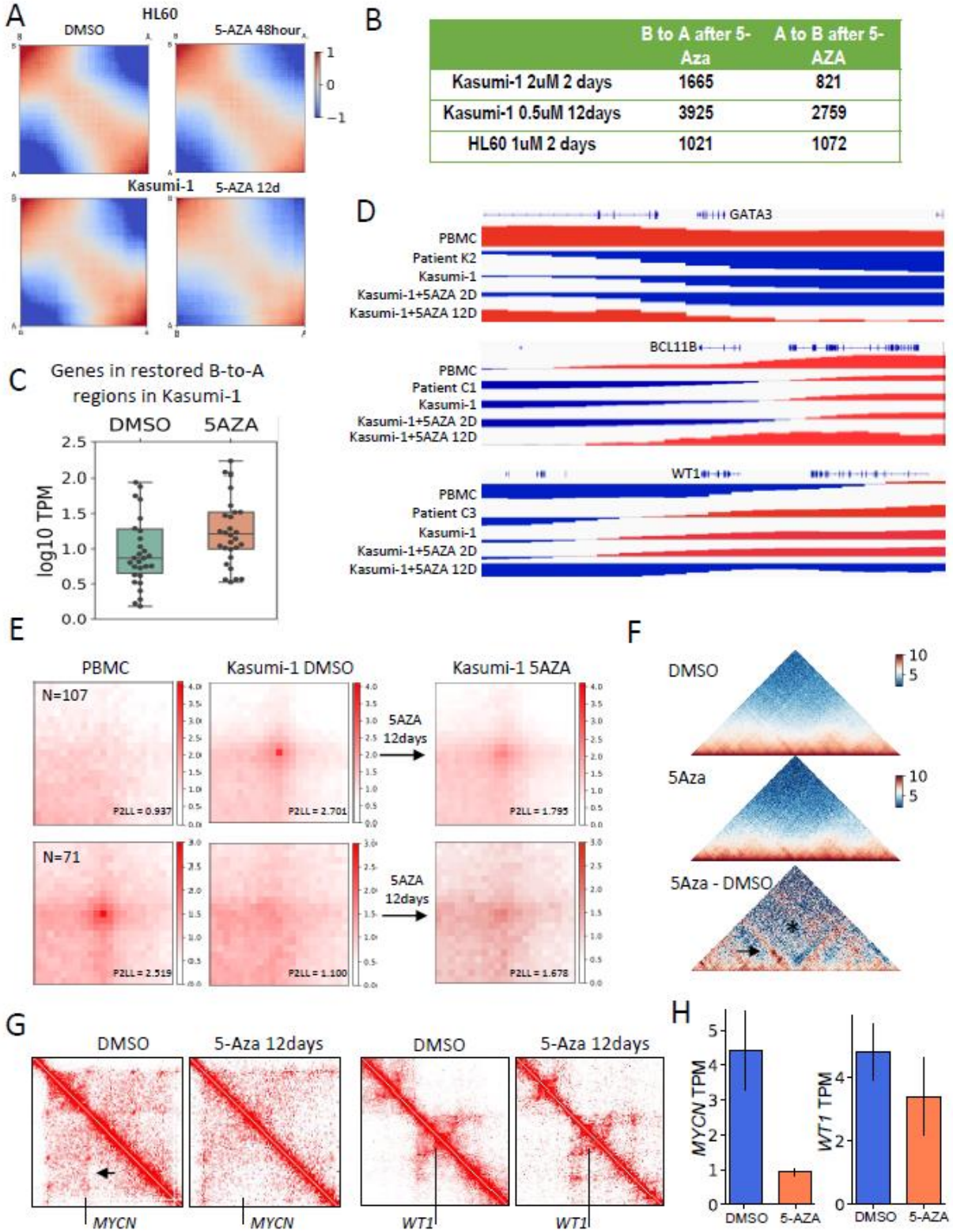
Discussion

AML is known for extensive heterogeneity in disease development and presentation. Patient can have onset of disease not rarely at all ages, facing highly distinct treatment and prognosis, with the myeloidblast cells originated from one of many differentiation stages and lineages[224]. We now know that the ultimate diverse disease phenotype is related to a variety of driver mutations, but how the genomic mutations distinctly contribute to this process at transcription level is understudied. Our work focusing on the chromatin spatial organization in AML demonstrated that AML samples not only extensively share recurrent alteration of chromatin structure, different subtypes also adopt distinct local conformation. We observed recurrent and subtype-specific compartment switch, shift of TAD boundaries, and massive gain of chromatin loops. All of those events are globally associated to dysregulation of gene transcription, including many leukemia and cancer-related genes. Moreover, the conformational change at different scales are highly consistent. For examples, gain of active chromatin loops accompanies B-to-A compartment switch for genes *MYCN* and *WT1*, or co-occur with TAD boundary shift for gene *ERG* and *MYC*. Inversely, loss of enhancer loop for gene *BCL11B* and *GATA3* co-occur with A-to-B compartment switch or shift of TAD boundary. We further identified hundreds of SV-induced neo-loops, and some genes are recurrently forming neo-loops although the SVs are different. This result indicates that neo-loops are more than rare or anecdotal events in AML, and

potentially are prevalent in all other cancers. We also found evidence of "enhancer-hijacking" as some genes interact with enhancers across the SVs, and globally correlated with transcription upregulation.

Our CUT&TAG for histone mark H3K27ac and H3K27me3 and ATAC-seq reveals a large number of AML-specific enhancers and repressor that are formed alongside gain of loops. Specifically, the finding of repressive loops is consistent with a previous study that showed a single risk element repressing transcription of *HOXA13* and *HOTTIP* in prostate cancer through a chromatin loop[225]. Our results further indicate that acquirement of repressive loops are widespread in the AML genome, mediating genome-wide transcription downregulation. Furthermore, we found that for a set of essential AML oncogenes, clusters of loops are gained in the form of FIRE or stripe. Comparing to regular loops, stripes are more enriched in gene promoters and cis-regulatory elements, particularly super enhancers and repressors. Overall, on AML context, we showed that gain of stripe or FIRE might represent misregulatory hubs for interacting *de novo* enhancers or repressor, which is usually adopted by the essential cancer-related genes for drastic transcription activation of suppression to take place.

Next, we identified that aberrant DNA methylation associated with AML change of chromatin structure. We have seen global and subtype-specific hypermethylation in the AML genome, correlated with prevalent A-to-B compartment switch, and loss and CTCF insulation that potentially drives excessive gain of chromatin loops. 5-AZA application was able to partially restore the chromatin structure including reverting switched compartment and dissociating gained loops, through rebuilding the genomic insulation. Understanding this mechanism of 5-AZA in leukemia is crucial, as the patient responsiveness can be better predicted based on the gene expression or 3D genome profile from the patient. For example, according to our data, patient with MYCN, WT1 activation or BCL11, GATA3 silencing are likely benefit from 5-AZA treatment.



	B to A after 5-Aza	A to B after 5-AZA
Kasumi-1 2uM 2 days	1665	821
Kasumi-1 0.5uM 12days	3925	2759
HL60 1uM 2 days	1021	1072

Figure 4- 7. Inhibition of DNA methylation restores chromatin structure and gene expression.

A. The heatmap of Hi-C interactions between A and B compartment under different drug treatmenet. Regions are ranked by the first principle component of Hi-C matrix, from top to bottom along the Y axis and from left to right along the X axis. The left upper corner indicates A-A interaction, the right lower corner indicates B-B interaction, whereas the right upper corner and left lower corner represent A-B interactions. **B.** Number of 40Kb genomic bins that show the corresponding compartment switch. **C.** mRNA expression for genes that are at B compartment in Kasumi-1 cells (treated with DMSO) and turned into A compartment upon 5-AZA treatment (2uM) for 48 hours. **D.** Examples of patient-representative restoration of compartment in Kasumi-1 cells upon 5-AZA treatment (0.5uM) for 12 days. Red and blue bars mark A and B compartment separately. **E.** Aggregated peak analysis plot for regions of Kasumi-1 specific loops, drawn from Hi-C in PBMC, Kasumi-1 cells treated with DMSO and Kasumi-1 cells treated with 5AZA (0.5uM) for 12 days. **F.** Restoration of insulation. Aggregated Hi-C maps centered at the most differentially hypermethylated CTCF sites in HL60 cells treated with DMSO (upper panel) and treated with 5-AZA (1uM) for 48 hours (middle panel). The bottom panel shows the difference between the two maps, calculated by the middle panel minus the upper panel. The arrow points to the restored insulation stripe and the asterisk marks the majorly reduced interactions across the centered CTCF sites. **G.** Complete erasion of *MYCN* strips (pointed by the arrow) and the weakening of the FIRE containing *WT1* (pointed by the vertical line) with 12-day treatment of 5-AZA (0.5uM) in Kasumi-1 cells. **H.** The mRNA expression of gene MYCN and WT1 with DMSO treatment or 48-hour 5AZA treatment (2uM) in Kasumi-1 cells.

In summary, we showed that subtype-specific chromatin conformational change and SV-induced neo-loop formation provides structure basis underlying the heterogeneity of transcriptional misregulation in AML. The 3D structure change is related to global and subtype-specific aberrant DNA methylation, which can be pharmaceutically targeted to restore the chromatin structure. A few thoughts rises as related to our findings. For example, as we observed enhancers gained along the chromatin loops, it will be crucial to uncover targetable master TFs. Also, 5-AZA was known for remarkable responsiveness at clinical trials, but its application is hindered by less satisfactory improvement of survival due to high relapse when using the drug alone. This might represent the epigenetic plasticity and redundancy in chromatin spatial organization, suggesting the importance of combined inhibition of other master regulator TFs. Last, the SV-induced neo-loops might be an important mechanism for disease onset, and how those events might inspire the disease intervention is question to be answered. Further exploration to the related questions will provides insight specifically into epigenetic therapies for AML.

Supplementary Figures

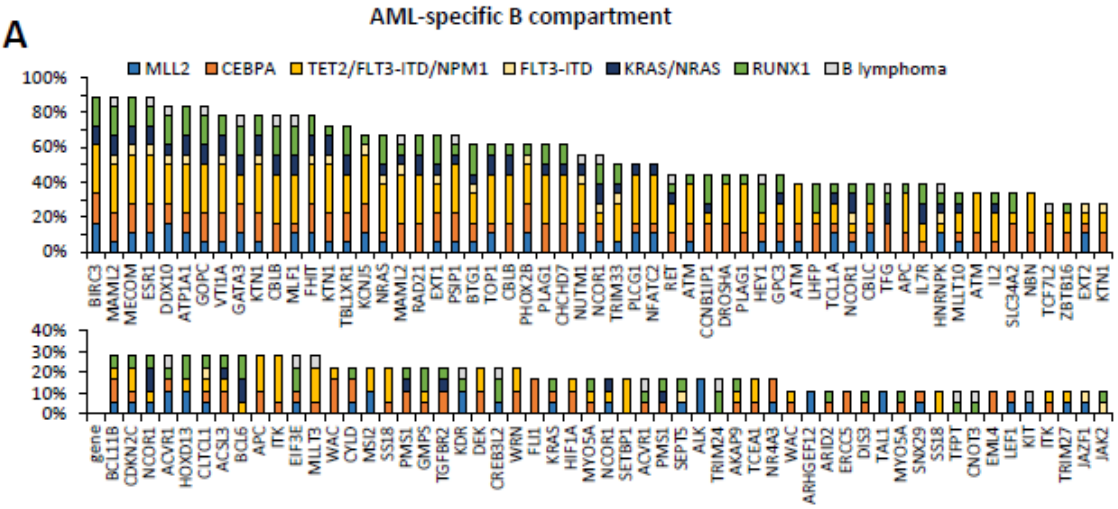


Figure 4-S 1. Recurrent compartment switch of COSMIC cancer-related genes.

A and B. Cancer related genes that recurrently switch promoters to B compartment (A) or to A compartment (B) in AML samples. Y axis indicates the percentage of recurrence across all AML samples. **C.** Example of genes with recurrent or subtype-specific compartment switch. The differential regions that overlap with the gene promoters are highlighted.

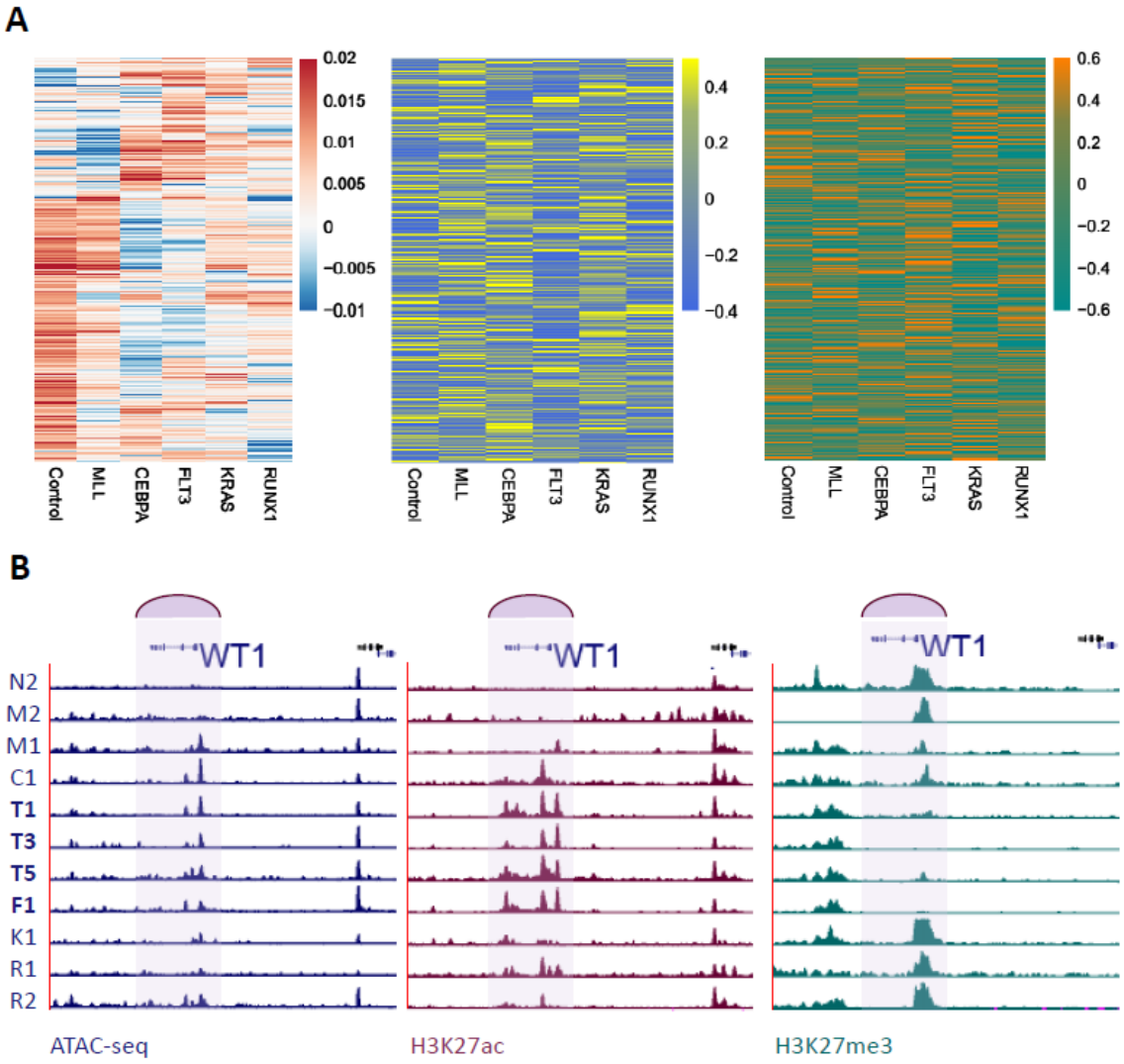


Figure 4-S 2. Compartment switch is correlated with gene expression and open chromatin.

A. Hierarchical clustering of the first principle component from Hi-C matrix for regions in Figure 2C, averaged within the same subtype (left panel), in parallel to the average mRNA expression (TPM) of the genes from the same region (middle panel) and average ATAC-seq normalized read count of distal regions (RPM) from the same region (right). **B.** B-to-A compartment switch (samples with bold font) at gene WT1 is correlated with open chromatin, active promoter and loss of repressive marks. The highlighted regions represents the FIRE with dense loop formation.

