IDENTIFICATION OF STRUCTURAL VARIATIONS FROM WHOLE GENOME SEQUENCING OF CANCER PATIENTS

Dissertation
Zur Erlangung des Grades
Doktor der Naturwissenschaften

Am Fachbereich Biologie
Der Johannes Gutenberg-Universität Mainz

vorgelegt von
Riccha Sethi
geboren am 24. April 1986 in New Delhi

Mainz, 2023

Prüfungskommission

Dekan:

1. **Berichterstatter:**
2. **Berichterstatter:**

Tag der mündlichen Prüfung: **23.11.2023**

Summary

Cancer is largely driven by accumulation of somatic mutations that can be subdivided into small mutations (single nucleotide variations (SNVs), small insertions and deletions) and large structural variations (SVs). While SNVs affect single nucleotide, SVs can affect large stretches of DNA. Reliable identification of all mutations is key to understanding genetic diseases like cancer.

SVs can be identified by whole genome sequencing with conventional Illumina short-read sequencing (cWGS) being the most widely used approach. However, reliable prediction of SVs with short-reads (50-150bp) from fragmented DNA (~0.5kb) is challenging due to ambiguous mapping reads at repetitive regions and typically only few short reads span rearranged SV breakpoints with limited sequence overlap (due to read length). The 10X Genomics linked-reads sequencing (10XWGS) technology aims to mitigate limitations by linking short-reads to the original larger fragment of DNA (~10kb). In this study, we performed an unbiased evaluation of these two technologies with different types and sizes of SVs and compared their performance. The SVs commonly identified by both the technologies were highly specific, while the validation rate dropped for uncommon SVs. Despite the technological advantage, a particularly high false discovery rate (FDR) was observed for SVs found only by 10XWGS without any significant improvement in sensitivity. We proposed a sensitive and specific statistical approach to improve SV predictions from both technologies and characterized SVs from MCF7 breast cancer cell line and a primary breast tumor with high precision.

Due to the limited benefit of 10XWGS for sensitivity, we trained a random forest classifier in FuseSV for accurate predictions only from cWGS sequencing data. FuseSV integrates SV predictions from multiple bioinformatics tools and mitigates high FDR of cWGS with a novel set of features derived from alignment of reads to the reference genome, biological mechanisms of SVs and breakpoints of SVs clustered together to consider complex genomic rearrangements (CGRs). The performance of FuseSV classifiers was superior to all individual bioinformatics tools as well as combined use with 10XWGS.

SVs whether simple or complex can form chimeric fusion transcripts (CMTs). CMTs can be predicted from RNA-sequencing (RNA-seq) data but include also transcripts that occur without underlying mutation and are also present in healthy tissues. Here we propose a novel pipeline, FUdGE, that predict three types of CMT directly from somatic SVs: These include direct fusion transcripts or classical fusion genes, transcripts with intron (IR) and intergenic region retained (INR). FUdGE allows independent confirmation of expressed CMTs from matched RNA-seq data. We validated the approach in the same MCF7 cell line and a primary breast tumor sample and investigate CMTs in a cohort of liposarcoma samples. Here we observed that the majority of confirmed SV driven CMTs were classical fusion genes with a much smaller number of IR and INR events.

Conclusively, FuseSV enables accurate prediction of somatic SVs in cancer using only cWGS. While FUdGE provides an RNA-seq independent strategy for direct prediction of CMTs formed due to somatic SV event. The respective expressed CMT candidates can be confirmed independently with RNA-seq data. This alternative approach only predicts tumor-specific somatic SV driven CMTs, which is advantageous for personalized immunotherapy interventions considering CMTs as neo-antigen candidates.

Zusammenfassung

Krebs wird weitgehend durch die Ansammlung somatischer Mutationen verursacht. Hier werden kleine Mutationen (Einzelnukleotidvariationen (SNVs), kleine Insertionen und Deletionen) und große strukturelle Variationen (SVs) unterschieden. Während SNVs einzelne Nukleotide betreffen, können SVs große Abschnitte der DNA verändern. Die akkurate Detektion aller Mutationen ist der Schlüssel zum Verständnis genetischer Krankheiten wie Krebs.

SVs können durch Genomsequenzierung identifiziert werden, wobei die Illumina Sequenzierung (cWGS), welche kurze Sequenzen erzeugt, die am häufigsten verwendete Methode ist. Eine zuverlässige Vorhersage von SVs aus kurzen Sequenzen (50-150bp) von fragmentierter DNA (~0,5kb) ist jedoch schwierig, weil kurze Sequenzen besonders in repetitiven Regionen oft nicht eindeutig zugeordnet werden können und weil nur wenige Sequenzen tatsächliche die Bruchpunkte von SVs überspannen und dann auch nur einen geringen Überlapp zu jeder Seite haben. Die 10X Genomics linked-reads Sequenzierungstechnologie (10XWGS) zielt darauf ab, diese Einschränkungen durch die Verknüpfung mehrerer kurzer Sequenzstücke von einem ursprünglichen größeren DNA-Fragment (~10kb) zu überwinden. Wir haben daher eine unvoreingenommene Testung beider Technologien für verschiedene Arten und Größen von SVs durchgeführt. Hier waren die von beiden Technologien gemeinsam identifizierten SVs hochspezifisch, während die Validierungsrate bei separat gefunden SVs geringer war. Trotz des technologischen Vorteils wurde eine besonders hohe Falschentdeckungsrate (FDR) für SVs beobachtet, die nur mit 10XWGS gefunden wurden, ohne dass sich die Sensitivität wesentlich verbesserte. Für beide Technologien haben wir einen sensitiven und spezifischen Vorhersagealgorithmus entwickelt und charakterisierten SVs aus der MCF7 Burstkrebszelllinie und einem primären Brusttumor mit hoher Präzision.

Aufgrund des geringen Nutzens von 10XWGS für die Sensitivität, trainierten wir einen Random-Forest-Klassifikator in FuseSV für die genaue Vorhersagen von SVs nur aus cWGS-Sequenzierungsdaten. FuseSV integriert SV-Vorhersagen aus mehreren bioinformatischen Tools und kompensiert eine hohe FDR mit einem neuartigen Set von Merkmalen, die aus dem Alignment von Sequenzen zum Referenzgenom, den biologischen Mechanismen zur Entstehung von SVs und gruppierten Bruchpunkten im Falle von komplexen genomischen Rearrangements (CGR) abgeleitet werden. Unter Verwendung dieser Merkmale ermöglicht der FuseSV Klassifikator eine akkuratere Vorhersage als die einzelnen bioinformatischen Tools und auch als die kombinierte Nutzung mit 10XWGS.

SVs, ob einfach oder komplex, können chimäre Fusionstranskripte (CMTs) bilden. CMTs können aus RNA-Sequenzierungsdaten (RNA-seq) vorhergesagt werden, enthalten dann aber auch Transkripte die ohne eine unterliegende Mutation entstehen und auch in gesunden Geweben vorkommen können. Hier stellen wir FUdGE vor, welches drei Arten von CMTs direkt aus somatischen SVs vorhersagen kann: Dazu gehören direkte Fusionstranskripte oder klassische Fusionsgene, Transkripte mit intronischen (IR) und intergenischen Sequenzen (INR). FUdGE ermöglicht eine unabhängige Bestätigung vorhergesagter und exprimierter CMTs anhand von RNA-seq Daten desselben Samples. Wir validierten unseren Ansatz in der MCF7 Zelllinie und einem primären Brusttumor und untersuchten die Häufigkeit von CMTs in einer Liposarkom-Kohorte. Die große Mehrheit der bestätigten CMTs waren klassische Fusionsgene, wobei nur eine geringe Anzahl IR und INR durch ein somatisches SV-Ereignis gebildet wurde.

Zusammenfassend, ermöglicht FuseSV eine genaue Vorhersage von somatischen SVs in Tumorproben von cWGS Daten alleine. FUdGE hingegen bietet eine RNA-seq unabhängige Strategie zur direkten Vorhersage von CMTs, die durch somatische SVs entstehen. Die exprimierten CMT-Kandidaten können unabhängig anhand der RNA-seq Daten bestätigt werden. Dieser alternative Ansatz sagt nur tumorspezifische durch somatische SV bedingte CMTs vorher, was von Vorteil ist, wenn für personalisierte Immuntherapien CMTs als Neoantigen-Kandidaten berücksichtigt werden sollen.