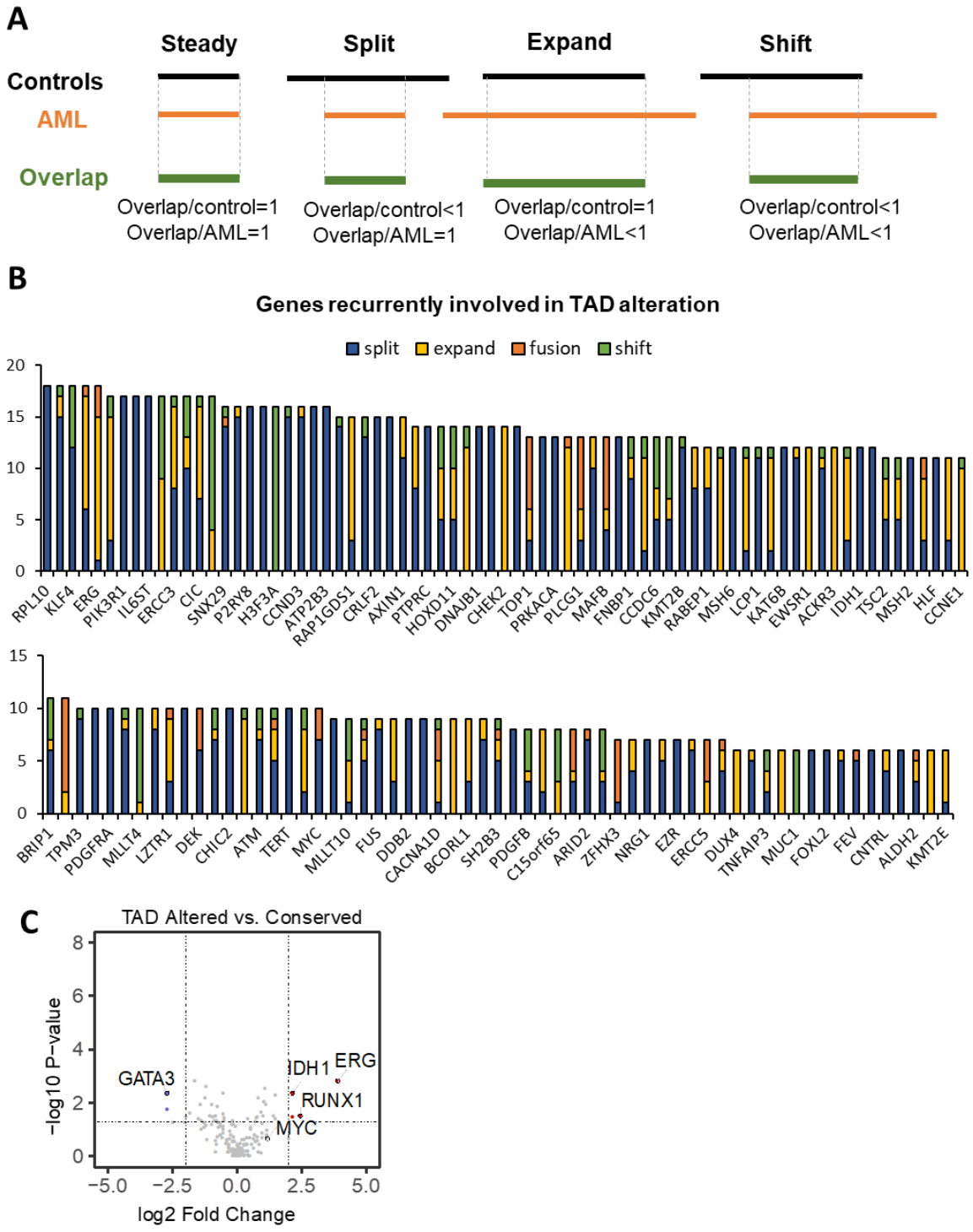


Figure 4-S 3. TAD boundary alteration and association with transcription.

A. Illustration for how to define alteration of TAD boundary. **B.** COSMIC cancer-related genes that are located in TADs with recurrent change of boundary. Y axis indicates the number of incidence across patient samples. **C.** Differential mRNA expression analysis for genes involved in recurrently changed TAD showing transcription of most genes are not much affected by boundary change.

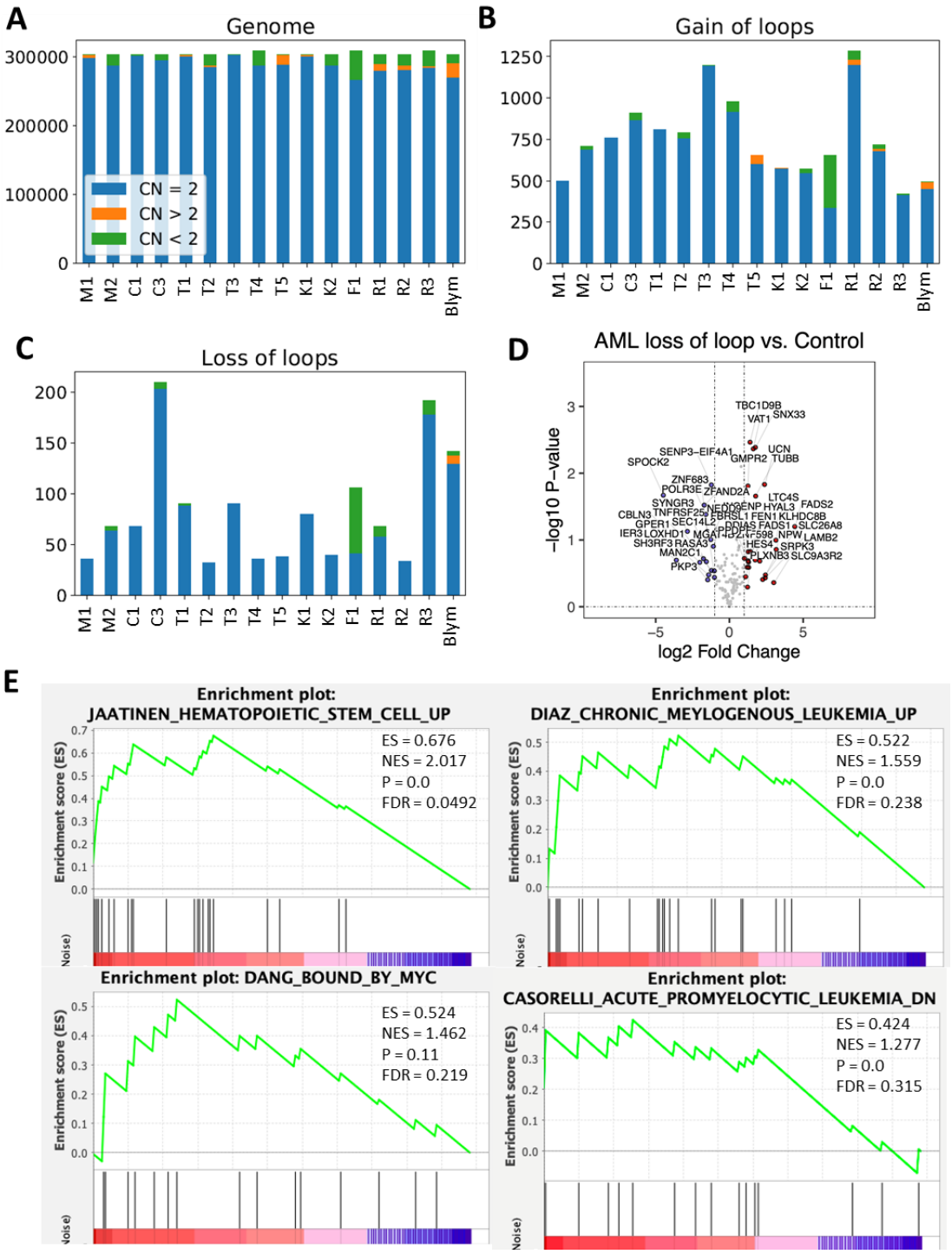


Figure 4-S 4. Differential loop analysis.

A-C. Excluding the confounding effect of CNV in calling differential loops. A shows number of genomic 10Kb bins with neutral, gain, or loss of copies. B shows number of anchors for gained loops in each sample (bar height) and the proportion for detecting gain of copy (orange) or loss of copy (green) in the anchor. C shows the same thing for loss of loops. **D.** Differential mRNA expression analysis for genes involved in recurrently lost loops in AML. **E**. GSEA analysis for differential expressed genes from the gained loops.

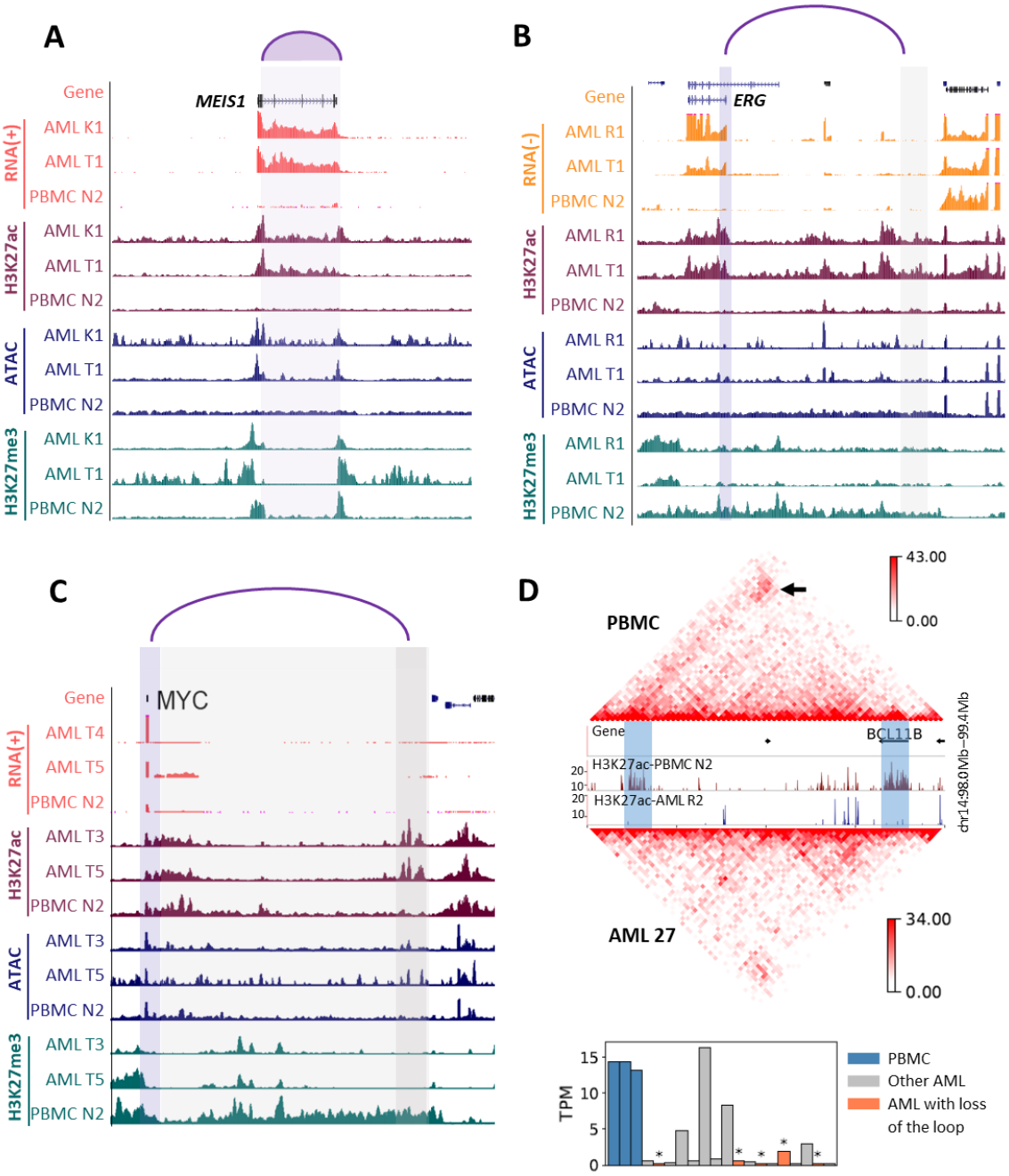


Figure 4-S 5. Correlation of differential loops with gene expression, open and repressive promoter, and distal enhancer.

A-C. Gain of loops correlated with activated expression, ATAC-seq and H3K27ac peaks at the gene (A, B, C) and the distal region (B, C). For each gain of loop, two representative AML samples are plotted, both with gain of loops whereas the demarcated loop is absent in three controls (only one is plotted). The purple arch points to the two loop anchors, with the purple shadow highlighting the affected gene promoter or the entire highly-interacting FIRE, and the grey shadow highlighting the distal regions. For C, the grey region represents the stretch of the whole stripe while the darker grey represents a region with especially high interaction frequency to its promoter. **D**. Loss of a loop (marked by the black arrow) between BCL11B and a distal region (highlighted by the blue shade) in AML samples correlate with loss of active promoter and enhancer mark on the distal region. The lower panels shows that the samples with loss of this loops (marked by orange color and an asterisk on top of the bar) also show dramatic loss of expression.

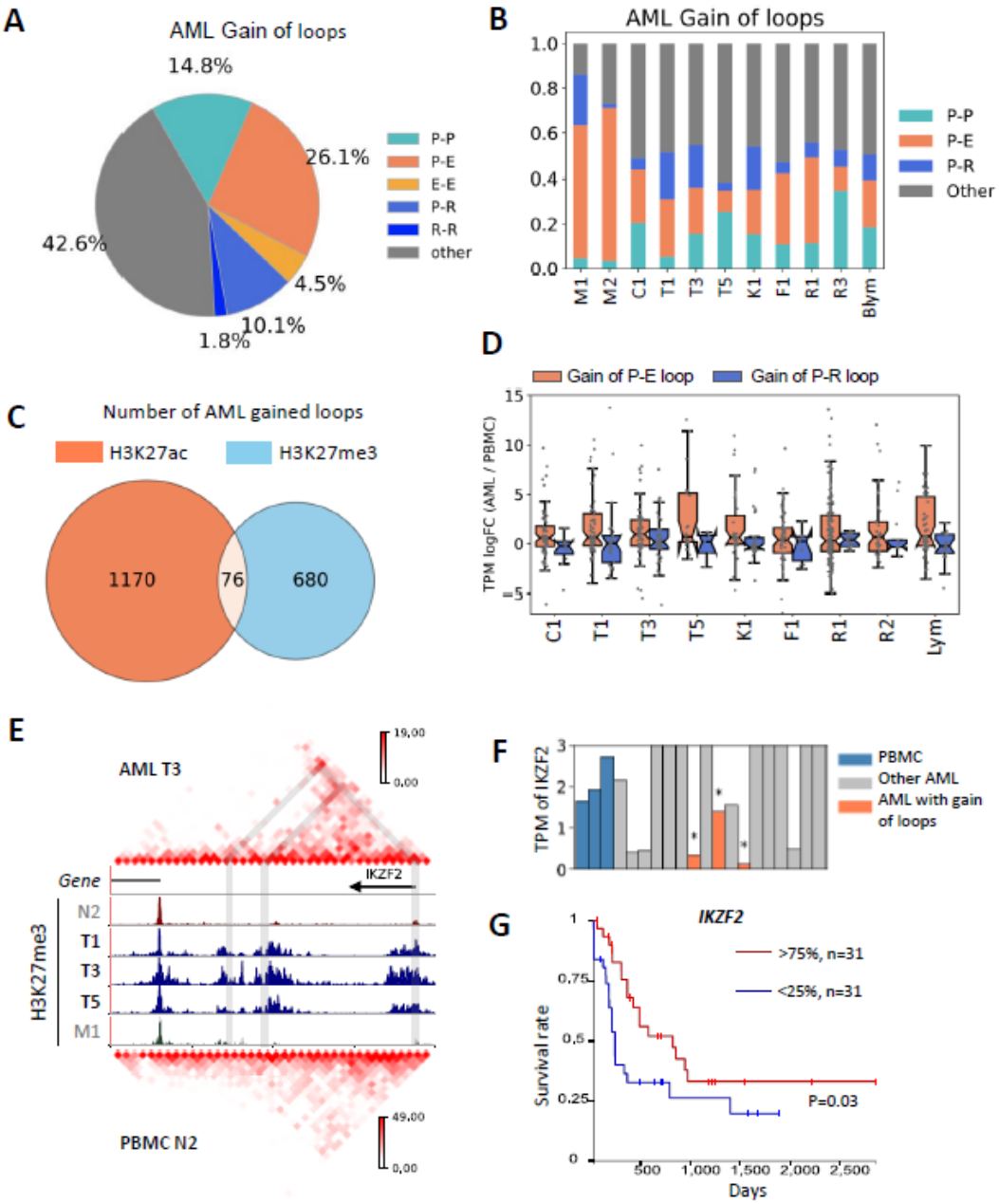


Figure 4-S 6. Component analysis of gained loops showing enhancer loops and repressor loops.

A-B. Percentage for each class of the gained loops across all AML samples (A) and each sample (B). Enhancer and repressor annotation are derived from our CUT&TAG of H3K27ac and H3K27me3 for the samples where loops are counted. Promoters are excluded from enhancer or repressor annotation. If both marks are present in a 10Kb bin, the distal region is defined as enhancer if the peak –log10(P) value of H3K27ac is at least three time higher than that of H3K27me3 peak, and the same for defining repressor. **C.** Number of loops with only enhancer mark, only repressive mark, or both at the distal region. **D.** Transcription fold change for genes that are in gained P-E loops or P-R loops versus the expression of the same gene in control PBMC without the loop. E-G. An example for gain of repressive loop (grey-shaded in E) involving a potential tumor suppressor IKZF2 correlated to lower expression (F) and poorer prognosis in TCGA AML GDC cohort (G). Sample marked with black and bold font in E are those with gained loops.

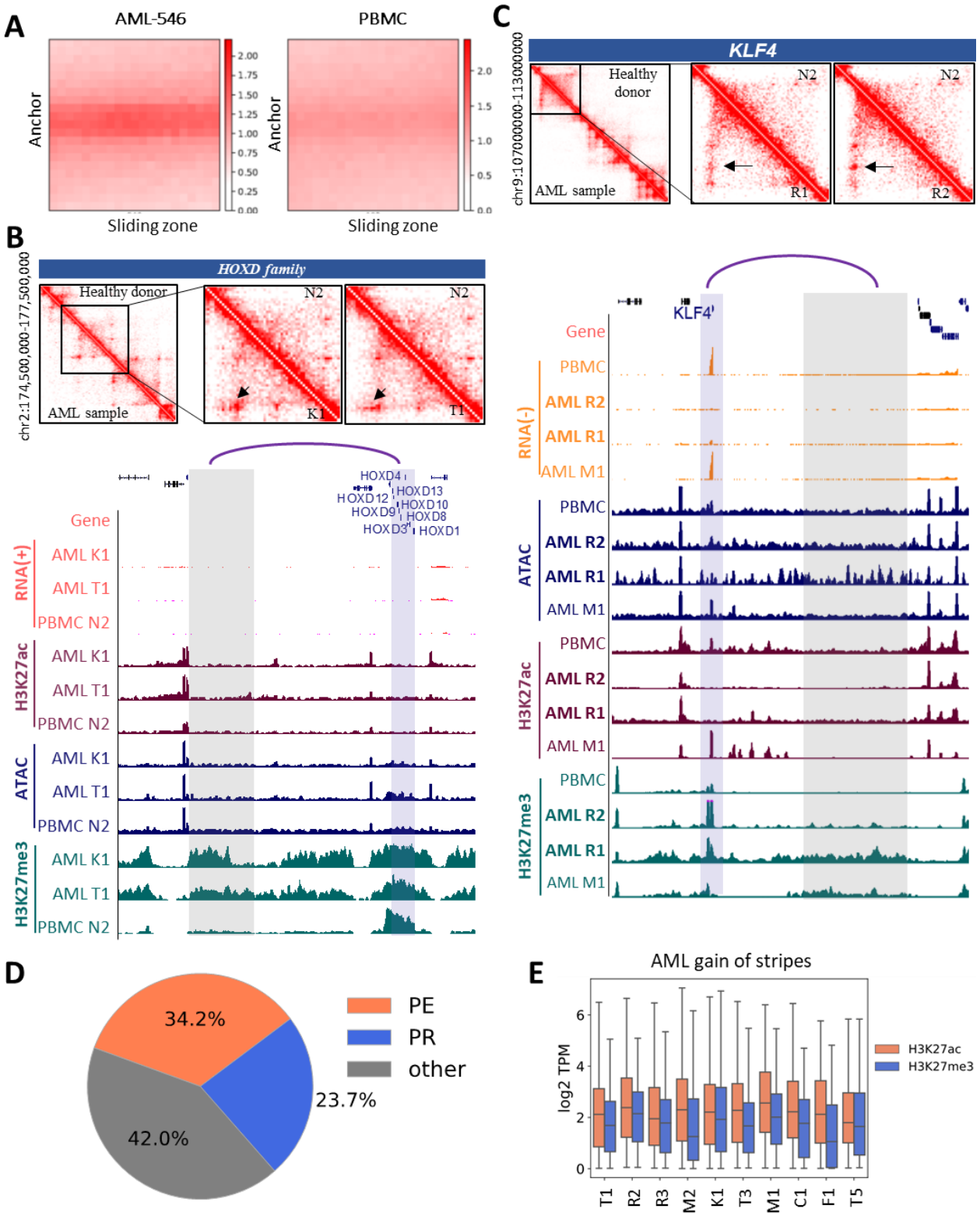


Figure 4-S 7. Stripe identification and characterization.

A. APA plot for regions with detected gain of stripes in AML samples and the same regions in control samples. **B and C.** Gain of repressive stripes for HOXD family and KLF4. The two AML samples in B have gain of stripes whereas only the AML samples with bold font in C have gain of stripe. **D.** Classification of stripes based on the anchored promoter and the histone marks on the stripe zones, annotated by CUT&TAG data from the same sample where stripes are detected. **E.** Genes with enhancer stripes have in average higher expression than genes with repressor stripes.

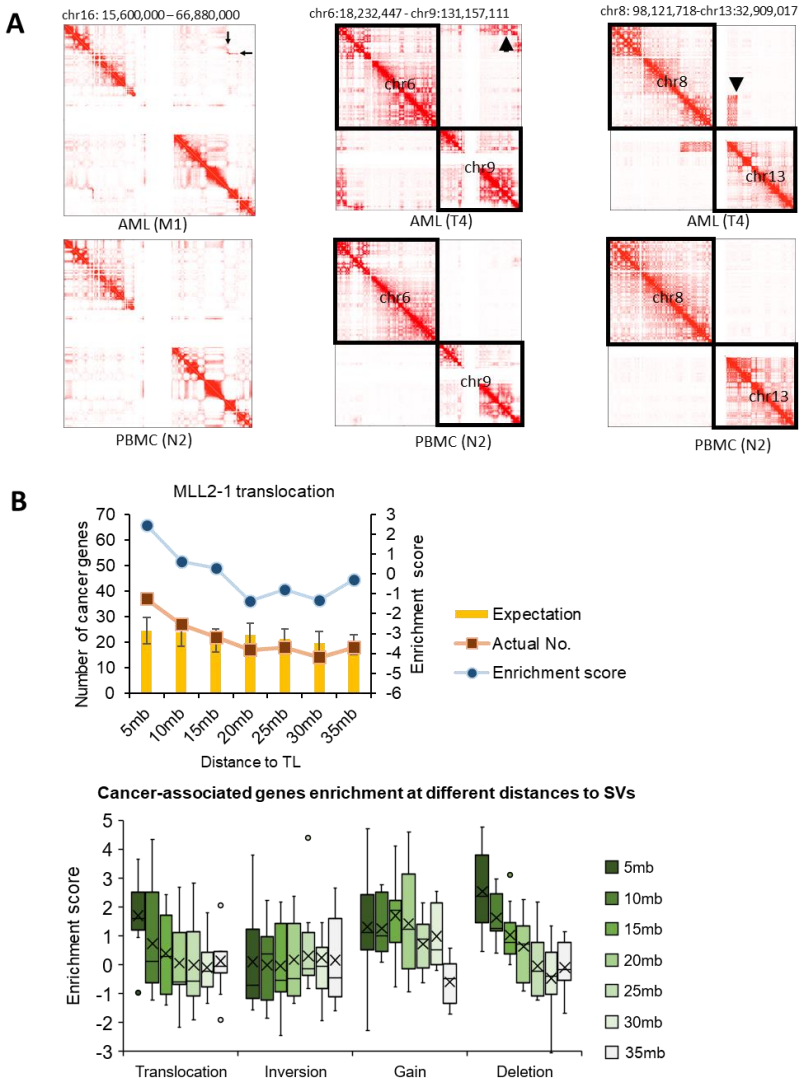


Figure 4-S 8. Sup Figure 8. SVs in AML samples are enriched nearby cancer or AML related genes.

A. Detection of SVs in AML samples from Hi-C data, marked by the black arrow in Hi-C maps. The aberrant signals shows the impact of SVs on chromatin structure.**B.** Higher density of translocations (upper) and other SVs (lower) than expectation nearby the cancer genes. Expectation is calculated by permutation of the cancer related genes in the genome 1000 times, and the number of permutated genes at each distance to the real translocations are counted. Genes are not overlapping with SVs when counted and the distance refers to the shortest distance. The enrichment score is the z-score for assessing the position of the actual number of genes in the distribution of the number of permutated genes. Enrichment larger than 1.96 indicates significant enrichment with 95% confidence.

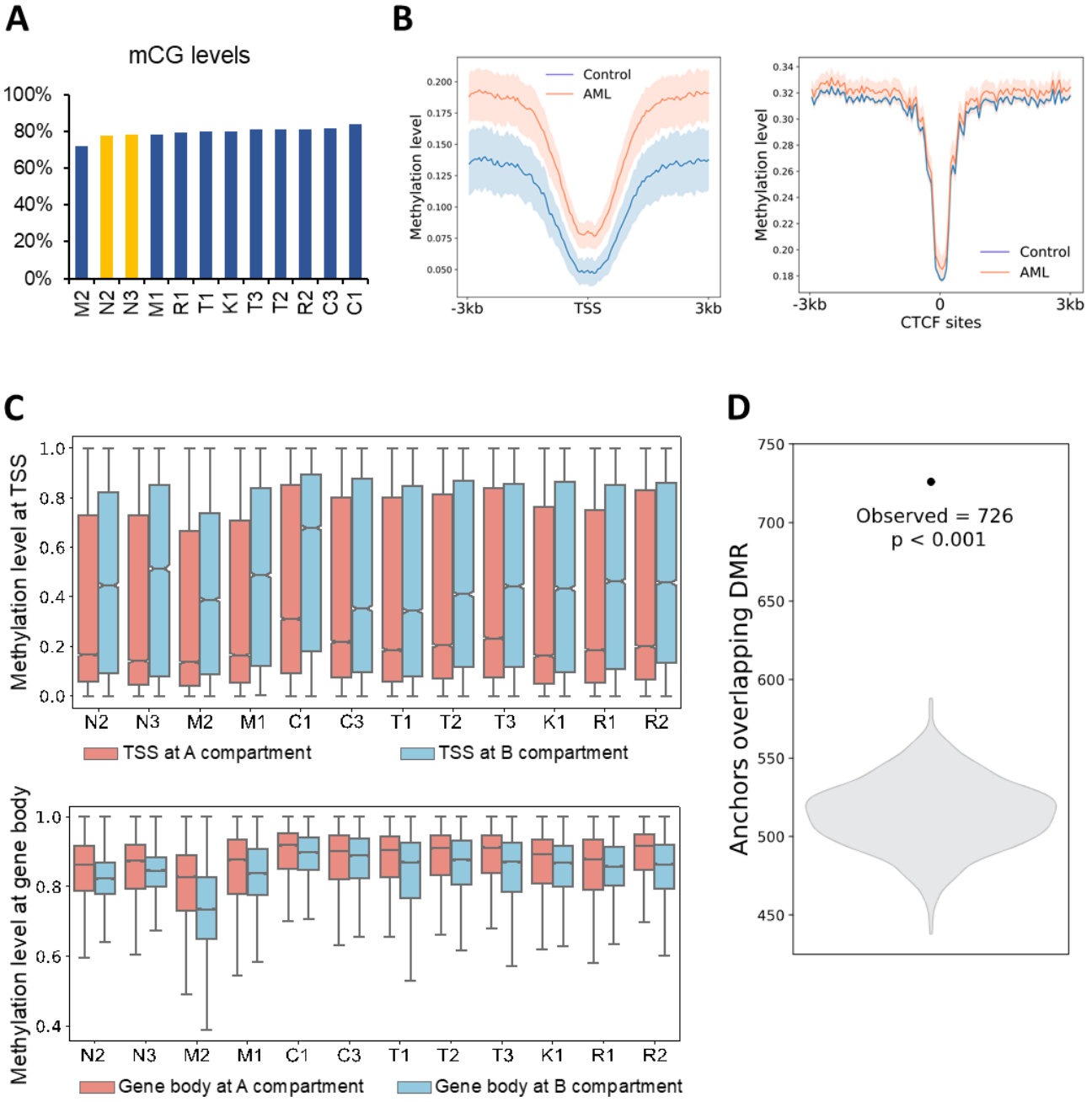


Figure 4-S 9. Association between DNA methylation and chromatin structure

A. Global higher methylation in AML samples (blue) than control (yellow). **B.** DNA methylation per 50bp bins centered at TSS (left panel) or CTCF sites profiled from control samples (right panel). The lighter ribbon shows the 95% confidence interval across all AML or control samples respectively. **C.** Methylation levels at TSS regions (upper) and gene body regions (lower panel) at A and B compartment. **D.** Anchors of differential loops in Figure 4-4C are enriched at DMR. The distribution of overlap is calculated by permutation of differential loop anchors for 1000 times in the genome.

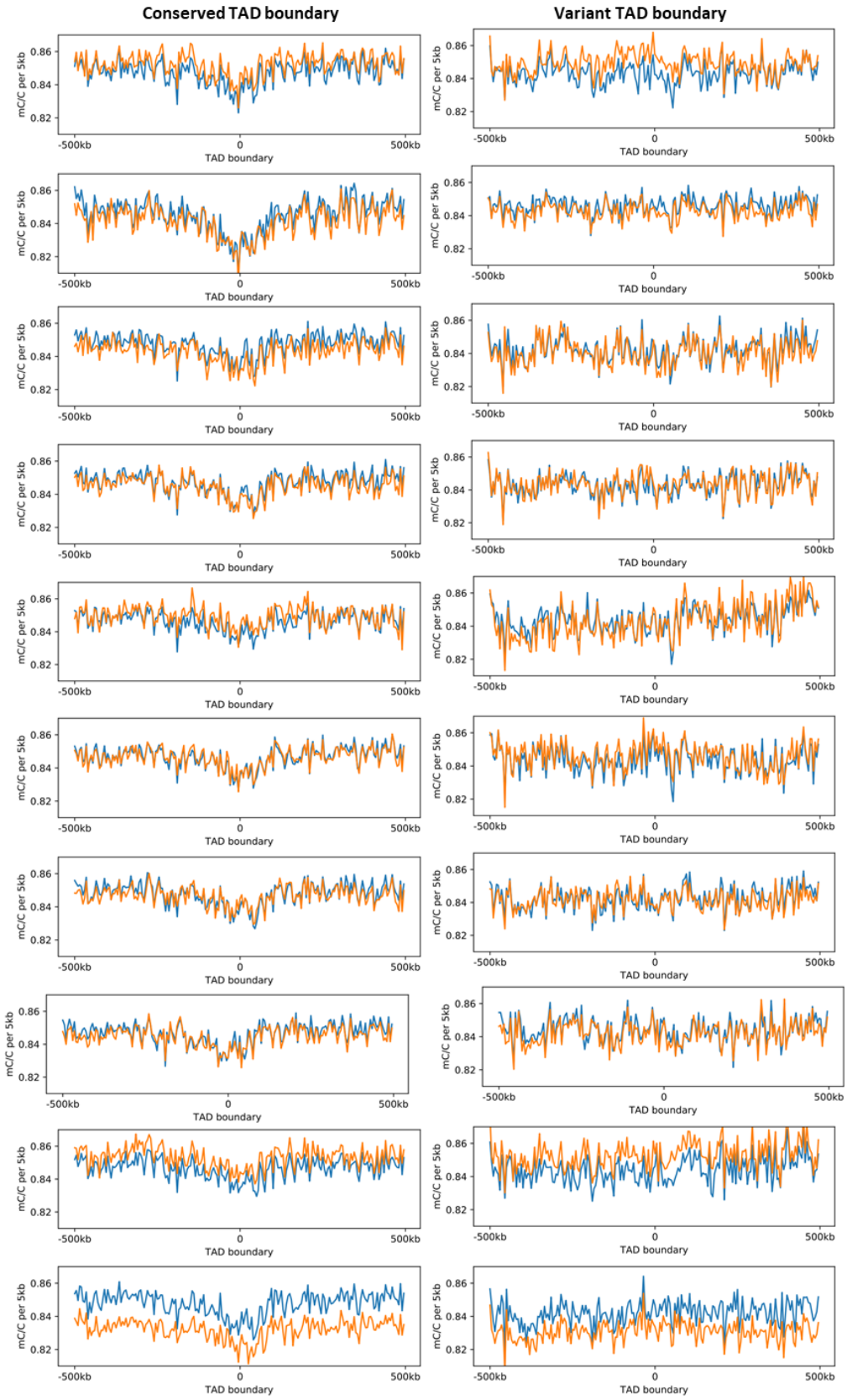


Figure 4-S 10. Conserved TAD boundary exhibits DNA hypo-methylation.