List of figures

Figure 3-1: FuseSV integrates calls and use machine learning for the accurate predictions of SVs. 35

Figure 3-2: The ensemble of SV prediction tools and the random forest (RF) classifier in FuseSV generates higher sensitivity and precision 37

Figure 3-3: DDLS type of sarcoma has significantly higher number of SV 39

Figure 3-4: Chromosome 12 in majority of the DDLS samples is shattered and rearranged 40

Figure 3-5: Majority of SVs across liposarcoma samples have 4-15bp homology around the breakpoints 41

Figure 3-6: SVs on chromosome 12 of the DDLS samples are created with mechanism that utilize homology of length 4-15bp around the breakpoints 43

Figure 4-1: Explored categories of the chimeric mRNA transcripts (CMTs) due to an SV event in this study 62

Figure 4-2: The scheme followed by FUdGE for the prediction of CMTs 63

Figure 4-3: Distribution of possible and confirmed chimeric mRNA transcripts (CMTs) in MCF7 and the primary breast tumor sample 65

Figure 4-4: The performance of FUdGE on qPCR validated fusion transcripts from classical fusion genes 66

Figure 4-5: Landscape of direct fusion transcripts in the liposarcoma cohort 68

Figure 4-6: Landscape of somatic intron-retained and intergenic-retained fusion transcripts in the liposarcoma cohort 69

List of supplementary figures

Supplementary figure 3-1: Representation of basic/requantification, homology and cluster features of FuseSV pipeline 47

Supplementary figure 3-2: Nested SV in MZ-GaBa-018 cell line with back-to-back deletion and inversion event 48

Supplementary figure 3-3: The relative importance of features in trained random forest classifier 49

Supplementary figure 3-4: Landscape of different sizes of SVs in three sub-types of liposarcoma (DDLS, MLS and WDLS) 50

Supplementary figure 3-5: The shattering of chromosome 12 in DDLS samples 51

Supplementary figure 3-6: The shattering of chromosome 12 in WDLS samples 52

Supplementary figure 3-7: The shattering of chromosomes in K02K-4WMX7Q MLS sample 53

Supplementary figure 3-8: Majority of large sized SVs (size>100kb) are formed with DNA repair mechanisms utilizing long stretches of homology around the breakpoints 54

Supplementary figure 4-1: The schema for prediction of direct fusion transcripts by FUdGE. 73

Supplementary figure 4-2: Various FUdGE predicted CMTs (direct, intron-retention, intergenic-retention) have support from RNA-seq data in MCF7 and the primary breast tumor sample.. 74

Supplementary figure 4-3: The underlying SVs for FUdGE predicted CMTs (direct, intron-retention, intergenic-retention) have support from WGS data in MCF7 and the primary breast tumor sample 75

Supplementary figure 4-4: The junction and spanning reads support FUdGE predicted direct fusion transcripts that were validated by qRT-PCR/qPCR 76

Supplementary figure 4-5: Most of the MLS samples in liposarcoma cohort are characterized with *FUS-DDIT3* fusion gene 77

Supplementary figure 4-6: The distribution of intron-retained transcripts expressed in the liposarcoma samples was confounded with fusions between paralogous genes 78

Supplementary figure 4-7: The distribution of intergenic-retained fusion transcripts expressed in the liposarcoma samples was characterized with fusion of an intergenic region with the *AK2* gene 79

List of supplementary tables

Supplementary table 3-1: The list of WGS samples and validated SVs used for training and testing FuseSV pipeline 55

Supplementary table 3-2: Explanation of different features included in three set of features (basic, homology and cluster) for training the random forest classifiers. 56

Supplementary table 3-3: Number of different type and size of SVs in liposarcoma samples 59

Supplementary table 4-1: Distribution of the CMTs in MCF7 and the primary breast tumor sample that were confirmed with RNA-seq data 80

Supplementary table 4-2: The distribution of recurrent direct fusion transcripts in DDLS, MLS and WDLS type of liposarcoma samples 81

List of abbreviations

SVs	Structural variations
SNV	Single nucleotide variation
INDEL	Insertion and deletion
bp	Base pair
CNV	Copy number variation
CGR	Complex genomic rearrangement
BFB	Breakage fusion bridge
CML	Chronic myeloid leukemia
FISH	Fluorescence in-situ hybridisation
CGH	Comparative genomic hybridization
SNP	Single nucleotide polymorphism
NGS	Next generation sequencing
10XWGS	10X Genomics linked-reads sequencing
GEM	Gel bead in emulsion
HMW	High molecular weight
DSB	Double stranded break
HR	Homologous recombination
NAHR	Non-allelic homologous recombination
BIR	Break induced replication
MMBIR	Microhomology mediated BIR
FoSTeS	Fork stalling and template switching
c-NHEJ	Classical non-homologous end-joining
alt-EJ	Alternate end joining
SSA	Single strand annealing
MMEJ	Microhomology-mediated end joining
SD-MMEJ	Synthesis dependent MMEJ
TAD	Topologically associated domain
IR	Intron-retention transcript
WGS	Whole genome sequencing
WDLS	Well-differentiated liposarcoma
DDLS	De-differentiated liposarcoma
MLS	Myxoid liposarcoma
PLS	Pleomorphic liposarcoma
MPLS	Myxoid pleomorphic liposarcoma
UPLS	Undifferentiated pleomorphic liposarcoma
STS	Soft tissue sarcoma
cWGS	Conventional whole genome sequencing by Illumina

FDR	False discovery rate
RP	Discordant read-pair
SR	Split reads
RD	Read depth
LA	Local *de-novo* assembly
AUC	Area under curve
ROC	Receiver operating curve
AMT	Altered mRNA transcript
CMT	Chimeric mRNA transcript
RNA-seq	RNA sequencing
INR	Intergenic region retained transcript

Table of Contents

Prüfungskommission II

Summary III

Zusammenfassung IV

Acknowledgements V

List of figures VI

List of supplementary figures VII

List of supplementary tables VIII

List of abbreviations IX

Table of contents XI

1 Introduction 1

1.1 Structural variations (SVs) 1

1.2 Technologies for identification of SVs 2

1.2.1 Illumina short-reads sequencing 2

1.2.2 10X Genomics linked-reads sequencing 3

1.3 Mechanisms of SVs 3

1.4 Functional consequences of SVs 4

1.4.1 Fusion gene 4

1.4.2 Gene dosage 4

1.4.3 Enhancer hijacking and altered expression 5

1.4.4 Intron-retention in expressed transcripts 5

1.5 Relevance of SVs in immunotherapy 5

1.6 SVs in liposarcoma 6

1.7 Outline of thesis 7

2 Integrative analysis of structural variations using short-reads and linked-reads yields highly specific and sensitive predictions 8

3 Reliable predictions of SVs from Illumina short-reads sequencing 33

3.1 Introduction 33

3.2 Results 34

3.2.1 Overview of FuseSV 34

3.2.2 FuseSV predicts SVs with high sensitivity and precision 36

3.2.3 Application of FuseSV to liposarcoma cohort 38

3.2.4 Chromosome shattering in DDLS and WDLS sub-types of liposarcoma 40

3.2.5 Mechanisms of SVs across liposarcoma 41

3.3 Discussion 43

3.4 Methods 44

3.4.1 Whole genome sequencing (WGS) samples and upstream processing 44

3.4.2 Prediction of SVs with an ensemble of bioinformatics tools 45

3.4.3 Integration of calls 45

3.4.4 Feature extraction 45

3.4.5 Random forest classifier 46

3.4.6 Liposarcoma samples 46

3.4.7 Shattering of chromosome 46

3.4.8 Statistical analysis and graphs 46

3.5 Supplementary figures 47

3.6 Supplementary tables 55
4 Direct prediction of neo-antigens from somatic SVs 61
4.1 Introduction 61
4.2 Results 63
4.2.1 FUdGE Scheme 63
4.2.2 Confirmation of FUdGE predicted CMTs by RNA-seq in breast cancer cell line and primary breast tumor sample 64
4.2.3 Confirmation of FUdGE predicted CMTs by qRT-PCR in breast cancer cell line and primary breast tumor sample 66
4.2.4 Analysis of somatic CMTs in liposarcoma cohort 67
4.2.5 Limited expression of intron and intergenic-retained CMTs in liposarcoma 68
4.2.6 Recurrent CMTs in liposarcoma 69
4.3 Discussion 70
4.4 Methods 71
4.4.1 Chimeric fusion transcripts (CMT) predictions from SVs in FUdGE 71
4.4.2 Application of FUdGE to MCF7 and primary breast tumor sample 72
4.4.3 Application of FUdGE to liposarcoma samples 72
4.5 Supplementary figures 73
4.6 Supplementary tables 80
5 Future outlook 84
6 References 86
7 Curriculum vitae 92