Each row represents one AML sample. The consensus TAD boundary from the three PBMC samples are classified by whether they are consistent with the TAD boundaries profiled from the AML sample, with the conserved TAD boundary in the left panel and the variant ones in the right. The averaged DNA methylation levels in the PBMC samples at their TAD boundaries are plotted with blue color, while the methylation at the same regions in the AML samples are plotted in orange color.

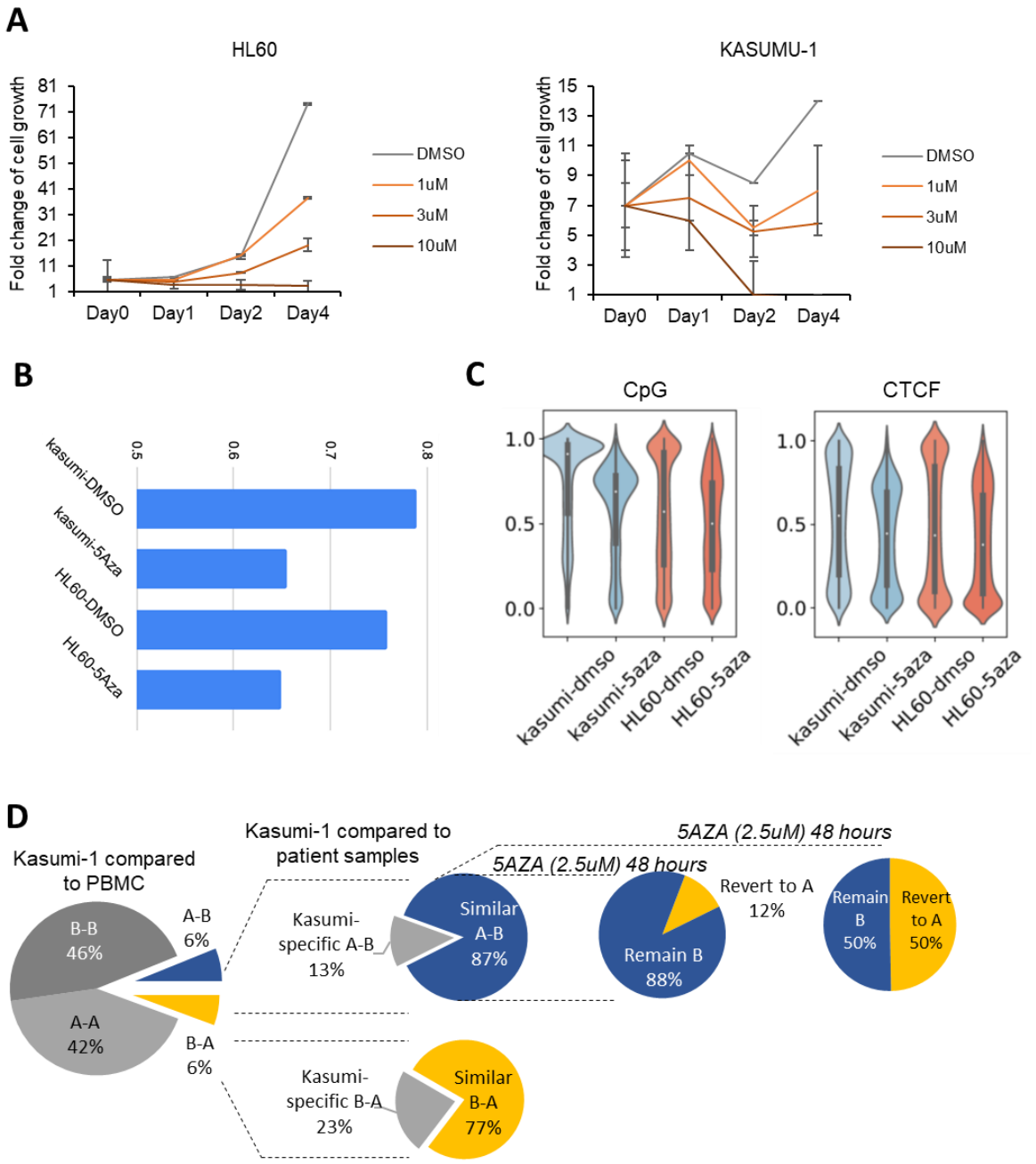


Figure 4-S 11. Results of 5-AZA treatment.

A. HL60 (left) and Kasumi-1 (right) proliferation restrained by 5-AZA treatment. Fold change is calculated by number of viable cells every 24 hour. MTT assay and CCK-8 assay had very similar results (data not shown). **B.** Global DNA methylation levels after 48-hour treatment of 5AZA (1uM for HL60 and 2uM for Kasumi-1). **C.** DNA methylation at top 10,000 most variable CpG sites (left) and top 10,000 CTCF sites with most variable methylation across the indicated four conditions, with the same treatment in B. **D.** Compartment switch and restoration in Kasumi-1 cells with comparison to patient data.

Materials and Methods

Materials and Experiments:

Primary sample collection

Human blood samples were obtained with consent from the AML patient or healthy donors, under the Penn State Hershey IRB-approved protocol. AML peripheral blood or bone marrow aspirates and peripheral blood from healthy donors were collected and immediately subjected to selection of mononuclear cells using Ficoll-Paque PLUS density gradient media (GE Healthcare, 17-1440-02) following manufacturer's instruction.

Cell culture

HL60 (ATCC CCL-240) and Kasumi-1 (ATCC CRL-2724) cells were given to us as a gift from Dr Sinisa Dovat lab, which are new vials purchased from ATCC. Cells are cultured following the manufacturer's culture method. Kasumi is cultured with RPMI-1640 (Gibco, 11875093) with 20% FBS. HL60 is cultured with Iscove's Modified Dulbecco's Medium (IMDM) (ATCC® 30-2005).

In-situ Hi-C

One to two million cryopreserved primary samples or cell lines in cell culture were spinned down with 500g and resuspended in 1ml/million RPMI 1640 medium with 10% fetal bovine serum, immediately crosslinked with 37% formaldehyde (MilliporeSigma 252549) to a final concentration of 2%, and incubated at room temperature for 10 minutes on a tube revolver at 16 rpm to mix, quenched by 2.5M glycine solution with a final concentration of 0.2M and incubated in room temperature for 5 minutes on revolvor. Cells were pelleted by centrifuge at 500g at 4°C for five minutes and washed once with 1ml cold 1X PBS by centrifuge at 500g at 4°C for five minutes and the supernatant was discarded.

Cells were lysed to extract nuclei with 250ul lysis buffer (10mM Tris-HCl pH8.0, 10mM NaCl, 0.2% Igepal CA630) mixed with 50ul 50x protease inhibitor (Sigma, P8340) and incubated on ice for 15 minutes, centrifuge to pellet at 2500g, 4°C for five minutes and washed with 500ul lysis buffer. Cell pellets were resuspended in 50ul 0.5% sodium dodecyl sulfate (SDS) and incubated at 62°C for 10 minutes, quenched by 145ul water and 25ul 10% Triton X-100 (Sigma, 93443), and incubated at 37°C for 15 minutes. 25µl of 10X NEBuffer2 (NEB B7207) and 100 unit of MboI restriction enzyme (NEB, R0147) was then added to the reaction for overnight DNA digestion at 37°C on the tube revolver. The digestion was then quenched by incubation at 62°C for 20 minutes. DNA was then end repaired and Biotin labeled with 50ul fill-in master mix (37.5µl of 0.4mM biotin-14-dATP (Life Technologies, 19524-016), 1.5µl of 10mM dCTP, 1.5µl of 10mM dGTP, 1.5µl of 10mM dTTP, 8µl of 5U/µl DNA Polymerase I, Large (Klenow) Fragment (NEB, M0210) and incubated at 37°C for 1.5 hours.Then DNA was ligated with 900ul of ligration master mix (669ul water, 120µl of 10X NEB T4 DNA ligase buffer (NEB, B0202), 100µl of 10% Triton X-100, 6µl of 20mg/ml Bovine Serum Albumin (MilliporeSigma B8667), 5µl of 400 U/ µl T4 DNA Ligase (NEB, M0202)), incubated at room temperature for 4 hours with slow rotation. Decrosslink DNA by adding 50ul of 20mg/ml proteinase K (QIAGEN 19133) and 120ul of 10% SDS, incubated at 55°C for 30 minutes. Quench the reaction by adding 130µl of 5M sodium chloride and incubate at 68°C overnight. Precipitate DNA by adding 1.6X volume of pure ethanol and 0.1X volume of 3M sodium acetate, pH 5.2 (MilliporeSigma S7899), incubate at -80°C for at least half hour, and centrifuge at max speed, 2°C for 15 minutes to discard supernatant. Wash DNA once with 700ul 80% ethanol. Dissolve dried DNA pellet in 130ul 10 mM Tris-HCl, pH 8. Sonicate the solution to shear DNA to average size of 300-500bp with Covaris sonicator, with parameters set as followings: PIP 140, duty factor 10, burst 200, and duration 58s-80s. Run 4ul sheared DNA in 16ul water on a 2% agarose to verify the size.

Pull down biotin-labeled DNA by washing 150µl of 10mg/ml Dynabeads MyOne Streptavidin T1 beads (Life technologies, 65602) with 400µl of 1X Tween Washing Buffer (TWB: 5mM Tris-HCl (pH 7.5); 0.5mM EDTA; 1M NaCl; 0.05% Tween 20), discard the solution. Resuspend the beads in 300µl of 2X Binding Buffer (10mM Tris-HCl (pH 7.5); 1mM EDTA; 2M NaCl) and add to the sheared DNA. Incubate at room temperature for 15 minutes with rotation. Separate beads and discard the supernatant with a magnetic rack. Wash beads with 600ul TWB buffer twice. End repair of sheared DNA by resuspending beads in 100ul 1X NEB T4 DNA ligase buffer (NEB, B0202), separating the beads, and resuspending in end repair master mix (88µl of 1X NEB T4 DNA ligase buffer with 10mM ATP (NEB B0202S), 2µl of 25mM dNTP mix, 5µl of 10U/µl NEB T4 PNK (NEB, M0201), 4µl of 3U/µl NEB T4 DNA polymerase I (NEB, M0203) , 1µl of 5U/µl NEB DNA polymerase I, Large (Klenow) Fragment (NEB, M0210)), and incubation at room temperature for half hour. Beads were washed twice with 500ul TWB buffer and resuspended in 100µl 1X Quick ligation reaction buffer (NEB, B6058), recollected, and proceeded with dATP attachment by resuspended in 100µl master mix (90µl of 1X NEBuffer 2, 5µl of 10mM dATP, 5µl of 5U/µl NEB Klenow exo minus (NEB, M0212)), incubated at 37°C for 30 minutes. Beads were washed twice with 500ul TWB buffer and resuspended in 100µl 1X Quick ligation reaction buffer (NEB, B6058), recollected, proceeded with adaptor ligation through resuspension in 50µl of 1X NEB Quick ligation reaction buffer, 2µl of NEB DNA Quick ligase (NEB, M2200), and 3ul of Illumina adaptor of choice, incubated in room temperature for 15 minutes. Beads were washed by 600ul TWB buffer and 100ul 1X Tris buffer, resuspended in 50ul 1X Tris buffer, heated on 98°C for 10 minutes to elute the DNA off the beads. Beads were discarded. Size selection was performed to remove small DNA fragments by adding 0.8X-0.9X KAPA beads to the DNA elution, incubation at room temperature for 5 minutes, and beads were collected with supernatant discarded. Wash beads twice with 500ul 80% ethanol and elute beads in 50ul 1X Tris buffer. Library amplification

was performed with 4-12 cycles of PCR with KAPA 2X library mix. Size selection was performed to remove small and large fragments using KAPA beads and maintain DNA fragments of 150bp-500bp. Libraries were sequenced as 150 bp paired-end reads with a raw sequencing depth between 300 million to 700 million read pairs per sample on platform Hiseq Xten or Novaseq.

CUT&TAG

The CUT&TAG experiments were performed exactly following the online protocol[226] [citation]: **https://www.protocols.io/view/bench-top-cut-amp-tag-z6hf9b6?version_warning=no.** For each targeted protein we use 0.1 million. We use the following primary antibodies: CTCF (Active motif 2899), H3K27ac (Active motif 39133), H3K27me3 (Cell signaling C36B11), Rabbit IgG (Cell signaling 2729), and the Guinea Pig anti-Rabbit IgG (H+L) secondary antibody (NBP1-72763). The pA-Tn5 fusion protein was kindly provided as a gift from Dr Steven Henikoff Lab. Final libraries were sequenced as 150 bp paired-end reads on platform Novaseq or Hiseq Xten, with a raw sequencing depth between 10 to 20 million read pairs.

ATAC-seq

ATAC-seq was performed following the published protocol with minimal modification[227]. Briefly, we also centrifuge down 50,000 cells, wash the cells with PBS, and perform nuclei extraction with cold lysis buffer. We added an extra step of washing the nuclei with another 500ul of lysis buffer to further remove mitochondrial DNA. We then proceed with the transposition reaction (Illumina Tagment DNA Enzyme and Buffer Large Kit 20034198) and purification steps (QIAGEN MinElute PCR Purification Kit Cat No./ID: 28004) . We elute the DNA in 20ul elution buffer instead of 10ul elution buffer to increase recovery rate. Then we proceed with the PCR amplification step, where we use 20ul transposed DNA, 2.5ul of Nextera PCR primer 1 and 2.5ul of Nexteral primer 2, and 25ul KAPA

HiFi HotStart Ready Mix master mix (KAPA KR0370). We use the PCR parameters indicated by the standard protocol with 11 cycles. We then perform size selection to remove small fragments using KAPA pure beads. We add 45ul KAPA beads to 50ul PCR solution, incubate for 15 minutes in room temperature, and use a magnet to capture the beads and discard the supernatant. We wash the beads with 200ul of 80% ethanol twice and remove the ethanol. We resuspend the beads in 20-50ul of prewarmed 10mM Tris-HCl, pH 8.0, incubate at 37 for 10 minutes, and use a magnet to recollect the supernatant as the final library. We use 1ul of the library to run on a 2% agarose gel to verify the footprint nucleosomes for successful assay. Libraries were sequenced as 150bp paired-end reads on platform Hiseq 4000 with 20 million raw read pairs per sample.

PCR-free whole genome sequencing:

Genomic DNA was isolated by QIAGEN DNeasy Blood & Tissue Kits (69504) using 0.5 million cells. Concentration was detected by fluorometer or Microplate Reader (e.g. Qubit Fluorometer, Invitrogen). Sample integrity and purity were detected by Agarose Gel Electrophoresis. 1μg genomic DNA was fragmented by Covaris. KAPA pure Magnetic beads (KK8000) was used to select DNA fragments with an average size of 300-400bp. DNA was quantified by Qubit fluorometer. The Fragments were subjected to end-repair and then was 3’ adenylated. Adaptors were ligated to the ends of these 3’ adenylated fragments.The double stranded products were heat denatured and circularized by the splint oligo sequence. The single strand circle DNA were formatted as the final library.Library was qualified by the Agilent Technologies 2100 bioanalyzer. The library was amplified to make DNA nanoball (DNB) which have more than 300 copies of each molecule. The DNBs were loaded into the patterned nanoarray and sequenced as 150bp paired-end reads by combinatorial Probe-Anchor Synthesis (cPAS).

Whole genome bisulfite sequencing

DNA bisulfite treatment was performed using the EZ DNA Methylation-Gold Kit (catalog D5005, Zymo Research Corporation) according to the manufacturer's instructions. The recovered bisulfite-converted single-stranded DNA was processed for library construction using the Accel-NGS@Methyl-seq DNA Library kit (catalog 30024, Swift BioSciences) as per manufacturer instructions. Briefly, using the Adaptase module, truncated adapter sequences were incorporated to the single-stranded DNA in a template-independent reaction through sequential steps. DNA was then enriched using 6 cycles of PCR with primers compatible with Illumina sequencing. The quantity and molecular size of the library was confirmed by Qubit HS DNA assay (ThermoFisher) and Tapestation 2200 system coupled with High Sensitivity D1000 ScreenTapes (Agilent). Illumina 8-nt dual-indices were used for multiplexing. Samples were pooled and sequenced on Illumina NovaSeq S4 sequencer for 150 bp read length in paired-end mode, with an output of 580 million reads per sample.

RNA-seq

RNA was extracted using QIAGEN Rneasy Plus kit (74034). RNA quality was assessed by Agilent RNA ScreenTape on Agilent 2200 Tapestation and quantified by Qubit. The mRNA was enriched by poly-A selection and the second strand synthesis was performed with NEBNext Ultra II Non-Directional RNA Second Strand Synthesis Module following manufacturer’s instructions (E6111S). Average final library size is between 380-400 bp. Illumina 8-nt dual-indices were used for multiplexing. Samples were pooled and sequenced on Illumina HiSeq X sequencer as 150 bp paired-end reads, with an output of 40 million reads per sample.

5-AZA treatment

5-azacytidine (MilliporeSigma A2385-100MG) was dissolved in DMSO to make 100mM stock solution, aliquoted and stored in -80°C. Working solutions (0.5uM-10uM) was made from further dilution of stock solution using complete cell culture media. Media was changed every 24 hours with freshly made 5AZA. Dead cells are removed by Ficoll-Paque PLUS density gradient media at the end of cell culture before cells are further processed for any profiling experiments.

Proliferation assay

Cell proliferation was independently performed and replicated by viable cell count with trypan-blue staining, MTT assay (abcam ab211091) and CCK-8 assay (ApexBio K1018) every 24 hours following the manufacturer's instruction.

Informatics analysis

Point mutation and structural variants analysis

WGS reads were first aligned to human genome reference GRCh38 with BWA MEM (v0.7.17-r1198). PCR duplicates were removed by Sambamba (v0.7.0) [190]. Uniquely mapped (MAPQ > 20) reads were retained for downstream variant detection. Point mutations including single nucleotide mutations and small indels were detected by Freebayes (v1.2.0-17-ga78ffc0) with parameters "--min-alternate-count 2 --min-alternate-fraction 0.05 --min-repeat-entropy 1", and minimal quality score of 20 was used to reduce false positive calls. To minimize the number of germline calls, point mutations that overlap with dbSNP150 mutations were removed. The functional effects of the filtered variants were annotated with SnpEff (v4.3T) [228]. Only variants annotated as high or moderate impact by SnpEff were used for downstream analysis. This final set of point mutations further confirm the clinical

molecular diagnosis of our AML samples, and both results were used for optimizing the subtyping of the AML samples.

Structural variants (SVs) were detected using WGS and Hi-C data as previously described[32]. Delly (v0.7.7) [116] and Speedseq (v0.1.2) [191] were used for detecting SVs in WGS data. Centromere, telomere and heterochromatin regions were excluded for SV detection. SV calls from Delly and Speedseq were merged and only SVs detected by both methods were kept to reduce false positives. Furthermore, the detected SVs were compared against the DGV database (version.2016-05-15) to reduce germline SV calls. For SV detection in Hi-C data, HiC Breakfinder [32] was used with default parameters. In this work, only large deletion, inversion (> 1Mb) and inter-chromosomal translocations were considered for Neo-loop analysis.

Hi-C data analysis

Paired-end reads were first trimmed by Trim_Galore! (v0.6.0) to remove adapters and low quality bases with parameter "--paired". Trimmed reads were then mapped to human genome GRCh38 by BWA MEM (v0.7.17-r1198) with parameter "-SP5M", and deduplicated with "pairtools dedup" (v0.3.0). Reads mapped to the same MboI restriction fragment or in outward direction are not informative to chromatin interactions and thus are removed for downstream analysis. Hi-C matrices and cooler files were generated with Cooler (v0.8.6.post0) [229]. Iterative correction and eigenvector decomposition (ICE) method was used for Hi-C normalization with "cooler balance" option. Higlass, Juicerbox, and 3D Genome Browser (http://3dgenome.org/) were used for visualization of Hi-C matrices[230-232].

HiC compartments were identified at 40kb resolution using a "sliding window" strategy as previously described [85]. First, the "exp" (expected) matrix was obtained by averaging Hi-C contacts at

the same distance. Then the “obs/exp” matrix was calculated by summing the observed Hi-C contacts within a window of 400kb centered at each bin divided by the sum of expected Hi-C contacts in the same window. A step size of 40kb was used to calculate the “obs/exp” value for all elements in the matrix. The “obs/exp” matrix was then converted to a Pearson Correlation matrix. The principal components were derived by calculating the covariance matrix of the Pearson Correlation matrix followed by eigenvector decomposition with ‘eigen’ function in R. The first principal component (PC1) was used to assign the A and B compartment where regions with positive PC1 values correspond to A compartment and negative to B compartment based on their association with gene density.

Topologically associating domains (TADs) were identified at 40kb resolution by DomainCaller[61]. Specifically, a Directionality Index (DI) was calculated for each genomic bins at a window size of 2Mb. Then a Hidden Markov model was used to determine the up-or-downstream biased status for each genomic bin based on the DI scores. TADs were defined as continuous genomic regions starting from the first bin of a series of consecutive downstream biased bins to the last bin of the next series of consecutive upstream biased bins.

Loop domains were identified at 10kb resolution using Peakachu, a machine-learning based method recently developed in our lab[213]. Peakachu reports a probability score associated with each loop which demarcates the likelihood of the loop being real. The probability score is also associated with the loop intensity, making it convenient for differential loop analysis. Peakachu detects loops as CTCF or H3K27ac loops depending on a pre-trained CTCF model or H3K27ac model. In this work, loops were first detected using both models and only those detected by both models were reported to decrease the false positive loops.

SV-induced neo-loop identification

Neo-loop formation induced by SVs were identified using npcaller developed by our lab (method not published yet).

CUT & Tag data analysis

CUT & Tag sequencing reads were processed using ENCODE ChIP-seq pipeline (https://github.com/ENCODE-DCC/chip-seq-pipeline2). Specifically, reads were first trimmed by Trim_Galore! with "--paired" option, and then aligned to human genome reference GRCh38 with Bowtie2 (v2.3.5.1)[233]. PCR duplicates were removed by Picard MarkDuplicates tool with "VALIDATION_STRINGENCY=LENIENT" option. MACS2 was used for peak calling for histone marks and TFs with "-p 1e-2 --nomodel --shift 0 --keep-dup all -B --SPMR" options. Peaks in the ENCODE hg38 blacklist regions (http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/hg38-human/hg38.blacklist.bed.gz) were filtered out. Peaks with MACS2-reported p value < 10e-5 were retained for downstream analyses. Log transformed p values (-log10 p-value) were used for track visualization in University of California Santa Cruz (UCSC) genome browser and Integrative Genomics Viewer (IGV).

Whole genome bisulfite sequencing data analysis

WGBS reads were processed using the Bismark pipeline[234]. Reads were first trimmed by 10 base pairs on the 5' end of both forward and reverse reads using Trim_Galore! with "--paired --clip_R1 10 --clip_R2 10" options. Trimmed reads were then mapped with "bismark" command and deduplicated using Bismark's "deduplicate_bismark" tool with default parameters. Per-cytosine methylation level was obtained using methylpy[235] "call-methylation-state" tool with "--binom-test True --paired-end True" options. The average CpG methylation level for a genomic region was calculated as the

accumulation of methylated reads over all CpG sites divided by the total reads in the region. Region-level methylation levels were normalized by global methylation levels of each sample.

ATAC-seq data analysis

ATAC-seq sequencing reads were processed using ENCODE ATAC-seq pipeline (https://github.com/ENCODE-DCC/atac-seq-pipeline). Specifically, reads were first trimmed by Cutadapt (v2.4) with "-m 5 -e 0.2" options. Trimmed reads were then aligned to human genome reference GRCh38 with "bowtie2 -X2000 --mm" and deduplicated using Picard MarkDuplicates tool with "VALIDATION_STRINGENCY=LENIENT" option. Read alignments were shifted +4bp on "+" strand and -5bp on "-" strand to account for Tn5 insertion before peak calling. Peaks were called by MACS2 with "--shift -75 --extsize 150 --nomodel -B --SPMR --keep-dup all --call-summits" options, and filtered against the ENCODE hg38 blacklist (http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/hg38-human/hg38.blacklist.bed.gz). Peaks were then filtered by MACS2-reported p value ($p < 10e-5$). Peak summits were extended by 250bp on both sides to a final width of 500bp for all downstream analyses. Overlapped peaks were handled using an iterative-removal approach similarly as previously described[236]. The most significant peak was examined, and any peaks directly overlapping with it were removed. Then this process iterates to the next most significant peak until all peaks are not overlapped. We performed a "score per million" normalization of MACS2 peak scores (-log10 p-value) by dividing each individual peak score by the sum of all peak scores in the sample divided by 1 million.

RNA-seq data analysis

RNA-seq sequencing reads were analyzed following the ENCODE standard pipeline. Raw sequencing reads were first adapter-trimmed by Trimm_Galore! with "--paired" option, and aligned to

human genome reference GRCh38 with STAR (v2.5.3a_modified)[195] with "--outSAMunmapped Within --outFilterType BySJout --outFilterMultimapNmax 20 --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --sjdbScore 1" options. RSEM (v1.2.31)[196] was used for transcript quantification with GENCODE v24 annotation and "--paired-end --estimate-rspd --calc-ci" options. TPM for all transcripts were quantile-normalized across all samples using the "normalize.quantiles" function in the "preprocessCore" library in R.

Hi-C based clustering of AML samples

The first principal component (PC1) of the Hi-C matrix was used for unsupervised clustering of AML samples. PC1 was calculated at 40kb resolution from the ICE-normalized Hi-C data. Then coefficients of variation of PC1 across all AML samples and PBMC controls was calculated for each 40kb genomic bins. The top 10% bins (n = 7013) with the largest variation of PC1 were selected for deriving a correlation matrix of PC1 across all samples. Hierarchical clustering was performed on the correlation matrix using "complete" linkage and "euclidean" distance metrics. The top 10% most variable bins were further clustered by hierarchical clustering to show sub-type specific A/B compartment patterns.

Identification of TAD alteration in AML

To identify TAD alteration in AML, a list of 898 TADs that are conserved (reciprocal overlap > 0.9) all three PBMC was first compiled. Then this list of conserved TADs were compared against the TADs in each AML sample to generate a per-sample altered TADs list. TADs that are altered in more than one AML samples were defined as recurrent altered TADs. Expression of genes in samples with altered TADs was compared with the same genes in samples without altered TADs to identify up- and

down-regulation of gene expression due to recurrent TAD alteration. The contacts between the promoter of representative deregulated genes and near regions were closely examined with virtual 4C plot. In virtual 4C plot, contacts were first averaged across AML samples and PBMC samples respectively, and then normalized as a proportion to the contacts within the anchor itself.

Identification of differential loops in AML

AML gained or lost loops specific to PBMCs were identified based on the loop probability from Peakachu with a Gaussian mixture model. First, Peakachu loops for each pair of AML and PBMC individuals were merged and deduplicated. Then for each of the merged list of loops, fold change of Peakchu probability between AML and PBMC, and its reciprocal were used as input for a Gaussian mixture model to determine significantly differential loops at a FDR of 0.05. Each AMl individual was compared to all three PBMC controls independently, and then the consensus from all three lists of differential loops were deemed as AML-specific gain or loss of loops. Differential loops specific to more than two AML individuals were defined as recurrent gain or loss of AML loops. Expression of genes in AML samples with recurrently differential loops were compared with the same genes in PBMC samples to identify dysregulation of gene expression due to differential loop formation.

Identification of AML subtype-specific loops

Peakachu loops from all AML and PBMC samples were first merged and deduplicated to form an ensembl list of 113,632 loops. Then a loop probability matrix was constructed for the ensembl loops across 13 AML samples and three PBMC samples. Four AML samples (M3, C2, R3, and F1) were excluded for this analysis due to significantly fewer loops detected compared with other samples. Loops that did not overlap with any gene promoter were removed, which left a total of 69,389 loops for downstream analyses. We then applied additional filters such that we required the maximum subtype-

wise average of loop probability is two-fold greater than the minimum subtype-wise average of loop probability. Then rows (loops) of the probability matrix were clustered by hierarchical clustering with "complete" linkage and "euclidean" distance.

Identification of most variable methylation CpG islands and CTCF sites

CpG island (n = 31,144) coordinates for hg38 were downloaded from UCSC Table Browser. We estimated mCG levels for each CpG island in 10 AML individuals and two PBMC controls. We then calculated the standard variation of mCG levels for each individual CpG island across all twelve individuals. We chose the top 10,000 CpG islands with highest variation among all sites. For CTCF sites, we first merged CTCF peaks from the same twelve individuals through the dba.count() function in the DiffBind R package. We further restricted CTCF peaks to those overlapped with CTCF motifs, resulting in 21,532 CTCF peaks. Similar to CpG islands, the mCG levels for each CTCF peak was generated and then top 10,000 most variable CTCF sites were selected.

Analysis of differentially methylated regions

Differentially methylated regions (DMRs) were identified by methylpy DMRfind. Samples within the same subtype were treated as replicates for DMR identification. A mCG matrix of all DMRs across AML and PBMC samples was constructed and row-wise scaled to be centered at 0, followed by hierarchical clustering on all DMRs.

Analysis of mCG at differential CTCF binding sites

Differential CTCF binding sites between AML and PBMC samples were identified by DiffBind. Additional filter was applied such that differential CTCF binding sites are overlapped with CTCF

motifs. We required hypermethylated CTCF sites to have at least 1.5 fold higher mCG levels in AML than in PBMC, as well as mCG in PBMC greater than 0.1.

Correlation of loss of CTCF binding with gain of loops in AML

We first identified AML-specific loss of CTCF binding compared with PBMC by DiffBind such that the normalized CTCF peak signal is at least two folds greater in PBMC than in AML. Next, within the AML-specific gain of loops that we previously defined, we searched for loops whose two anchors are at the opposite sides of the AML-specific loss of CTCF sites. This gives us a list of loop-CTCF pairs. We then calculated the Pearson correlation of loop intensity (measured by Peakachu probability) with CTCF peak signals for each loop-CTCF pair.

Saddle plot of Hi-C data with DMSO and 5-AZA treatment

We generated HiC saddle plots following guides from cooltools (https://github.com/mirnylab/cooltools). Specifically, the first eigenvector (E1) calculated at 25kb resolution was ranked ascendingly and then equally binned into 38 groups. Next, each genomic bin was assigned to these groups based on its E1 value. We then calculated the average of "obs/exp" contacts for bins in each group with bins in all other groups.

Chapter 5

Overall discussion

Summary and innovation

SV Identification has been challenging due to the many limitations we discussed about in the introduction, and so is the downstream analysis that aims to characterize and interpret the SVs. Therefore, one focus of our work is to develop advantageous new method, with the overall goal to improve SV detection and to better understand the functional impact of SVs.

In Chapter 2, we showed that we are the first effort that uses BioNano optical mapping to identify SVs in cancer genome, and it exhibit great advantage in identifying large SVs or complex SVs. Moreover, we were able to correct the sizes of gaps in human reference genome, and we revealed previously unrecognized polymorphism in gap sizes in human genome. For large gaps (>50Kb), this is unlikely achievable by any other technologies except BAC scaffolding. We were also able to pinpoint the contigs in human reference genome that are mis-position or inversely placed, benefitted from the megabase-scale long reads from BioNano. This findings could greatly improve the accuracy of the human reference genome. Second, we are also the first work that developed the algorithm and software for using Hi-C to detect SVs. We showed its robustness in detection of translocations or large intra-chromosomal SVs, since its performance will not be interfered by complex genomic sequences around the SV breakpoints. We also took an unprecedented trial to integrate WGS, Hi-C and BioNano for SV detection, and quantitatively and systematically compared their performance, with extension to PET-seq, karyotyping and fusion of transcription. One by one we revealed their advantages and limitations with regard to resolution, mappability, SV size and complex SVs. Based on the integration, we demonstrated

how to reconstruct the haplotype-resolved local genome with a series of chained SVs. We accomplished this tough mission for a very complicated region in K562 cells as a proof of concept. Importantly, all previous genomic phasing-related work has been focusing on SNVs due to the challenge of SV phasing.

With the successful experience of the integrative detection of SVs on the model ENCODE cancer cell lines, we then applied BioNano optical mapping onto primary leukemia samples, as shown in Chapter 3. We showed that we were able to capture thousands of SVs missed by routinely used methods in clinical setting. A realistic challenge for somatic SV characterization in leukemia is the lack of normal blood controls, so we developed a pipeline to filter germline SVs, which is also the first solution to SV stratification for BioNano data. We revealed novel genes that are frequently mutated by SVs, which were either previously known as generally related to cancer but not specifically related to AML or not recognized for any association to cancer. Nevertheless we uncovered that the expression of 15 such genes are significantly associated with survival of AML patients.