1 Introduction

Cancer is the uncontrolled growth of normal cells in the body that can mutate, expand and spread to other regions of body from the local site. A continuous effort in the cancer research began as early as 460-370 B.C. with coining of the term "carcinos" and "carcinoma" by Greek physician Hippocrates. A key milestone in this effort was contributed by Percivall Pott in 1775 with his study of squamous cell carcinoma by exposure to chimney soot (1). As reviewed in (2), a simultaneous effort on the origin of cancer were proposed in the late 19th and early 20th century. A theory proposed by David Paul von Hansemann and Theodor Boveri involved observation of abnormality in the numbers of chromosomes and mitosis in cancer cells. However, the research by Paul Ehrlich in 1909 linked cancer with suppression of immune system through the concept immune surveillance. This serves as a basis for treatment of cancer by immunotherapy where the inherent immune system is trained and enhanced to fight cancer.

The cancer cells are characterized by a broad spectrum of mutations, varying from single nucleotoide variants (SNV), small insertions and deletions (INDEL) to structural variations (SVs) that affect more than 50 base pairs (bp). Several studies have associated SVs to a genetic disease (3,4) but not all SVs have functional impact. Generally, mutations in cancer can be germline i.e., inherited from parents or acquired from environmental factors leading to somatic mutations solely found in cancer cells. These mutations can be beneficial, neutral (passenger mutations) or harmful (driver mutations) based on its functional impact. The genetic disease like cancer is driven by higher ratio of driver mutations in comparison to the passenger mutations (5). These driver mutations affect normal functions of the cell and are often localized in the tumor suppressor genes (example *BRCA1*, *BRCA2*, *p53* genes), oncogenes (example *RAS*, *HER2* genes) and DNA repair genes (example *BRCA1*, *BRCA2*). Thus, correct identification such driver mutations is crucial for better understanding of the cancer biology, its diagnosis and treatment.

1.1 Structural variations (SVs)

Simple SVs include deletion of genomic segments, duplication, inversion and translocation of genomic segments between different chromosomes. Each SV contains at least two breakpoints that merges two distant genomic locations. Deletions and duplications are also known as copy number variations (CNV) as the genomic segment is deleted or duplicated respectively. While copy neutral variations include inversions where a segment is inverted and inserted at same or any other location, and translocations where a genomic segments from different chromosomes merge. Translocation can be balanced when segments from different chromosomes are exchanged or unbalanced when segment from one chromosome is either deleted, duplicated or inverted and inserted at another chromosome. SVs can also be complex in nature where multiple events are incorporated in the genome. One of such complex genomic rearrangement (CGR) is chromothripsis where the chromosome is shattered in a single catastrophic event (6). It is characterized by localization of multiple breakpoints in a confined genomic section with an alternating oscillation of 2 or 3 copy numbers. It is prominently found on one chromosome while another type of CGR like chromoplexy involves several chromosomes being joined together with multiple DNA strand breakage and joining (7). This causes clustering of multiple breakpoints from more than five chromosomes with balanced rearrangements. One of the earliest discovered forms of CGR was breakage fusion bridge (BFB) cycles (8). It was first detected by Barbara McClintock in 1930s where two sister chromatids fuse together due to lack of telomeres and form mitotic bridges leading to DNA breakage. BFB cycles are characterized with fold-back-inversions with copy number changes. With the advancement in algorithms and technologies many new types of CGR are discovered in various cancer types (4). These include chains, cycles and bridges of templated insertions that involve shuffling of different chromosome sections in a string that can either form bridges (when start chromosome has a gap), chains (when string of inserted sections do not revert to start chromosome) and cycles (a cycle of inserted sections from multiple chromosomes is observed). Furthermore, (9) explored other types of CGR like pyrgo and rigma. The pyrgo contains clusters of duplication like events with low number of junction-copy numbers in a confined genomic segment while rigma has clusters of deletion like events with low number of junctions merged where an interval section also reached zero copy number. In the pan-cancer analysis of CGR, it was seen that endometrial, ovarian and breast cancer were significantly enriched with pyrgo while rigma was significantly enriched in Barrett's esophagus and esophageal adenocarcinoma. Henceforth, most cancer genomes are enriched with different types of SVs and CGRs that are being discovered with the advancements in cancer research.

1.2 Technologies for identification of SVs

The advancement in cancer research is also linked to the advancement in technologies used to study the cancer cells. One of the earliest technologies used for the detection of cancer-related abnormalities includes cytogenetic techniques. It was Peter Nowell and David Hungerford in 1960 who first detected Philadelphia chromosome in chronic myeloid leukemia (CML) by visualization of cancer cells under the microscope (10). Later in 1980 researchers developed DNA fluorescence in situ hybridization (FISH) technology that binds fluorescent probes to section of DNA (11) to visualize rearrangements under the fluorescence microscope. This cost-effective technique allowed cancer biologists to visualize several abnormalities in cancer cell including CGR like breakage-fusion-bridges (12).

Apart from cytogenetic techniques, hybridization-based techniques became popular since 90s. These technique use labeled complementary DNA or RNA probes to bind specific DNA or RNA sequence on a plate. It has been used extensively to identify DNA copy number changes and point mutations in tumor cells with comparative genomic hybridization (CGH) (13) and single nucleotide polymorphism (SNP) arrays (14). These techniques provide better resolution than cytogenetic technique, but they are inadequate for detection of balanced copy neutral variations like inversions.

A big revolution in cancer research came with development of next generation sequencing (NGS) techniques. With the base pair sequence information at disposal, it has been possible to link mutations in genome (including both coding and non-coding/intergenic regions) with the evolution of cancer and its metastasis. This enabled pan-cancer analysis of mutations (both SNPs and SVs) in 38 sub-types of cancer cells from 2658 patients under PCAWG consortium (4). Such joint efforts have revealed several patterns and signatures of SVs and CGR in different cancer by usage of paired-end short reads sequencing from the Illumina platform.

1.2.1 Illumina short-reads sequencing

One of the most predominant NGS technologies used in research is Illumina's paired–end short reads sequencing. The detection of mutations with Illumina's NGS involves following steps:

Library preparation: A short insert of ~0.5kb is prepared from the isolated genomic DNA or cDNA and ligated with 5' and 3' adapter sequences. The tagged fragments are then amplified by PCR and purified.

Cluster generation: The amplified libraries are loaded on flow cells where they bind to oligos complementary to the adapters. Next, the bridge amplification amplifies attached libraries in the clusters.

Sequencing-by-synthesis: Illumina technology uses base-by-base sequencing where a terminator-bound dNTPs is added. The dNTPs are fluorescently labeled that allows visualization of sequence when they are incorporated in sequence of reads that would be synthesized as complementary to the library's sequence. This is repeated for "n" cycles that generates a read of length "n" bases. In case of paired-end sequencing, the fragment is sequenced from both the sides as depicted in figure.

Bioinformatics analysis: After the generation of reads, they are aligned to the reference genome and signals from aligned reads can be used to detect different types of variations.

A major shortcoming of this technology is contributed by the short-fragment DNA library preparation that are sequenced with even shorter reads of length typically 2x150 bp. As a result, this technique proves inefficient in aligning reads originating from repetitive elements in the human genome that are often associated with SVs (15) .