From these two projects, we demonstrated the importance of SVs at non-coding regions in the cancer genome by disrupting cis-regulatory elements like enhancers or insulator. This is also an innovative endeavor in characterizing non-coding SVs genome-wide in many cancer genomes. We found SVs on enhancers are functionally highly associated with the specific cancer type. We for the first time showed that the cancer-related genes that are located nearby the SVs often have extremely high or low expression in leukemia, and further analysis revealed the allelic unbalanced expression, indicating the heterogenous SVs are disrupting cis-regulation of the nearby genes. Further, we presented a new finding that SVs will alter the replication timing. Last, we created a way of reconstructing Hi-C maps according to information of the SVs, in order to analyze and visualize impact of SVs on 3D genome structure. We illustrated that SV can cause fusion of TADs and activation of tumor-related oncogene.

In chapter 4, we presented a first investigation of chromatin spatial reorganization in primary acute myeloid leukemia in a subtype-specific manner. We also identified and revealed the intriguing co-occurrence of sub-type specific cis-regulatory elements, specifically enhancer, repressors, and CTCF bindings sites, with the gained or lost chromatin loops in AML, providing evidences that the physical chromatin structure accommodates the subtype-specific cis-regulation. Importantly, repressive loops is understudied and less appreciated for their function before. This is the first time that genome-wide gain of pathogenic repressive loops have been identified, which is associated with global transcription downregulation. This is also the first study that identifies gain of stripes genome-wide, and in addition to the initial work that demonstrated the enrichment of super enhancers at stripes, we also showed that stripes can be a hub for associating repressive loops.

This work also for the first time systematically identified the SV-induced neo-loops and enhancer hijacking genome-wide using our newly developed method. We investigated the association between aberrant DNA methylation and altered chromatin structure in depth. We revealed how the methylation abolishes chromatin insulation and causes the excessive gain of chromatin loops. We also uncovered the unexplored mechanism of the chemical 5-AZA in AML treatment. While 5-AZA can activate gene expression through demethylating promoters, why 5AZA can suppress oncogenes and when it will be useful is understudied. Now we showed that through restoring compartment switch and dissociating active chromatin loops, it can also silence oncogenes. This provides a new angle to determine other targets and to develop similar treatment.

Significance and contribution to the field

First, the work presented in Chapter 2 has great resource value to the whole community. We generated 14 Hi-C, 8 optical mapping, 7 WGS, 7 replication timing and 2 karyotyping for ENCODE cancer cell line that are fully accessible to any future work. The comprehensive SV list we generated has become an important reference for researchers to check the genomic background for the cell lines of interest. For example, in design of a CRIPSR experiment, it is critical know the copy number, the SNVs and SVs in the targeted region, to avoid off-target consequence. Also, as SVs extensively drive transcriptional alteration, without knowing the genomic background a lot of data will be hardly interpretable.

Different technologies read SVs differently and many times their outputs are not readily comparable, due inconsistent SV definition or resolution. We provided to the community a detailed standardized pipeline for platform integration. Also, as we revealed the advantage and limitation of each technologies, researchers can more easily make a decision for what technique is the best fit for their purpose. Moreover, our findings and the strategy we applied might also inspire the field that commits to genome assembly in various species. Essentially, SV detection and genome assembly share many common challenges like the existence of large genomic repeats and the unresolved gaps that demand for contig scaffolding. We showed that Hi-C in combination with BioNano can thread large genomic regions with longer contiguity to improve the accuracy in decrypting any genome. The corrected gap sizes we generated will also contribute to the next version of human reference genome.

Further, our work presented a model of applying the advanced technology to improve the clinical diagnosis. The novel genes we identified that are frequently mutated in AML, which are associated to the disease survival, might provide new targets for disease treatment to the cancer field. We also showed the functional importance of SVs located in the non-coding regions. We noticed that a recent set of

publications from TCGA that investigated the SVs in 38 primary cancer types have a substantial part of work focusing on non-coding variants in driving cancers and driving chromatin reorganization, showing the filed's growing interest and comprehension in the role of inter-genic SVs in cancer.

In Chapter 4, we revealed the chromatin structural basis underlying the subtype-specific disruption of cis-regulation in AML. It suggests that similar chromatin reorganization might also take place in other types of cancers and associate with subtypes defined by different classification system. Moreover, we revealed a previously understudied mechanism, which is restoring chromatin structure and cis-regulation, of the well-known drug 5-AZA. Understanding this mechanism of 5-AZA in leukemia is crucial, as the patient responsiveness can be better predicted based on the gene expression or 3D genome profile from the patient. For example, according to our data, patient with *MYCN* and *WT1* activation, or *BCL11* and *GATA3* silencing are likely benefit from 5-AZA treatment. We believe this work can trigger more interesting findings that are related to oncogenic mechanism and treatment of a variety of cancer types.

Future questions and perspectives

Our development of Hi-C *breakfinder* turned out highly innovative and useful in identifying SVs, but its application could be hindered by the cost. The sequencing is affordable but still not ideally cheap, and right now this method needs around 100 million sequencing reads for approaching optimal performance. It will be useful to improve the algorithm in order to sensitize the SV detection with lower sequencing depth. One interesting idea is to apply *HiCPlus*, a machine-learning based Hi-C imputation method that only needs shallow sequencing data to predict the map generated with deep sequencing[237]. An alternative resolution is to incorporate region selection into the protocol so only the

SVs on genes of interest will be detected with very shallow sequencing depth. This can enable the SV screening for hundreds to thousands of targets each time across many samples in parallel, which could greatly increase the detection efficiency while lowering the cost.

In chapter 1, we also showcased reconstruction of a local map containing complex SVs with the integrative method, which is done manually. An algorithm that automatically recognize the existence of chained SVs in a haplotype-resolved manner, or even thread them with SNVs to realize genome-wide phasing, will be very useful to picture the cancer genome. This also applies to our detection of neo-TAD events induced SVs. Instead of manually reconstructing the rearranged Hi-C maps and visually looking for neo-TAD, the field needs an algorithm that can automatically construct all SV-related Hi-C maps and systematically and statistically detect neo-TAD and neo-loops. Indeed, we later resolved this task taking into account SV information like locus, orientation, complex SV and copy number, and we also streamlined TAD and loop detection on such genomes all at once.

In chapter 2, we carried out a pilot study for advancing SV detection in clinical diagnosis and we revealed many novel genes potentially related to leukemia. Due to the small scale of sample size, our result might provide some representative snapshot of the whole picture in leukemia SVs. Efforts based on large sample size is needed in the future to reveal the overall landscape of SVs in AML.

A few thoughts also rises as related to our findings in Chapter 4. For example, as we observed enhancers gained along the chromatin loops, it will be crucial to uncover targetable master TFs. Also, 5-AZA was known for remarkable responsiveness at clinical trials, but its application is hindered by less satisfactory improvement of survival due to high relapse when using the drug alone. In fact, 5-AZA and its deoxy derivative, decitabine, have been applied clinically in combination with other chemicals to improve the treatment effect in recent years, such as vorinostat, valproic acid, all-trans retinoic acid and sodium phenylbutyrate[206-208]. This might represent the epigenetic plasticity and redundancy in

chromatin spatial organization, suggesting the importance of recognition of other master regulator TFs and combined inhibition with 5-AZA. Further, the SV-induced neo-loops might be an important mechanism for disease onset, and how those events might inspire the disease intervention is question to be answered. Last, as we saw highly distinct chromatin structure, cis-regulation, and gene expression in different subtypes in AML, we felt that there are two directions for exploring epigenetic therapies, which could facilitate each other: One is to find and target the most recurrent alteration across all subtypes for ensuring high responsiveness, and the other is to identify the highly subtype-specific but essential altercations for sub-type specific treatment.

Reference

1. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
2. Futreal, P.A., et al., *A census of human cancer genes.* Nat Rev Cancer, 2004. **4**(3): p. 177-83.
3. 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.
4. 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.
5. 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.
6. 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.
7. Arber, D.A., et al., *The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia.* Blood, 2016. **127**(20): p. 2391-405.
8. Fielding, A.K., *How I treat Philadelphia chromosome-positive acute lymphoblastic leukemia.* Blood, 2010. **116**(18): p. 3409-17.
9. Sudmant, P.H., et al., *An integrated map of structural variation in 2,504 human genomes.* Nature, 2015. **526**(7571): p. 75-81.
10. Consortium, I.T.P.-C.A.o.W.G., *Pan-cancer analysis of whole genomes.* Nature, 2020. **578**(7793): p. 82-93.
11. Cortes-Ciriano, I., et al., *Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing.* Nat Genet, 2020. **52**(3): p. 331-341.
12. Zhang, Y., et al., *High-coverage whole-genome analysis of 1220 cancers reveals hundreds of genes deregulated by rearrangement-mediated cis-regulatory alterations.* Nat Commun, 2020. **11**(1): p. 736.
13. Rheinbay, E., et al., *Analyses of non-coding somatic drivers in 2,658 cancer whole genomes.* Nature, 2020. **578**(7793): p. 102-111.
14. Wan, T.S., *Cancer cytogenetics: methodology revisited.* Ann Lab Med, 2014. **34**(6): p. 413-25.
15. Zack, T.I., et al., *Pan-cancer patterns of somatic copy number alteration.* Nat Genet, 2013. **45**(10): p. 1134-40.
16. Mardis, E.R. and R.K. Wilson, *Cancer genome sequencing: a review.* Hum Mol Genet, 2009. **18**(R2): p. R163-8.
17. Inaki, K., et al., *Transcriptional consequences of genomic structural aberrations in breast cancer.* Genome Res, 2011. **21**(5): p. 676-87.
18. Maher, C.A., et al., *Transcriptome sequencing to detect gene fusions in cancer.* Nature, 2009. **458**(7234): p. 97-101.
19. Zhang, J., et al., *INTEGRATE: gene fusion discovery using whole genome and transcriptome data.* Genome Res, 2016. **26**(1): p. 108-18.
20. Campbell, P.J., et al., *Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing.* Nat Genet, 2008. **40**(6): p. 722-9.
21. Alkan, C., B.P. Coe, and E.E. Eichler, *Genome structural variation discovery and genotyping.* Nat Rev Genet, 2011. **12**(5): p. 363-76.
22. Peifer, M., et al., *Telomerase activation by genomic rearrangements in high-risk neuroblastoma.* Nature, 2015. **526**(7575): p. 700-4.
23. Nik-Zainal, S., et al., *Landscape of somatic mutations in 560 breast cancer whole-genome sequences.* Nature, 2016. **534**(7605): p. 47-54.

24. Van de Weyer, A.L., et al., *A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis thaliana.* Cell, 2019. **178**(5): p. 1260-1272 e14.
25. Hsieh, P., et al., *Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes.* Science, 2019. **366**(6463).
26. Gilpatrick, T., et al., *Targeted nanopore sequencing with Cas9-guided adapter ligation.* Nat Biotechnol, 2020. **38**(4): p. 433-438.
27. Gong, L., et al., *Picky comprehensively detects high-resolution structural variants in nanopore long reads.* Nat Methods, 2018. **15**(6): p. 455-460.
28. Jing Lu, L.d.P., Zhe Liu1,2,10, Verity Hill4,10,Min Kang2, Huifang Lin1,2, Jiufeng, et al., *Genomic epidemiology of SARS-CoV-2 in Guangdong Province, China.* Cell, 2020.
29. Fauver, J.R., et al., *Coast-to-coast spread of SARS-CoV-2 during the early epidemic in the United States.* Cell, 2020.
30. Tyson, J.R., et al., *MinION-based long-read sequencing and assembly extends the Caenorhabditis elegans reference genome.* Genome Res, 2018. **28**(2): p. 266-274.
31. Rhoads, A. and K.F. Au, *PacBio Sequencing and Its Applications.* Genomics Proteomics Bioinformatics, 2015. **13**(5): p. 278-89.
32. Dixon, J.R., et al., *Integrative detection and analysis of structural variation in cancer genomes.* Nat Genet, 2018. **50**(10): p. 1388-1398.
33. Jiao, Y., et al., *Improved maize reference genome with single-molecule technologies.* Nature, 2017. **546**(7659): p. 524-527.
34. Kawakatsu, T., et al., *Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions.* Cell, 2016. **166**(2): p. 492-505.
35. Bevan, M.W., et al., *Genomic innovation for crop improvement.* Nature, 2017. **543**(7645): p. 346-354.
36. Mascher, M., et al., *A chromosome conformation capture ordered sequence of the barley genome.* Nature, 2017. **544**(7651): p. 427-433.
37. Wu, S., et al., *Circular ecDNA promotes accessible chromatin and high oncogene expression.* Nature, 2019. **575**(7784): p. 699-703.
38. Bickhart, D.M., et al., *Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome.* Nat Genet, 2017. **49**(4): p. 643-650.
39. Kronenberg, Z.N., et al., *High-resolution comparative analysis of great ape genomes.* Science, 2018. **360**(6393).
40. Seo, J.S., et al., *De novo assembly and phasing of a Korean human genome.* Nature, 2016. **538**(7624): p. 243-247.
41. Alberts, B., *Molecular Biology of the Cell*.
42. Mather, K., *Crossing over and Heterochromatin in the X Chromosome of Drosophila Melanogaster.* Genetics, 1939. **24**(3): p. 413-35.
43. Strom, A.R., et al., *Phase separation drives heterochromatin domain formation.* Nature, 2017. **547**(7662): p. 241-245.
44. Meaburn, K.J. and T. Misteli, *Cell biology: chromosome territories.* Nature, 2007. **445**(7126): p. 379-781.
45. van Steensel, B. and A.S. Belmont, *Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression.* Cell, 2017. **169**(5): p. 780-791.
46. Edelman, L.B. and P. Fraser, *Transcription factories: genetic programming in three dimensions.* Curr Opin Genet Dev, 2012. **22**(2): p. 110-4.
47. Pope, B.D., et al., *Topologically associating domains are stable units of replication-timing regulation.* Nature, 2014. **515**(7527): p. 402-5.
48. Ryba, T., et al., *Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types.* Genome Res, 2010. **20**(6): p. 761-70.
49. Lieberman-Aiden, E., et al., *Comprehensive mapping of long-range interactions reveals folding principles of the human genome.* Science, 2009. **326**(5950): p. 289-93.

50. Jin, F., et al., *A high-resolution map of the three-dimensional chromatin interactome in human cells.* Nature, 2013. **503**(7475): p. 290-4.
51. 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.
52. Ong, C.T. and V.G. Corces, *CTCF: an architectural protein bridging genome topology and function.* Nat Rev Genet, 2014. **15**(4): p. 234-46.
53. Rubio, E.D., et al., *CTCF physically links cohesin to chromatin.* Proc Natl Acad Sci U S A, 2008. **105**(24): p. 8309-14.
54. Fudenberg, G., et al., *Formation of Chromosomal Domains by Loop Extrusion.* Cell Rep, 2016. **15**(9): p. 2038-49.
55. Rao, S.S.P., et al., *Cohesin Loss Eliminates All Loop Domains.* Cell, 2017. **171**(2): p. 305-320.e24.
56. Kubo, N., et al., *Preservation of Chromatin Organization after Acute Loss of CTCF in Mouse Embryonic Stem Cells.* bioRxiv, 2017: p. 118737.
57. Yan, J., et al., *Histone H3 lysine 4 monomethylation modulates long-range chromatin interactions at enhancers.* Cell Res, 2018. **28**(2): p. 204-220.
58. Nichols, M.H. and V.G. Corces, *A CTCF Code for 3D Genome Architecture.* Cell, 2015. **162**(4): p. 703-5.
59. Davidson, I.F., et al., *DNA loop extrusion by human cohesin.* Science, 2019. **366**(6471): p. 1338-1345.
60. Vian, L., et al., *The Energetics and Physiological Impact of Cohesin Extrusion.* Cell, 2018. **175**(1): p. 292-294.
61. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions.* Nature, 2012. **485**(7398): p. 376-80.
62. Kagey, M.H., et al., *Mediator and cohesin connect gene expression and chromatin architecture.* Nature, 2010. **467**(7314): p. 430-5.
63. Phillips-Cremins, J.E., et al., *Architectural protein subclasses shape 3D organization of genomes during lineage commitment.* Cell, 2013. **153**(6): p. 1281-95.
64. Nagano, T., et al., *Single-cell Hi-C reveals cell-to-cell variability in chromosome structure.* Nature, 2013. **502**(7469): p. 59-64.
65. Beagan, J.A., et al., *YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment.* Genome Res, 2017. **27**(7): p. 1139-1152.
66. Haarhuis, J.H.I., et al., *The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension.* Cell, 2017. **169**(4): p. 693-707.e14.
67. Dixon, J.R., D.U. Gorkin, and B. Ren, *Chromatin Domains: The Unit of Chromosome Organization.* Mol Cell, 2016. **62**(5): p. 668-80.
68. Richmond, T.J. and C.A. Davey, *The structure of DNA in the nucleosome core.* Nature, 2003. **423**(6936): p. 145-50.
69. Robinson, P.J., et al., *EM measurements define the dimensions of the "30-nm" chromatin fiber: evidence for a compact, interdigitated structure.* Proc Natl Acad Sci U S A, 2006. **103**(17): p. 6506-11.
70. Schalch, T., et al., *X-ray structure of a tetranucleosome and its implications for the chromatin fibre.* Nature, 2005. **436**(7047): p. 138-41.
71. Song, F., et al., *Cryo-EM study of the chromatin fiber reveals a double helix twisted by tetranucleosomal units.* Science, 2014. **344**(6182): p. 376-80.
72. Ou, H.D., et al., *ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells.* Science, 2017. **357**(6349).
73. Langer-Safer, P.R., M. Levine, and D.C. Ward, *Immunological method for mapping genes on Drosophila polytene chromosomes.* Proc Natl Acad Sci U S A, 1982. **79**(14): p. 4381-5.
74. Westphal, V., et al., *Video-rate far-field optical nanoscopy dissects synaptic vesicle movement.* Science, 2008. **320**(5873): p. 246-9.
75. Betzig, E., et al., *Imaging intracellular fluorescent proteins at nanometer resolution.* Science, 2006. **313**(5793): p. 1642-5.