1.2.2 10X Genomics linked-reads sequencing

To deal with limitations of Illumina short-reads sequencing, recently linked-reads sequencing technology (10XWGS) was introduced by 10X Genomics. This utilizes reads derived from high molecular weight (HMW) DNA with typical fragment size between 50 – 100 kb in order to supply long-range information (16). This approach captures high molecular weight (HMW) DNA molecules in “Gel beads in EMulsion (GEM)”. After encapsulation, HMW DNA is sheared into smaller fragments (0.5 kb), labeled with GEM specific barcodes and subjected to Illumina short-reads sequencing (paired-end). The attached barcodes link each short read-pair to its originating HMW DNA. The 10XWGS bioinformatics pipeline utilizes this information to reconstruct the initial long HMW DNA molecule that links longer sections of the genome together into a phased haplotype and resolving SVs in low complexity regions of the genome. Theoretically, this enables highly specific and sensitive prediction of SVs.

1.3 Mechanisms of SVs

The cellular mechanisms behind the generation of SVs are an active field of research. Primarily, an SV occurs due to inefficient DNA repair of a double stranded break (DSB) in DNA, collapse of the replication fork, telomere decay or enzymatic activity. The genomic sequence around the repaired breakpoints can indicate the active DNA repair mechanism in the cell. One of the key genomic signatures suggesting the repair mechanism is the length of microhomology around the broken DNA and presence of small INDEL at the repaired DNA section (17). Microhomology is defined as the length of nucleotides that are same around the broken DNA that eventually constitutes the two breakpoints of an SV. These can be located at the junction or at some distance from the merged genomic segments of the SV.

Different DNA repair mechanisms are categorized into break and ligate, and template and ligate. As reviewed in (17) and (18), template and ligate repair mechanisms typically begin with the resection of DNA by BLM helicase and DNA2 endonuclease or EXO1 exonuclease, next is the formation of 3’single stranded tails, RAD51-mediated strand invasion and DNA synthesis by polymerase. One of the mechanisms included in this category is homologous recombination (HR) that uses sister chromatid or corresponding homologous sequence as template to fix broken DNA section. This is usually error free and requires long stretches of homologous sequences around the breakpoints. The DNA-RAD51 nucleoprotein combination searches for suitable template by the formation of displacement-loop. Next in the synthesis dependent strand annealing, the D-loop is dissociated after synthesis of several hundred base pairs and nascent strand pairs with single-stranded DNA on opposite side of break. The double-strand break repair (DSBR) model involves creation of a double holliday junction structure which completes the repair (19). Although HR is accurate in repair, but sometimes RAD51-DNA complex can use nearly similar sequence from non-allelic template. This is called non-allelic homologous recombination (NAHR). NAHR can generate a lot of chromosomal rearrangements as nearly 45% of human genome is rich in repeats. One of the NAHR derived variation is seen in *BRCA1* gene whose intronic region is covered with *Alu* repeats. The inaccurate DNA repair is the major cause of mutations in *BRCA1* deficient cancers like breast cancer (20). Another error-prone HR pathway is break-induced replication (BIR) that involves invasion of the homologous template by one broken DNA end and synthesis of extremely long DNA >100kb. Such breaks are often observed in the collapsed replication fork or eroded telomeres that were reported in many cancers (21). The BIR pathway prefers one ended DSB, however, this mechanism is also active when there is limited homology of <150bp around the two ended DSB (22). One of the variations of BIR is microhomology mediated BIR (MMBIR) pathway that relies on smaller homologous sequence (or microhomology) of 2-5bp around the break for annealing the dissociated single-stranded DNA. MMBIR is also referred to as fork stalling and template switching (FoSTeS) that involves multiple rounds of template switching often leading to CGR (23).

The break and ligate mechanisms include classical non-homologous end-joining (c-NHEJ), alternate end-joining (alt-EJ) and single strand annealing (SSA). The c-NHEJ repair mechanism is normally accurate or with small INDEL around the DNA breaks and require minimal microhomology (0-4bp) (24). It is active during the interphase and begins by attachment of Ku70/80 heterodimers to blunt DNA ends or ssDNA with less than 5 bp. Further, the breaks are ligated via DNA ligase 4 or other enzymes depending on whether DNA break can be ligated directly or not (25). Another break and ligate mechanism include alternate end-joining (alt-EJ) which is originally described for all the repair mechanisms active in the absence of c-NHEJ. It is an error-prone mechanism that require microhomology of 1-8bp around the DNA breaks. Alt-EJ can be further classified based on the properties of repair junctions. The first one is microhomology-mediated end joining (MMEJ) that begins by the resection of DNA to ssDNA, annealing of microhomologous sequences at the breaks, deletion of 3' tails of non-homologous ends, fill-in-synthesis and ligation. It requires microhomologous sequence of length between 1-8 bp for annealing. The second one is the synthesis-dependent MMEJ (SD-MMEJ) that introduces *de novo* microhomologous sequence as an additional step during the DNA synthesis. SD-MMEJ is characterized with the presence of INDELs around the DNA breaks once repaired (26). Another version of MMEJ uses DNA Polymerase theta for end joining of the nicked DNA (27). The DNA Polymerase theta uses very short microhomology of 1-2 bp to prime DNA synthesis. Because of the multiple rounds of annealing, synthesis and dissociation, template insertions can be introduced in the genome. Apart from the above-mentioned break and ligate mechanisms, another type is the SSA. It is like HR based DNA repair as it requires DNA resection and formation of 3' single strand DNA ends (28). However, instead of RAD51-mediated DNA repair, the SSA uses small section of homologous sequences for annealing, 3' single-stranded tails are clipped and ligated. It is characterized with presence of homologous sequence of length greater than 15bp but less than 100bp.

The cancer cells undergo extreme stress imposed by high degree of DNA replication or stress imposed by chemotherapeutic agents. Consequently, they rely on error prone DSB repair for survival that introduces mutations in genome. The mutational signatures observed in several cancer sub-types can indicate the type of repair mechanism. One well known example is "BRCAness" signature which is reported as Signature 3 in Catalog of Somatic Mutations in Cancer (COSMIC). This signature is observed in HR-deficient cancer types and is characterized by base substitutions, INDEL>3bp with microhomology around the breakpoints which indicates alt-EJ repair pathway (29). Hence, understanding the genomic features around the breakpoints of SVs can point to the most prominent DNA repair mechanism used by cancer cells for its survival.

1.4 Functional consequences of SVs

SVs in the cancer genome can rearrange the genome and have several forms of functional impact. These include following:

1.4.1 Fusion gene

A fusion transcript is generated when a gene is placed next to other gene due to SV and is expressed. Such chimeric fusion transcripts can drive cancer as seen in following cases: *BCR-ABL* fusion gene in CML (10), *TMPRSS2-ERG* in prostate cancer (7) and *EML4-ALK* in non-small cell lung cancer (30).

1.4.2 Gene dosage

SVs can increase or decrease copy number of sections of the genome. This cause transcriptional dosage changes like higher expression of oncogenes or reduced to no expression of tumor suppressor genes. A canonical example of this case is the overexpression of *MYC* oncogene in 13-17% cases of the breast cancer (31) and loss of *CDKN2A* tumor suppressor gene in the brain cancer (32).

1.4.3 Enhancer hijacking and altered expression

The disruption of topologically associated domain (TAD) boundaries by SVs can also alter the expression of genes. TAD are the DNA sequences on chromosomes that interact physically with each other during the tightly packed interphase stage of cell cycle. The two different TAD boundaries can be separated by megabases but bring an enhancer close to a distant gene when packed tightly. Moreover, deletion of a TAD boundary can establish new promoter-enhancer relationships that can alter gene expression and drive diseases (33–35).

Enhancer hijacking and remodeling of chromatin topology has been observed in several cancer types like acute myeloid leukemia (36), medulloblastoma (37) and T-cell acute lymphoblastic leukemia (38).