76. Hess, S.T., T.P. Girirajan, and M.D. Mason, *Ultra-high resolution imaging by fluorescence photoactivation localization microscopy.* Biophys J, 2006. **91**(11): p. 4258-72.
77. Ma, H., et al., *Multiplexed labeling of genomic loci with dCas9 and engineered sgRNAs using CRISPRainbow.* Nat Biotechnol, 2016. **34**(5): p. 528-30.
78. Shachar, S., et al., *Identification of Gene Positioning Factors Using High-Throughput Imaging Mapping.* Cell, 2015. **162**(4): p. 911-23.
79. Dekker, J., et al., *Capturing chromosome conformation.* Science, 2002. **295**(5558): p. 1306-11.
80. Simonis, M., et al., *Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C).* Nat Genet, 2006. **38**(11): p. 1348-54.
81. Dostie, J., et al., *Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements.* Genome Res, 2006. **16**(10): p. 1299-309.
82. Sridhar, B., et al., *Systematic Mapping of RNA-Chromatin Interactions In Vivo.* Curr Biol, 2017. **27**(4): p. 610-612.
83. Li, X., et al., *GRID-seq reveals the global RNA-chromatin interactome.* Nat Biotechnol, 2017. **35**(10): p. 940-950.
84. Du, Z., et al., *Allelic reprogramming of 3D chromatin architecture during early mammalian development.* Nature, 2017. **547**(7662): p. 232-235.
85. Dixon, J.R., et al., *Chromatin architecture reorganization during stem cell differentiation.* Nature, 2015. **518**(7539): p. 331-6.
86. Schmitt, A.D., et al., *A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome.* Cell Rep, 2016. **17**(8): p. 2042-2059.
87. Lin, Y.C., et al., *Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate.* Nat Immunol, 2012. **13**(12): p. 1196-204.
88. Siersbaek, R., et al., *Dynamic Rewiring of Promoter-Anchored Chromatin Loops during Adipocyte Differentiation.* Mol Cell, 2017. **66**(3): p. 420-435 e5.
89. Doynova, M.D., et al., *Linkages between changes in the 3D organization of the genome and transcription during myotube differentiation in vitro.* Skelet Muscle, 2017. **7**(1): p. 5.
90. Valton, A.L. and J. Dekker, *TAD disruption as oncogenic driver.* Curr Opin Genet Dev, 2016. **36**: p. 34-40.
91. Northcott, P.A., et al., *Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma.* Nature, 2014. **511**(7510): p. 428-34.
92. Taberlay, P.C., et al., *Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations.* Genome Res, 2016. **26**(6): p. 719-31.
93. Liu, E.M., et al., *Identification of Cancer Drivers at CTCF Insulators in 1,962 Whole Genomes.* Cell Syst, 2019. **8**(5): p. 446-455 e8.
94. Kloetgen, A., et al., *Three-dimensional chromatin landscapes in T cell acute lymphoblastic leukemia.* Nat Genet, 2020. **52**(4): p. 388-400.
95. Rosa-Garrido, M., et al., *High-Resolution Mapping of Chromatin Conformation in Cardiac Myocytes Reveals Structural Remodeling of the Epigenome in Heart Failure.* Circulation, 2017. **136**(17): p. 1613-1625.
96. Javierre, B.M., et al., *Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters.* Cell, 2016. **167**(5): p. 1369-1384 e19.
97. Flavahan, W.A., et al., *Insulator dysfunction and oncogene activation in IDH mutant gliomas.* Nature, 2016. **529**(7584): p. 110-4.
98. Flavahan, W.A., et al., *Altered chromosomal topology drives oncogenic programs in SDH-deficient GISTs.* Nature, 2019. **575**(7781): p. 229-233.
99. Lupianez, D.G., et al., *Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions.* Cell, 2015. **161**(5): p. 1012-1025.
100. Hnisz, D., et al., *Activation of proto-oncogenes by disruption of chromosome neighborhoods.* Science, 2016. **351**(6280): p. 1454-1458.

101. Groschel, S., et al., *A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia.* Cell, 2014. **157**(2): p. 369-381.
102. Mostovoy, Y., et al., *A hybrid approach for de novo human genome sequence assembly and phasing.* Nat Methods, 2016. **13**(7): p. 587-90.
103. Pendleton, M., et al., *Assembly and diploid architecture of an individual human genome via single-molecule technologies.* Nat Methods, 2015. **12**(8): p. 780-6.
104. Burton, J.N., et al., *Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions.* Nat Biotechnol, 2013. **31**(12): p. 1119-25.
105. Kaplan, N. and J. Dekker, *High-throughput genome scaffolding from in vivo DNA interaction frequency.* Nat Biotechnol, 2013. **31**(12): p. 1143-7.
106. Mak, A.C., et al., *Genome-Wide Structural Variation Detection by Genome Mapping on Nanochannel Arrays.* Genetics, 2016. **202**(1): p. 351-62.
107. Jiao, W.B., et al., *Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data.* Genome Res, 2017.
108. Bickhart, D.M., et al., *Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome.* Nat Genet, 2017.
109. Jarvis, D.E., et al., *The genome of Chenopodium quinoa.* Nature, 2017. **542**(7641): p. 307-312.
110. Lam, E.T., et al., *Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly.* Nat Biotechnol, 2012. **30**(8): p. 771-6.
111. Selvaraj, S., et al., *Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing.* Nat Biotechnol, 2013. **31**(12): p. 1111-8.
112. Engreitz, J.M., V. Agarwala, and L.A. Mirny, *Three-dimensional genome architecture influences partner selection for chromosomal translocations in human disease.* PLoS One, 2012. **7**(9): p. e44196.
113. Naumova, N., et al., *Organization of the mitotic chromosome.* Science, 2013. **342**(6161): p. 948-53.
114. Xu, H., et al., *Integrative Analysis Reveals the Transcriptional Collaboration between EZH2 and E2F1 in the Regulation of Cancer-Related Gene Expression.* Mol Cancer Res, 2016. **14**(2): p. 163-72.
115. Layer, R.M., et al., *LUMPY: a probabilistic framework for structural variant discovery.* Genome Biol, 2014. **15**(6): p. R84.
116. Rausch, T., et al., *DELLY: structural variant discovery by integrated paired-end and split-read analysis.* Bioinformatics, 2012. **28**(18): p. i333-i339.
117. Boeva, V., et al., *Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data.* Bioinformatics, 2012. **28**(3): p. 423-5.
118. Wang, Z., et al., *The properties of genome conformation and spatial gene interaction and regulation networks of normal and malignant human cell types.* PLoS One, 2013. **8**(3): p. e58793.
119. Barutcu, A.R., et al., *Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells.* Genome Biol, 2015. **16**: p. 214.
120. Barutcu, A.R., et al., *RUNX1 contributes to higher-order chromatin organization and gene regulation in breast cancer cells.* Biochim Biophys Acta, 2016. **1859**(11): p. 1389-1397.
121. Guo, Y., et al., *CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function.* Cell, 2015. **162**(4): p. 900-10.
122. Krzywinski, M., et al., *Circos: an information aesthetic for comparative genomics.* Genome Res, 2009. **19**(9): p. 1639-45.
123. Seaman, L., et al., *Nucleome Analysis Reveals Structure-Function Relationships for Colon Cancer.* Mol Cancer Res, 2017. **15**(7): p. 821-830.
124. Harewood, L., et al., *Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours.* Genome Biol, 2017. **18**(1): p. 125.
125. Wu, H.J. and F. Michor, *A computational strategy to adjust for copy number in tumor Hi-C data.* Bioinformatics, 2016. **32**(24): p. 3695-3701.
126. Chakraborty, A. and F. Ay, *Identification of copy number variations and translocations in cancer cells from Hi-C data.* Bioinformatics, 2017.

127. Naumann, S., et al., *Complete karyotype characterization of the K562 cell line by combined application of G-banding, multiplex-fluorescence in situ hybridization, fluorescence in situ hybridization, and comparative genomic hybridization.* Leuk Res, 2001. **25**(4): p. 313-22.
128. O'Doherty, A., et al., *An aneuploid mouse strain carrying human chromosome 21 with Down syndrome phenotypes.* Science, 2005. **309**(5743): p. 2033-7.
129. Gribble, S.M., et al., *Massively parallel sequencing reveals the complex structure of an irradiated human chromosome on a mouse background in the Tc1 model of Down syndrome.* PLoS One, 2013. **8**(4): p. e60482.
130. Nora, E.P., et al., *Spatial partitioning of the regulatory landscape of the X-inactivation centre.* Nature, 2012. **485**(7398): p. 381-5.
131. Rhind, N. and D.M. Gilbert, *DNA replication timing.* Cold Spring Harb Perspect Biol, 2013. **5**(8): p. a010132.
132. Dileep, V., et al., *Large-Scale Chromatin Structure-Function Relationships during the Cell Cycle and Development: Insights from Replication Timing.* Cold Spring Harb Symp Quant Biol, 2015. **80**: p. 53-63.
133. Pope, B.D., et al., *Replication-timing boundaries facilitate cell-type and species-specific regulation of a rearranged human chromosome in mouse.* Hum Mol Genet, 2012. **21**(19): p. 4162-70.
134. Ryba, T., et al., *Abnormal developmental control of replication-timing domains in pediatric acute lymphoblastic leukemia.* Genome Res, 2012. **22**(10): p. 1833-44.
135. Dileep, V., et al., *Topologically associating domains and their long-range contacts are established during early G1 coincident with the establishment of the replication-timing program.* Genome Res, 2015. **25**(8): p. 1104-13.
136. Rivera-Mulia, J.C., et al., *Dynamic changes in replication timing and gene expression during lineage specification of human pluripotent stem cells.* Genome Res, 2015. **25**(8): p. 1091-103.
137. Sima, J. and D.M. Gilbert, *Complex correlations: replication timing and mutational landscapes during cancer and genome evolution.* Curr Opin Genet Dev, 2014. **25**: p. 93-100.
138. Chiarle, R., et al., *Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells.* Cell, 2011. **147**(1): p. 107-19.
139. Struski, S., et al., *Identification of chromosomal loci associated with non-P-glycoprotein-mediated multidrug resistance to topoisomerase II inhibitor in lung adenocarcinoma cell line by comparative genomic hybridization.* Genes Chromosomes Cancer, 2001. **30**(2): p. 136-42.
140. Strefford, J.C., et al., *A combination of molecular cytogenetic analyses reveals complex genetic alterations in conventional renal cell carcinoma.* Cancer Genet Cytogenet, 2005. **159**(1): p. 1-9.
141. Peng, K.J., et al., *Characterization of two human lung adenocarcinoma cell lines by reciprocal chromosome painting.* Dongwuxue Yanjiu, 2010. **31**(2): p. 113-21.
142. Beheshti, B., et al., *Identification of a high frequency of chromosomal rearrangements in the centromeric regions of prostate cancer cell lines by sequential giemsa banding and spectral karyotyping.* Mol Diagn, 2000. **5**(1): p. 23-32.
143. Liu, J., et al., *Modeling of lung cancer by an orthotopically growing H460SM variant cell line reveals novel candidate genes for systemic metastasis.* Oncogene, 2004. **23**(37): p. 6316-24.
144. Espino, P.S., et al., *Genomic instability and histone H3 phosphorylation induction by the Ras-mitogen activated protein kinase pathway in pancreatic cancer cells.* Int J Cancer, 2009. **124**(3): p. 562-7.
145. Sirivatanauksorn, V., et al., *Non-random chromosomal rearrangements in pancreatic cancer cell lines identified by spectral karyotyping.* Int J Cancer, 2001. **91**(3): p. 350-8.
146. Rondon-Lagos, M., et al., *Differences and homologies of chromosomal alterations within and between breast cancer cell lines: a clustering analysis.* Mol Cytogenet, 2014. **7**(1): p. 8.
147. Hillmer, A.M., et al., *Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes.* Genome Res, 2011. **21**(5): p. 665-75.
148. Hampton, O.A., et al., *Long-range massively parallel mate pair sequencing detects distinct mutations and similar patterns of structural mutability in two breast cancer cell lines.* Cancer Genet, 2011. **204**(8): p. 447-57.

149. Yang, L., et al., *Diverse mechanisms of somatic structural variations in human cancer genomes.* Cell, 2013. **153**(4): p. 919-29.
150. Forbes, S.A., et al., *COSMIC: exploring the world's knowledge of somatic mutations in human cancer.* Nucleic Acids Res, 2015. **43**(Database issue): p. D805-11.
151. Mifsud, B., et al., *Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C.* Nat Genet, 2015. **47**(6): p. 598-606.
152. Franke, M., et al., *Formation of new chromatin domains determines pathogenicity of genomic duplications.* Nature, 2016. **538**(7624): p. 265-269.
153. 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.
154. Huang, R., et al., *MYCN and MYC regulate tumor proliferation and tumorigenesis directly through BMI1 in human neuroblastomas.* FASEB J, 2011. **25**(12): p. 4138-49.
155. Barretina, J., et al., *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature, 2012. **483**(7391): p. 603-7.
156. Marchal, C., et al., *Repli-seq: genome-wide analysis of replication timing by next-generation sequencing.* bioRxiv, 2017.
157. Kim, D. and S.L. Salzberg, *TopHat-Fusion: an algorithm for discovery of novel fusion transcripts.* Genome Biol, 2011. **12**(8): p. R72.
158. Haas, B., et al., *STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq.* bioRxiv, 2017.
159. Benelli, M., et al., *Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript.* Bioinformatics, 2012. **28**(24): p. 3232-9.
160. Klijn, C., et al., *A comprehensive transcriptional portrait of human cancer cell lines.* Nat Biotechnol, 2015. **33**(3): p. 306-12.
161. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nat Protoc, 2012. **7**(3): p. 562-78.
162. Berger, M.F. and E.R. Mardis, *The emerging clinical relevance of genomics in cancer medicine.* Nat Rev Clin Oncol, 2018. **15**(6): p. 353-365.
163. Vogelstein, B., et al., *Cancer genome landscapes.* Science, 2013. **339**(6127): p. 1546-58.
164. Metzeler, K.H., et al., *Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia.* Blood, 2016. **128**(5): p. 686-98.
165. Papaemmanuil, E., et al., *Genomic Classification and Prognosis in Acute Myeloid Leukemia.* N Engl J Med, 2016. **374**(23): p. 2209-2221.
166. Dixon, J.R., et al., *Integrative detection and analysis of structural variation in cancer genomes.* Nat Genet, 2018.
167. MacDonald, J.R., et al., *The Database of Genomic Variants: a curated collection of structural variation in the human genome.* Nucleic Acids Res, 2014. **42**(Database issue): p. D986-92.
168. Levy-Sakin, M., et al., *Genome maps across 26 human populations reveal population-specific patterns of structural variation.* Nature Communications, 2019. **in press**.
169. Chiaretti, S., et al., *CRLF2 overexpression identifies an unfavourable subgroup of adult B-cell precursor acute lymphoblastic leukemia lacking recurrent genetic abnormalities.* Leuk Res, 2016. **41**: p. 36-42.
170. Harvey, R.C., et al., *Rearrangement of CRLF2 is associated with mutation of JAK kinases, alteration of IKZF1, Hispanic/Latino ethnicity, and a poor outcome in pediatric B-progenitor acute lymphoblastic leukemia.* Blood, 2010. **115**(26): p. 5312-21.
171. Russell, L.J., et al., *Deregulated expression of cytokine receptor gene, CRLF2, is involved in lymphoid transformation in B-cell precursor acute lymphoblastic leukemia.* Blood, 2009. **114**(13): p. 2688-98.
172. Dong, X., et al., *RSPO2 suppresses colorectal cancer metastasis by counteracting the Wnt5a/Fzd7-driven noncanonical Wnt pathway.* Cancer Lett, 2017. **402**: p. 153-165.
173. Wilhelm, F., et al., *Novel Insights into Gastric Cancer: Methylation of R-spondins and Regulation of LGR5 by SP1.* Mol Cancer Res, 2017. **15**(6): p. 776-785.