1.4.4 Intron-retention in expressed transcripts

Intron-retention (IR) in transcripts is one of the classes of transcripts generated due to alternative splicing. This involves expression of transcripts with intron region due to mis-splicing of the introns in mRNA and is primarily active after transcription. Such IR transcripts can play regulatory role in controlling the expression of genes (39). Additionally, a recent study in the cancer patients have inferred nearly 18% of SNV related to splicing lead to IR transcripts (40). While several studies have indicated widespread evidence of IR transcripts in the cancer cells (41,42), the expression of IR transcripts in normal/wild-type cells have also been reported (41). This indicates that their expression in cancer studies might be overestimated and requires effort in finding IR transcripts expressed only in the cancer cells. The detection of somatic IR transcripts can help link several mutations in the genome (like SNV and SV) with their functional impact as creation of chimeric transcript with retained intron region.

1.5 Relevance of SVs in immunotherapy

The cancer cells adapt to survive attacks from the immune system by various mechanisms. One of the mechanisms include blockage of immune checkpoints on T lymphocytes to mimic immunosuppressive activity. The discovery of such checkpoint inhibitors and the mechanism to hijack this pathway led to Nobel Prize award in 2018 to two immunologists, namely, James P. Allison and Tasuku Honjo. Their work led to immunotherapies blocking immune checkpoints (PD-1 (43) and CTLA-4 (44)) with monoclonal antibodies that enhanced anti-tumor immune response in the cancer patients.

Immune checkpoint-based therapies are promising treatment for cancer, but they have lower efficacy in the solid tumors. Henceforth, novel personalized interventions with neoantigen based immunotherapy can be particularly attractive in such cases. As reviewed in (45) and discussed at the beginning, the cancer cells are driven by genomic mutations like SNV, SVs, INDEL etc. These variants and their derived mutant proteins are presented on the surface of antigen presenting cells via major histocompatibility complex (MHC)-I or MHC-II in the human body. The mutated protein-MHC combination can elicit an immune response by interacting with T lymphocytes (CD8+ or CD4+) and eliminate those cancer cells by expansion of T-cell clones recognizing the mutant protein. The derived mutant proteins from somatic mutations that occur only in the cancer cells have particularly strong immunological response that is also exempt from central tolerance. Such cancer specific antigens are called as neoantigens. The neoantigens are further classified into shared or personalized neoantigens. The shared neoantigens is derived from mutated proteins that are common amongst different cancer patients while personalized neoantigen is derived from mutated protein that are uniquely present in individual cancer patient. The cancer vaccines targeting neoantigens can decrease the probability of immune escape of cancer cells. Some of the examples of cancer vaccination programmes targeting shared neoantigens include mutRas and mutP53. On the other hand, the personalized neoantigen based vaccine are designed for specific patient and thus, exploit their complete mutanome for best treatment of cancer. Many such personalized vaccine development programmes are under clinical trials that hold an optimistic future in cancer treatment.

The current repertoire of neoantigens based cancer vaccine programmes is primarily derived from SNV with little to no exploitation of neoantigens derived from SVs. This can be attributed to several reasons like unreliable predictions of SVs by existing bioinformatics tools, lack of bioinformatics tools to directly predict functional consequences of SVs, enrichment of SVs in repetitive regions of the human genome that are difficult to resolve by NGS, complex nature of cancer derived mutations that can be clonal and sub-clonal, and cost associated with whole genome sequencing (WGS) for prediction of SVs. Nevertheless, few recently published research studies have shown promising immune response to neoantigens derived from SVs or fusion genes that can be formed due to SV event. The first study demonstrated T cell response in PBMC of the mesothelioma patient that was treated with peptides of potential neoantigens derived from chromosomal rearrangements (46). The other studies have shown stimulatory T cell response to the neoantigens derived from fusion genes (47,48). This implies promising potential of undiscovered neoantigens derived from SVs that can elicit an immunological response, including cancers with low mutation burden like head and neck cancer (47).

1.6 SVs in liposarcoma

In this thesis we focus on expanding the genomic landscape of liposarcoma. It is the most common form of adult soft-tissue sarcoma (STS) that has mesenchymal origin and is highly heterogenous (~70 subtypes classified by WHO (49)). This cancer type is difficult to diagnose and treat because of its great diversity. There are six aggressive forms of liposarcoma with complex karyotypes: 1. Well-differentiated liposarcoma (WDLS), 2. De-differentiated liposarcoma (DDLS), 3. Myxoid liposarcoma (MLS), 4. Pleomorphic liposarcoma (PLS), 5. Myxoid pleomorphic liposarcoma (MPLS), and 6. Undifferentiated pleomorphic liposarcoma (UPLS).

WDLS covers 40-45% of liposarcoma cases in adults and are characterized with ring chromosomes amplifying *MDM2* proto-oncogene. It can develop to a poorly differentiated form of sarcoma in non-adipocytes, known as DDLS. Like WDLS, DDLS is also characterized with ring chromosomes that amplifies MDM2 along with other oncogenes on chromosome 12q13~15 arm (50–52). DDLS is more aggressive than WDLS with some common mutational landscape. Even though both these form of liposarcoma have *MDM2* amplification, its association with prognosis of disease is debatable. One study observed negligent prognosis effect with *MDM2* amplification (53) while two other studies reported low survival rate of patients with high *MDM2* amplification (54,55). Moreover, 90% of WDLS/DDLS patients have *CDK4* amplified (also located on amplified chromosome 12 arm) (56). But, in this case, higher amplification levels of *CDK4* are linked to prognosis in WDLS/DDLS. However, only high-grade DDLS is established to have higher *CDK4* amplification levels in comparison to WDLS and low-grade DDLS (57). Since amplification of chromosome 12q arm is common between WDLS/DDLS cases, majority of genes in this section are amplified that also include other genes like *HMGA2*, *TSPAN31*, *CPM* and *YEATS4* (57). Moreover, mutational load of both WDLS and DDLS is low and very few DDLS patients have mutated *TP53* as published in TCGA dataset (58). One difference in the genomic landscape of WDLS and DDLS is presence of *CTDSP1/2-DNM3OS* fusion gene in DDLS patients which is completely absent in WDLS patients (59).

MLS constitutes 15-20% of liposarcoma cases in world. It is characterized by translocation between chromosome 12 and 16 and poorly differentiated round cell morphology (60). The translocation causes fusion of *FUS-DDIT3* genes that is present in majority of MLS patients (61). In the remaining cases, a chromosome translocation between chromosome 12 and 22 is reported that fuses *DDIT3* with EWSR1 (62,63). Both these fusions have been reported in number of studies but none of these fusions has prognostic value (64).

PLS is highly malignant and rare type of sarcoma that occurs in 5-10% of liposarcoma cases and is characterized with pleomorphic lipoblasts (65). Another rare form of liposarcoma includes myxoid pleomorphic liposarcoma (MPLS) is the most recent, aggressive subtype of sarcoma that is prominent in children and adolescents. This tumor’s histologic features are similar to MLS and PLS but doesn't contain fusion genes and amplified regions that are established in MLS, PLS, WDLS and DDLS. MPLS has been observed to have whole chromosome gains in chromosomes 1, 6-8 and 18-21 with losses in chromosomes 13, 16 and 17 that also cause loss of tumor suppressor gene *RB1* (65). Undifferentiated pleomorphic liposarcoma (UPLS) is another rare, highly aggressive, high-grade myofibroblastic sarcoma whose cell of origin is unclear (49).

The general treatment of liposarcoma is removal of localized tumor, radiotherapy for reduction of tumor size and chemotherapy to treat metastatic disease. Some of the types of liposarcoma are sensitive to chemotherapy like MLS (66) but many others (25-50%) sarcoma patients redevelop tumor or metastatic tumor. The heterogeneity of sarcoma makes the usual treatment of disease inadequate. Hence, it is required to understand both genomic and clinical aspects of different liposarcoma types and find novel treatments. Two past studies have compared the landscape of genomic mutations (considering SNV, CNV, fusion genes and expressed genes) in adult STS (58,67). It was reported that STS had lower somatic mutation burden in terms of SNV (1.7 SNVs in STS in comparison to 6.1 in melanoma). Contrastingly, STS have higher percentage of CNV and fusion genes in comparison to many other cancer types like renal carcinoma and melanoma respectively. Moreover, immune cell infiltration was often detected in DDLS and UPS patients. Considering liposarcoma have lower mutation burden but higher number of expressed fusion genes, this type of cancer can benefit from neoantigen based immunotherapy.

1.7 Outline of thesis

This thesis is further divided into following chapters: a) **Chapter 2** covers a published benchmarking study for prediction of SVs from Illumina short-reads and 10X Genomics linked-reads sequencing. In this study, a logistic regression machine learning model was trained for accurate prediction of SVs when a sample is sequenced by either or both the technologies; b) **Chapter 3** focuses on the development of machine learning pipeline for reliable prediction of SVs using Illumina short-reads sequencing. This approach is an improvement over logistic regression model by inclusion of novel features derived from mechanisms of SV and CGR; c) **Chapter 4** focusses on functional impact of SVs in terms of direct prediction of expressed chimeric fusion transcripts with an underlying SV event. I further analyzed a cohort of liposarcoma patients to study the landscape of genomic variations contributed by SVs and chimeric fusion transcripts in this low mutational burden class of cancer.

2 Integrative analysis of structural variations using short-reads and linked-reads yields highly specific and sensitive predictions

Sethi R, Becker J, Graaf Jd, Löwer M, Suchan M, et al. (2020). Integrative analysis of structural variations using short-reads and linked-reads yields highly specific and sensitive predictions. PLOS Computational Biology 16(11): e1008397.

Abstract:

Genetic diseases are driven by aberrations of the human genome. Identification of such aberrations including structural variations (SVs) is key to our understanding. Conventional short-reads whole genome sequencing (cWGS) can identify SVs to base-pair resolution, but utilizes only short-range information and suffers from high false discovery rate (FDR). Linked-reads sequencing (10XWGS) utilizes long-range information by linkage of short-reads originating from the same large DNA molecule. This can mitigate alignment-based artefacts especially in repetitive regions and should enable better prediction of SVs. However, an unbiased evaluation of this technology is not available. In this study, we performed a comprehensive analysis of different types and sizes of SVs predicted by both the technologies and validated with an independent PCR based approach. The SVs commonly identified by both the technologies were highly specific, while validation rate dropped for uncommon events. A particularly high FDR was observed for SVs only found by 10XWGS. To improve FDR and sensitivity, statistical models for both the technologies were trained. Using our approach, we characterized SVs from the MCF7 cell line and a primary breast cancer tumor with high precision. This approach improves SV prediction and can therefore help in understanding the underlying genetics in various diseases.

My contribution:

Algorithm development: 90%

Data Processing: 100%

Data Analysis: 85%

Manuscript Writing: 85%

RESEARCH ARTICLE

Integrative analysis of structural variations using short-reads and linked-reads yields highly specific and sensitive predictions

Riccha Sethi[1], **Julia Becker**[1], **Jos de Graaf**[1], **Martin Löwer**[1], **Martin Suchan**[1], **Ugur Sahin**[1,2‡]*, **David Weber**[1‡]*

1 TRON—Translational Oncology at the University Medical Center of Johannes Gutenberg University Mainz gGmbH, Mainz, Germany, **2** University Medical Center of the Johannes Gutenberg University, Mainz, Germany

‡ These authors are joint senior authors on this work.
* sahin@uni-mainz.de (US); david.weber@tron-mainz.de (DW)

OPEN ACCESS

Citation: Sethi R, Becker J, Graaf Jd, Löwer M, Suchan M, Sahin U, et al. (2020) Integrative analysis of structural variations using short-reads and linked-reads yields highly specific and sensitive predictions. PLoS Comput Biol 16(11): e1008397. https://doi.org/10.1371/journal.pcbi.1008397

Editor: Florian Markowetz, University of Cambridge, UNITED KINGDOM

Received: January 10, 2020

Accepted: September 24, 2020

Published: November 23, 2020

Copyright: © 2020 Sethi et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The sequencing data for MCF7 and primary tumor is available at European Genome-phenome Archive: EGAD00001005724. The code for pipeline is available at GitHub repository (https://github.com/TRON-Bioinformatics/SV-LRvsSR).

Funding: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 789256) awarded to US. The funders had no role in

Abstract

Genetic diseases are driven by aberrations of the human genome. Identification of such aberrations including structural variations (SVs) is key to our understanding. Conventional short-reads whole genome sequencing (cWGS) can identify SVs to base-pair resolution, but utilizes only short-range information and suffers from high false discovery rate (FDR). Linked-reads sequencing (10XWGS) utilizes long-range information by linkage of short-reads originating from the same large DNA molecule. This can mitigate alignment-based artefacts especially in repetitive regions and should enable better prediction of SVs. However, an unbiased evaluation of this technology is not available. In this study, we performed a comprehensive analysis of different types and sizes of SVs predicted by both the technologies and validated with an independent PCR based approach. The SVs commonly identified by both the technologies were highly specific, while validation rate dropped for uncommon events. A particularly high FDR was observed for SVs only found by 10XWGS. To improve FDR and sensitivity, statistical models for both the technologies were trained. Using our approach, we characterized SVs from the MCF7 cell line and a primary breast cancer tumor with high precision. This approach improves SV prediction and can therefore help in understanding the underlying genetics in various diseases.

Author summary

Cancer and many other diseases are often driven by structural rearrangements in the patients. Their precise identification is necessary to understand evolution and cure for the disease. In this study, we have compared two sequencing technologies for the identification of structural variations i.e. Illumina's short-reads and 10X Genomics linked-reads sequencing. Short-reads sequencing is already known to have high false discovery rate for structural variations, while, an unbiased performance evaluation of linked-reads sequencing is missing. Hence, we evaluate the performance of these two technologies using

study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: Ugur Sahin is co-founder and shareholder of TRON, co-founder and CEO of BioNTech SE.

computational and PCR based methodologies. Moreover, we also present a statistical approach to increase their performance, supporting better detection of structural variations and thus further research into disease biology.

This is a *PLOS Computational Biology* Benchmarking paper.

Introduction

Structural variations (SVs) are large rearrangements in the genome, including deletions, duplications, inversions, translocations and insertions, and drive the development of diseases like cancer, autism and mendelian disorders [1]. One well-known example is the Philadelphia chromosome, an interchromosomal rearrangement (translocation) between chromosome 22 and chromosome 9 in chronic myeloid leukemia. This SV causes the fusion of two distantly located genes, BCR and ABL1, forming an active tyrosine kinase which leads to uncontrolled growth of cells [2]. Even a single SV can alter the expression of genes by functional impacts such as enhancer hijacking, truncation or disruption of tumor suppressor genes and amplifications of oncogenes. Hence, resolving such chromosomal rearrangements holds the key to understanding the causes behind genetic diseases [1].

Historically, large genomic alterations could be identified microscopically using karyotyping that allows genome wide identification but only at a very low resolution. More recently, SVs that lead to copy number variations (CNVs) could also be identified using array-comparative genomic hybridization, but without breakpoint information.

The onset of next-generation sequencing enabled a genome-wide read out for all SV types at base pair resolution. In theory, conventional whole genome sequencing (cWGS) by Illumina allows the identification of all SVs in an individual sample. However, a major shortcoming of this technology is contributed by the short-fragment DNA library preparation for sequencing with DNA fragment of size typically below 0.5 kb. Moreover, these short-fragments are sequenced with even shorter reads of length typically 2x150 bp. Therefore, this technique proves inefficient in aligning reads originating from repetitive elements in the human genome that are often associated with SVs [3]. Multiple tools and algorithms exist for prediction of SVs from cWGS data [4], but due to the described limitations, they often lack sensitivity and have high false discovery rates (FDR), especially in repetitive regions [5]. To reduce FDR, many studies consider SVs predicted by multiple bioinformatics tools in consensus as true positives [6–8] at the cost of losing sensitivity. This approach might not be appropriate in a clinical setting where the treatment of a patient relies on sensitive discovery of true somatic variants. In general, these bioinformatics tools identify SVs by using up to three different signals from aligned reads: (a) Read-depth information for inferring CNVs from non-uniform coverage in the regions, (b) discordant read-pairs that map with unexpected distance or orientation, and (c) split reads that have portions of a read mapping to different locations.