174. Yoon, J.K. and J.S. Lee, *Cellular signaling and biological functions of R-spondins.* Cell Signal, 2012. **24**(2): p. 369-77.
175. Cleveland, D.W., *NuMA: a protein involved in nuclear structure, spindle assembly, and nuclear re-formation.* Trends Cell Biol, 1995. **5**(2): p. 60-4.
176. Wells, R.A., C. Catzavelos, and S. Kamel-Reid, *Fusion of retinoic acid receptor alpha to NuMA, the nuclear mitotic apparatus protein, by a variant translocation in acute promyelocytic leukaemia.* Nat Genet, 1997. **17**(1): p. 109-13.
177. Cho, Y., et al., *AFAP1 Is a Novel Downstream Mediator of TGF-beta1 for CCN2 Induction in Osteoblasts.* PLoS One, 2015. **10**(9): p. e0136712.
178. Gatesman, A., et al., *Protein kinase Calpha activates c-Src and induces podosome formation via AFAP-110.* Mol Cell Biol, 2004. **24**(17): p. 7578-97.
179. Qian, Y., et al., *Analysis of the role of the leucine zipper motif in regulating the ability of AFAP-110 to alter actin filament integrity*. J Cell Biochem, 2004. **91**(3): p. 602-20.
180. Liu, S., et al., *ATX-LPA receptor axis in inflammation and cancer.* Cell Cycle, 2009. **8**(22): p. 3695-701.
181. Onallah, H., et al., *Activity and clinical relevance of autotaxin and lysophosphatidic acid pathways in high-grade serous carcinoma.* Virchows Arch, 2018. **473**(4): p. 463-470.
182. Lu, J.R., et al., *FOG-2, a heart- and brain-enriched cofactor for GATA transcription factors.* Mol Cell Biol, 1999. **19**(6): p. 4495-502.
183. Porcher, C., H. Chagraoui, and M.S. Kristiansen, *SCL/TAL1: a multifaceted regulator from blood development to disease.* Blood, 2017. **129**(15): p. 2051-2060.
184. Chan, E.K.F., et al., *Optical mapping reveals a higher level of genomic architecture of chained fusions in cancer.* Genome Res, 2018. **28**(5): p. 726-738.
185. Jaratlerdsiri, W., et al., *Next generation mapping reveals novel large genomic rearrangements in prostate cancer.* Oncotarget, 2017. **8**(14): p. 23588-23602.
186. Blank, U. and S. Karlsson, *TGF-beta signaling in the control of hematopoietic stem cells.* Blood, 2015. **125**(23): p. 3542-50.
187. Bachegowda, L., et al., *Signal transduction inhibitors in treatment of myelodysplastic syndromes.* J Hematol Oncol, 2013. **6**: p. 50.
188. Zhou, L., et al., *Inhibition of the TGF-beta receptor I kinase promotes hematopoiesis in MDS.* Blood, 2008. **112**(8): p. 3434-43.
189. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
190. Tarasov, A., et al., *Sambamba: fast processing of NGS alignment formats.* Bioinformatics, 2015. **31**(12): p. 2032-4.
191. Chiang, C., et al., *SpeedSeq: ultra-fast personal genome analysis and interpretation.* Nat Methods, 2015. **12**(10): p. 966-8.
192. Faust, G.G. and I.M. Hall, *SAMBLASTER: fast duplicate marking and structural variant read extraction.* Bioinformatics, 2014. **30**(17): p. 2503-5.
193. Cingolani, P., et al., *Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift.* Front Genet, 2012. **3**: p. 35.
194. Krzywinski, M.I., et al., *Circos: An information aesthetic for comparative genomics.* Genome Research, 2009.
195. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
196. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC Bioinformatics, 2011. **12**: p. 323.
197. van de Geijn, B., et al., *WASP: allele-specific software for robust molecular quantitative trait locus discovery.* Nat Methods, 2015. **12**(11): p. 1061-3.
198. Dohner, H., et al., *Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel.* Blood, 2017. **129**(4): p. 424-447.

199. Dohner, H., D.J. Weisdorf, and C.D. Bloomfield, *Acute Myeloid Leukemia.* N Engl J Med, 2015. **373**(12): p. 1136-52.
200. Cancer Genome Atlas Research, N., et al., *Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia.* N Engl J Med, 2013. **368**(22): p. 2059-74.
201. Assi, S.A., et al., *Subtype-specific regulatory network rewiring in acute myeloid leukemia.* Nat Genet, 2019. **51**(1): p. 151-162.
202. Despang, A., et al., *Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture.* Nat Genet, 2019. **51**(8): p. 1263-1271.
203. Kraft, K., et al., *Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations.* Nat Cell Biol, 2019. **21**(3): p. 305-310.
204. Viny, A.D., et al., *Cohesin Members Stag1 and Stag2 Display Distinct Roles in Chromatin Accessibility and Topological Control of HSC Self-Renewal and Differentiation.* Cell Stem Cell, 2019. **25**(5): p. 682-696 e8.
205. Hashimshony, T., et al., *The role of DNA methylation in setting up chromatin structure during development.* Nat Genet, 2003. **34**(2): p. 187-92.
206. Burke, M.J., et al., *Decitabine and Vorinostat with Chemotherapy in Relapsed Pediatric Acute Lymphoblastic Leukemia: A TACL Pilot Study.* Clin Cancer Res, 2020.
207. Soriano, A.O., et al., *Safety and clinical activity of the combination of 5-azacytidine, valproic acid, and all-trans retinoic acid in acute myeloid leukemia and myelodysplastic syndrome.* Blood, 2007. **110**(7): p. 2302-8.
208. Maslak, P., et al., *Pilot study of combination transcriptional modulation therapy with sodium phenylbutyrate and 5-azacytidine in patients with acute myeloid leukemia or myelodysplastic syndrome.* Leukemia, 2006. **20**(2): p. 212-7.
209. Van Vlierberghe, P. and A. Ferrando, *The molecular basis of T cell acute lymphoblastic leukemia.* J Clin Invest, 2012. **122**(10): p. 3398-406.
210. Spassov, B.V., et al., *Wilms' tumor protein and FLT3-internal tandem duplication expression in patients with de novo acute myeloid leukemia.* Hematology, 2011. **16**(1): p. 37-42.
211. Akdemir, K.C., et al., *Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer.* Nat Genet, 2020. **52**(3): p. 294-305.
212. Bahr, C., et al., *A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies.* Nature, 2018. **553**(7689): p. 515-520.
213. Salameh, T.J., et al., *A supervised learning framework for chromatin loop detection in genome-wide contact maps.* bioRxiv, 2019: p. 739698.
214. Cai, Y., et al., *H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions.* bioRxiv, 2019: p. 684712.
215. Vian, L., et al., *The Energetics and Physiological Impact of Cohesin Extrusion.* Cell, 2018. **173**(5): p. 1165-1178 e20.
216. Phanstiel, D.H., et al., *Static and Dynamic DNA Loops form AP-1-Bound Activation Hubs during Macrophage Development.* Mol Cell, 2017. **67**(6): p. 1037-1048 e6.
217. Prange, K.H.M., et al., *MLL-AF9 and MLL-AF4 oncofusion proteins bind a distinct enhancer repertoire and target the RUNX1 program in 11q23 acute myeloid leukemia.* Oncogene, 2017. **36**(23): p. 3346-3356.
218. Bell, A.C. and G. Felsenfeld, *Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene.* Nature, 2000. **405**(6785): p. 482-5.
219. Figueroa, M.E., et al., *DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia.* Cancer Cell, 2010. **17**(1): p. 13-27.
220. Akalin, A., et al., *Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia.* PLoS Genet, 2012. **8**(6): p. e1002781.
221. Neri, F., et al., *Intragenic DNA methylation prevents spurious transcription initiation.* Nature, 2017. **543**(7643): p. 72-77.

222. McKeown, M.R., et al., *Superenhancer Analysis Defines Novel Epigenomic Subtypes of Non-APL AML, Including an RARalpha Dependency Targetable by SY-1425, a Potent and Selective RARalpha Agonist.* Cancer Discov, 2017. **7**(10): p. 1136-1153.
223. Yang, A.S., et al., *DNA Methylation Changes after 5-Aza-2′-Deoxycytidine Therapy in Patients with Leukemia.* Cancer Research, 2006. **66**(10): p. 5495-5503.
224. Patel, J.P., et al., *Prognostic relevance of integrated genetic profiling in acute myeloid leukemia.* N Engl J Med, 2012. **366**(12): p. 1079-89.
225. Luo, Z., et al., *A Prostate Cancer Risk Element Functions as a Repressive Loop that Regulates HOXA13.* Cell Rep, 2017. **21**(6): p. 1411-1417.
226. Kaya-Okur, H.S., et al., *CUT&Tag for efficient epigenomic profiling of small samples and single cells.* Nat Commun, 2019. **10**(1): p. 1930.
227. 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.
228. Cingolani, P., et al., *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.* Fly (Austin), 2012. **6**(2): p. 80-92.
229. Abdennur, N. and L.A. Mirny, *Cooler: scalable storage for Hi-C data and other genomically labeled arrays.* Bioinformatics, 2020. **36**(1): p. 311-316.
230. Kerpedjiev, P., et al., *HiGlass: web-based visual exploration and analysis of genome interaction maps.* Genome Biol, 2018. **19**(1): p. 125.
231. Durand, N.C., et al., *Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom.* Cell Syst, 2016. **3**(1): p. 99-101.
232. Wang, Y., et al., *The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions.* Genome Biol, 2018. **19**(1): p. 151.
233. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.
234. Krueger, F. and S.R. Andrews, *Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.* Bioinformatics, 2011. **27**(11): p. 1571-2.
235. Schultz, M.D., et al., *Human body epigenome maps reveal noncanonical DNA methylation variation.* Nature, 2015. **523**(7559): p. 212-6.
236. Corces, M.R., et al., *The chromatin accessibility landscape of primary human cancers.* Science, 2018. **362**(6413).
237. Zhang, Y., et al., *Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus.* Nat Commun, 2018. **9**(1): p. 750.

VITA: Jie Xu

EDUCATION

Ph.D program in Biomedical Sciences, **Penn State University,** PA, USA	Aug.2014-May.2020
Pre-doctoral visiting scholar, **Northwestern University**, IL, USA	Jul.2019-May.2020
Undergrad Exchange Program, **Stanford University**, CA, USA	Jan.2014~April.2014
B.S. in Basic Medical Sciences, **Fudan University**, Shanghai, China	Sep.2009~Jun.2014

SELECED AWARDS AND HONORS

The Charles W. Hill Graduate Student Award	Aug. 2018
Outstanding Poster Award and Travel Award, Penn State Biochem Dept.	Aug. 2017
Travel award: Penn State College of Medicine	Sept. 2017
Full scholarship: VARI 2017 advanced training program	May. 2017
Outstanding thesis award, School of Medical Sciences, Fudan Uni.	Jun.2014
Full scholarship: Stanford Exchange Program	Jan.2014
Designated: *Outstanding Leader* by Fudan University	2011,2012
Selected as one of three *Medical Students of the Year (2010)*	Jun.2011
College Entrance Examination-Exempted Admission to Fudan Uni.	Jan.2009
First Prize in Chinese National Biology Olypiad	May.2008

SELETED ABSTRACT PRESENTATIONS

J. Xu, J. Dixon, F. Song & F. Yue. *Structural variation and its impact on 3D genome structure in cancer cells.* ●2018 American Society of Human Genetics (ASHG), San Diego, USA, Oct.2018 (Platform Oral Presentation). ●ASHG, Vancouver, Canada, Oct.2016 (Poster). ●4DN Annual meeting, Bethesda, USA, Sep. 2017 (Poster).

J. Xu, F. Song, J. R. Broach & F. Yue. *Detection of structure variations in cancer cell lines and leukemia patient samples.* ●2018 Workshop on emerging methods for sequence analysis. Penn State, State College, USA. June. 2018. (Oral presentation). ●Penn State College of Medicine Dept. of Biochemistry Annual Retreat. Hershey, USA, Aug.2018. (Oral Presentation)

SELECTED PUBLICATION AND PREPRINTS

J. Xu, F. Song, B. Zhang, X. Wang, B. Jia, A. Kazmer, N. Birch, A. Shilatifard, R. Levine, J. R. Broach, H. Zheng, F. Yue. Subtype-specific and structure variation-induced chromatin spatial reorganization in acute myeloid leukemia. (*Manuscript ready for submission*)

J. Xu, F. Song, E. Schleicher, C. Pool, D. Bann, …, B. Miller, D. Claxton, G. Moldovan, F. Yue, J. R. Broach. Whole genome optical mapping reveals previously unrecognizable structural variants in leukemia patients' samples. (*Revision in **Cancer Research***)

J. Dixon*, **J. Xu*,** V. Dileep*, Y. Zhan*, F. Song*, …, C. Ernst, S. Hadjur, D. T. Odom, J. A. Stamatoyannopoulos, J. R. Broach, R. Hardison, F. Ay, W. S. Noble, J. Dekker, D. M. Gilbert and F. Yue, An integrative detection and analysis of structural variation in cancer genomes, **Nature Genetics.** 50 (2018), no. 10, 1388-1398. (*****co-first author***)

J. Zhang*, D. Lee*, V. Dhiman*, Peng Jiang*, **J. Xu***, P. McGillivray*, H. Yang*, …, F. Yue, X. S. Liu, K. White, M. Gerstein. An integrative ENCODE resource for cancer genomics. **Nature Communication.** *In press* (****co-first author***)

Y. Zhang, L. An, **J. Xu,** B. Zhang, W. J. Zheng, M. Hu, J. Tang and F. Yue, Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus, **Nature Communication.** 9 (2018), no. 1, 750.

F. Xian, **J. Xu**, L. Nakhleh. Detecting large indels using optical map data. Comparative Genomics. RECOMB-CG 2018. **Lecture Notes in Computer Science**, vol 11183.

Y. Wang, F. Song, B. Zhang, L. Zhang, L. An, **J. Xu**, …, F. Yue, The 3D Genome Browser: A web-based browser for visualizing 3D genome organization and long-range chromatin interactions, **Genome Biology**. 19 (2018), no. 1, 151.

Z. Yan, N. Huang, W. Wu, W. Chen, Y. Jiang, J. Chen, X. Huang, X. Wen, **J. Xu**, Q. Jin, K. Zhang, Z. Chen, S. Chien and S Zhong. Genome-wide co-localization of RNA-DNA interactions and fusion RNA pairs. **PNAS.** 2019 116 (8) 3328-3337.

ENCODE Consortium. Expanded Encyclopedias of DNA Elements in the Human and Mouse Genomes**. Nature**. *In press.*

H. Yang, Y. Luan, T. Liu, Y. Wang, X. Wang, B. Zhang, Q. Jin, **J. Xu**, F. Song, C. Khunsriraksakul, …, R. Hardison, T. Wang, K. C. Cheng, F. Yue. A comprehensive map of cis-regulatory elements and 3D structure of the zebrafish genome. (*Revision in **Nature***)

H. Yang, H. Zhang, Y. Luan, T. Liu, K. G. Roberts, M. Qian, B. Zhang, W. Yang, V. Perez-Andreu, **J. Xu**, …, F. Yue, J. Yang. Non-coding germline GATA3 variants alter chromatin topology and contribute to pathogenesis of acute lymphoblastic leukemia (*Under review in **Nature Genetics***)

R. Fu, R. J. Gill, E. Y. Kim, N. E. Briley, E. R. Tyndall, **J. Xu**, …, F. Tian, Spherical nanoparticle supported lipid bilayers for the structural study of membrane geometry-sensitive molecules, **JACS,** 137 (2015), no. 44, 14031-14034.

L. Jiang, M. Yin, **J. Xu,** M. Jia, S. Sun, X. Wang, J. Zhang and D. Meng, The transcription factor bach1 suppresses the developmental angiogenesis of zebrafish, **Oxidative Medicine and Cellular Longevity** 2017 (2017), 2143875.

L. Jiang, M. Yin, X. Wei, J. Liu, X. Wang, C. Niu, X. Kang, **J. Xu,** …, R. Qian, N. Sun, A. Chen, R. Wang, J. Zhang, S. Chen and D. Meng, Bach1 represses Wnt/beta-Catenin signaling and angiogenesis, **Circulation Research** 117 (2015), no. 4, 364-375.

J. Xu, J. Liu, L. Jiang, D. Meng Roles of reactive oxygen species in pluripotent stem cell functions. **Chinese Journal of Pathophysiology**, 2013.