To deal with limitations of cWGS, recently "linked-reads sequencing" (10XWGS) technology was introduced. This utilizes reads derived from high molecular weight (HMW) DNA with typical fragment size between 50–100 kb in order to supply long-range information [7]. This approach captures HMW DNA molecules in so-called "Gel beads in Emulsion (GEM)". After encapsulation, HMW DNA is sheared into smaller fragments (0.5 kb), labelled with GEM specific barcodes and subjected to cWGS (2x150 bp). The attached barcodes link each short read-pair to its originating HMW DNA. The 10XWGS bioinformatics pipeline (Long Ranger) utilizes this information to reconstruct the initial long HMW DNA molecule. This also allows linking longer sections of the genome together into a phased haplotype and resolving SVs in low complexity regions of the genome. Theoretically, this should enable highly specific and sensitive prediction of SVs.

Several studies have recently used 10XWGS for molecular characterization of either large-sized SVs [8,9] or complex genomic rearrangements [10]. This is not limited to the normal human genome [11] but also feasible for different types of cancer and other diseases [12–14]. However, these studies predominantly use 10XWGS technology for orthogonal validation of SVs, but a comprehensive comparison of all SVs identified with 10XWGS and cWGS as an independent finding is currently not available.

Here, we performed an in-depth analysis of SVs from the MCF7 breast cancer cell line and a primary breast cancer sample. The goals of this study were: a) to evaluate and compare 10XWGS and cWGS technology for the prediction of different types and sizes of SVs; b) to identify an approach to predict highly specific SVs from both the technologies; c) to analyse GEM count as a predictor of true positive SVs. With this analysis, we also propose a statistical approach to determine highly specific and sensitive SVs amongst many false positive calls from both technologies that can also serve as a high confidence benchmarking set.

Materials and methods

Genomic DNA samples

The MCF7 breast cancer cell line was obtained from American Type Culture Collection (ATCC), Manassas, VA. Cells were maintained in EMEM medium with 0.01 mg/ml of insulin and 10% fetal bovine serum (FBS). The cells were incubated at 37˚C and in a 5% CO_2 humidified environment.

The primary tumor tissue was purchased from BioIVT (https://www.bioivt.com/) and was available as a fresh frozen sample. The sample is a triple negative breast cancer primary tissue with 50% tumor content based on histopathological examination. The data was analysed anonymously.

cWGS

DNA from MCF7 and the primary tumor sample was extracted with Qiagen's DNeasy blood and tissue kit (Qiagen, Hilden, Germany). Whole genome libraries for NGS were prepared by fragmenting 1 μg genomic DNA to achieve an average fragment size of 550 bp. Subsequently, the library was prepared using KAPA hyper prep kit (Roche, Basel, Switzerland) using 8 bp single-index NEXTflex DNA barcodes and sufficient library yield was achieved by 4 cycles of PCR. Leftover adaptors were removed with 1X bead purification performed with Agencourt AMPure XP beads (Beckman Coulter, Brea, USA). The Qubit dsDNA HS assay kit (Invitrogen, Carlsbad, USA) and Bioanalyzer high sensitivity DNA kit (Agilent Technologies, Santa Clara, USA) were used for quality control. The libraries were sequenced on Illumina's NovaSeq 6000 platform with S2 Reagent Kit for 300 cycles with a sequencing length of 2x150 bp (paired-end reads sequencing) with coverage as in S1 Table.

10XWGS

HMW genomic DNA was extracted from MCF7 and primary tumor tissue with MagAttract HMW DNA kit (Qiagen, Hilden, Germany). With 1 ng of HMW DNA, 10X Chromium reagents and gel beads library was prepared using the 10X Genomics Chromium genome reagent kit V2 user guide. Initial library construction takes place within droplets containing beads with unique barcodes. During library construction, a unique barcode (16 bp in length) is incorporated adjacent to Read-1. Final libraries were quantified on the Qubit using dsDNA HS assay kit (Invitrogen, Carlsbad, USA) and fragment length was determined using Bioanalyzer high sensitivity DNA kit (Agilent Technologies, Santa Clara, USA).

Prediction of SVs from cWGS

The Illumina paired-end reads were aligned to the GRCh38 reference genome using BWA-MEM (version 0.7.17) [15], duplicates were removed using Samblaster v0.1.24–0 [16] and alignment files were sorted using Samtools v1.3.1 [17]. We referred to two review studies [18,19] for the selection of tools for prediction of SVs from cWGS. An ensemble of tools was chosen for better sensitivity and specificity that utilized multiple sources of evidence like discordant read-pairs, split reads, read depth and local *de novo* assembly. Since there is no single ensemble of tools that outperforms other ensembles [18], we selected three tools based on their popularity, easy usability, prediction of all SV types that can also be predicted by 10XWGS tools and inclusion of an assembly based tool. This ensemble included Delly (v0.7.6) [20], Lumpy (v0.2.13) [21] and SvABA (v0.2.1) [22]. All these tools utilize discordant read-pairs and split-reads, while Delly also utilizes read-depth and SvABA utilizes local *de novo* assembly. After the predictions from all the tools, SVs of the same type (deletion, duplication, inversion and translocation), sharing the same orientation (3'to5', 5'to3', 3'to3' 5'to5') and breakpoints within a 500-bp window were merged as a single SV call. This window size was selected as short-fragment sequence analysis can confidently relate breakpoints that are within the median fragment size (~500 bp) [23]. The CNVs predicted only by read-depth methodology were not analysed here, as exact breakpoints necessary for further comparison could not be inferred. In order to maximize sensitivity we considered all high quality calls (predicted with filter "PASS") along with low quality calls (predicted without filter "PASS") from all the three tools. Moreover, to assess the confidence level of calls from cWGS pipeline, we allotted high confidence calls to the predictions that were predicted with filter "PASS" by at least one of the tools.

Prediction of SVs from 10XWGS

The sequenced linked-reads were analysed and processed using Long Ranger v2.2.2 wgs command with–somatic flag. The reads were aligned to the GRCh38 reference genome using Lariat and SNPs were predicted by freebayes v0.9.21-7-g7dd41db-dirty. The Long Ranger from 10X Genomics performs haplotype phasing and predicts SV after estimating a probability of barcode overlap between linked-reads and split reads for refining the breakpoints of rearrangements. The Long Ranger reports following types of SVs: deletion, duplication, inversion, translocation and some unresolved variants labelled as 'Unknown-UNK'. The CNVs predicted only by read-depth were not considered for analysis here. For a fair comparison with cWGS pipeline and to maximize sensitivity, we included two more tools utilizing linked-reads for prediction of SVs. The tool NAIBR v1.0 also performs haplotype phasing and constructs a probabilistic model to find novel adjacencies using discordant read-pairs and split barcoded molecules from linked-reads sequencing [24]. While GROC-SV v0.2.5 [25] utilizes a similar approach as Long Ranger additionally with local assembly at breakpoints using linked-reads. All the high quality calls (reported with filtered "PASS") and low quality calls (reported without filter "PASS") were considered for the comparison. The SVs from three tools were merged with the same scheme followed for intersection by cWGS pipeline. In order to estimate the confidence level of SVs from 10XWGS pipeline, each call was allotted high confidence when predicted with filter "PASS" by at least one of the tools.

Requantification of supporting reads for SVs

In order to evaluate the two technologies, we used an approach that quantifies the number of supporting reads for the SVs. The workflow (S1A Fig) involves construction of a synthetic genomic template from the sequence of reference genome. For SVs larger than 1 kb, a 1 kb template is constructed by retrieving 500 bp reference genome sequences to either side of the breakpoints, which are then fused according to the orientation of reported SV (S2 Fig). For SVs below 1 kb, the size of genomic template is reduced to atleast twice the size of SV. Next, short-reads are aligned to this synthetic genomic template with BWA-aln (version 0.7.17). From each SV alignment, we calculate the number of reads overlapping the fusion breakpoint for at least 15 bp (junction reads, JR) and read-pairs that span breakpoints (spanning pairs, SP). Only the reads with at least 70% of its bases aligning to the genomic template were considered for JR and SP. JR and SP were normalized as:

$$Normalized\ junction\ reads\ (JR) = \frac{Number\ of\ junction\ reads\ supporting\ SV}{Total\ number\ of\ reads} * 10^8 \tag{1}$$

$$Normalized\ spanning\ pairs\ (SP) = \frac{Number\ of\ spanning\ pairs\ supporting\ SV}{Total\ number\ of\ read - pairs} * 10^8 \tag{2}$$

$$Joint\ requantification\ support\ (JRS) = JR + SP \tag{3}$$

The requantification support was calculated from reads from both the technologies. Since, cWGS samples were sequenced at higher coverage than 10XWGS samples, we downsampled cWGS reads for calculation of requantification support. Moreover, read-1 from 10XWGS contains a 16 bp barcode sequence. Thus, for calculation of requantification support we trimmed the reads to a length of 125 bp, thereby removing the barcode. JR, SP and JRS were labelled with their sources as cWGS or 10XWGS.

GEM quantification for SVs

We also calculated the number of unique barcodes or GEMs containing read-pairs that support SVs reported from both the technologies. For this we used 10XWGS generated alignment file to first separate read-pairs that are aligned without a normal alignment FLAG. This was done using tool Samblaster v0.1.24–0 [16]. Next we counted number of unique barcodes or GEM (with BX tag in BAM file) that support a particular type and orientation of SV (S1B Fig). The unique GEMs were retrieved in the window w_i around breakpoints. The window size was selected as the ratio of average molecule length and N50 linked-reads per molecule from 10XWGS experiment. The GEM count was normalized as:

$$Normalized\ GEM\ count = \frac{Number\ of\ GEM\ supporting\ SV}{Total\ GEM\ detected} * 10^6 \tag{4}$$

Annotation of SVs and comparison from cWGS and 10XWGS

Each breakpoint of the SV was annotated with repeat region masked in RepeatMasker and poor mappability region [26]. In order to investigate the advantage of 10XWGS technology, we also calculated local coverage around the breakpoints in a window of size 400 bp for each SV. This was calculated using samtools pileup command and the local coverage was normalized by average coverage of the sequenced sample.

The SVs with size greater than 50 bp from both technologies were compared based on their breakpoint positions (within a window of 500 bp), type and orientation. As the 10XWGS pipeline reports inversions and duplications with size greater than 10 kb only, comparison was performed for those size ranges of inversions and duplications.

PCR confirmation of SVs

Some of the SVs that were common and uncommon between the technologies were selected for validation by PCR. We randomly selected a comparable number of candidate SVs from shared, 10XWGS only and cWGS only identified SVs. PCR primers were designed according to the predicted breakpoint spanning the junction site of the rearrangement with one primer positioned upstream and the corresponding primer downstream of the fusion. The genomic template for primer designing was produced according to the type and orientation of SV (S3 Fig).

Each PCR contained 10 ng sample DNA and primers with a final concentration of 0.333 μM each. The final volume was 30 μl using HotStarTaq Master Mix Kit (QIAGEN Cat. No. 203443) and 3 step-PCR with an annealing temperature of 60˚C for 40 cycles according to the manufacturer's recommendation.

Subsequently, the PCR products were analyzed on a QIAxcel capillary gel electrophoresis instrument using QIAxcel DNA Screening Kit (QIAGEN Cat. No. 929004). For alignment and size determination, a 15 bp / 500 bp marker (QIAGEN Cat.No. 929520) was used.

Sanger sequencing

To further confirm the PCR products, Sanger sequencing was performed in forward and reverse direction with primers used for the PCR. Samples were sent to Eurofins genomics (https://www.eurofinsgenomics.eu/) for sequencing.

Statistical analysis

All statistical tests were performed in R (version 3.6). The nonparametric Wilcoxon Rank sum test was used to compare positive and negative groups of PCR validated SVs. It was also used to compare local coverage around the breakpoints of SV derived from cWGS and 10XWGS alignments. While pairwise Kruskal-Wallis test was used to compare three groups of SVs: common SVs (predicted by both the technologies), only 10XWGS SVs (predicted only by 10XWGS) and only cWGS SVs (predicted only by cWGS).

Logistic regression model

Two logistic regression models were trained for filtering true positive calls from the cWGS and 10XWGS technology respectively. The features common between models were type of SVs (deletion, duplication, inversion and translocation), normalized junction reads (JR), spanning read-pairs (SP), size of the SV and local coverage around the positions. These were calculated from reads originating from the respective sequencing technology. Comparatively, the 10XWGS model also included GEM count as another feature. Only the SVs internally tested by PCR and predicted with respective technology were used for training and testing the model (for cWGS: Positive SVs = 178, Negative SVs = 75; and for 10XWGS: Positive SVs = 131, Negative SVs = 106). The respective data set was divided in 70:30 ratio as training and test data set.

The performance of models was measured on test data chosen with bootstrap resampling with 10 resamples (S17 Fig). Since the training data set for cWGS model was unbalanced, we also tested the performance of models with different type of sampling strategies (down sampling, up sampling and SMOTE). However, different samplings to balance the unbalanced data did not improve the performance of original cWGS model. Hence, we trained the cWGS model with unbalanced data only. Finally, we predicted true SVs as the ones predicted by either model with probability greater than 60%. The training of the classification model was carried out with the package caret in R v3.6 and importance of individual features was calculated with varImp function of caret. The varImp function calculates importance based on the absolute value of their t-statistics. The relative importance of features was calculated using dominance analysis [27] that derives importance of one feature over others by creating a subset of models with different combinations of features.

Results

cWGS and 10XWGS predict different numbers and classes of SVs

We compared cWGS and 10XWGS in terms of the numbers and classes of SVs predicted in two samples: a breast cancer cell line (MCF7) and a primary breast cancer sample. MCF7 and primary breast cancer sample was sequenced with 51X and 92X by cWGS technology. Their sequencing coverage was 17.4X and 17.7X respectively, by 10XWGS technology. The physical fragment coverage achieved by 10XWGS technology was 87X and 88.5X for MCF7 and primary breast cancer (nearly equivalent to average coverage of samples sequenced by cWGS) (S1 Table).

SVs were predicted by combining calls from an ensemble of three SV detection tools for cWGS data (SvABA, Delly and Lumpy) and three tools for 10XWGS data (Long Ranger, NAIBR, GROC-SV). The set of cWGS tools included Delly and Lumpy that use discordant read-pairs, split reads for detection of SVs and are widely accepted tools. Additionally, SvABA, a local assembly tool, was also included as Cameron *et. al.* [18] proposed an ensemble with a local de novo assembly tool for best performing collection of cWGS tools for SVs. Considering this, we created an ensemble of 10XWGS tools that use discordant read-pairs, split barcode molecules, barcode overlap and local de novo assembly. This included Long Ranger, GROC-SV and NAIBR. All the high and low quality SV calls from tools were considered and merged according to the type, orientation and their breakpoints. They are also referred to as high and low confidence calls respectively.

First, we investigated the different types of SVs identified by the cWGS and 10XWGS pipelines in both samples (Figs 1 and S4). There was significant difference in the number of different types of SVs predicted by the two pipelines (irrespective of high or low confidence calls). The ensemble of cWGS tools predicted comparatively higher number of all SV types (especially translocations). When looking in more detail into different size ranges, both the cWGS and 10XWGS pipelines identified deletion of all size range (S4E and S4F Fig) but the 10XWGS pipeline predicted nearly 5 times less deletions. The highest number of deletions in the cWGS pipeline came from low quality calls of SvABA while in the 10XWGS pipeline they came from high quality calls of Long Ranger (S4E and S4F Fig). Moreover, the 10XWGS pipeline predicted about 6 times less duplications in comparison to the cWGS pipeline when we consider both high and low confidence calls. This can also be attributed to the fact that tools in the 10XWGS pipeline predicted duplications with size>10 kb only (S4E and S4F Fig). However, tools in the cWGS pipeline predicted all sizes of duplications where most of them are low quality calls from SvABA and Delly (S5B and S6B Figs). Similar to the duplications, the 10XWGS pipeline predicted inversions greater than 10 kb only.

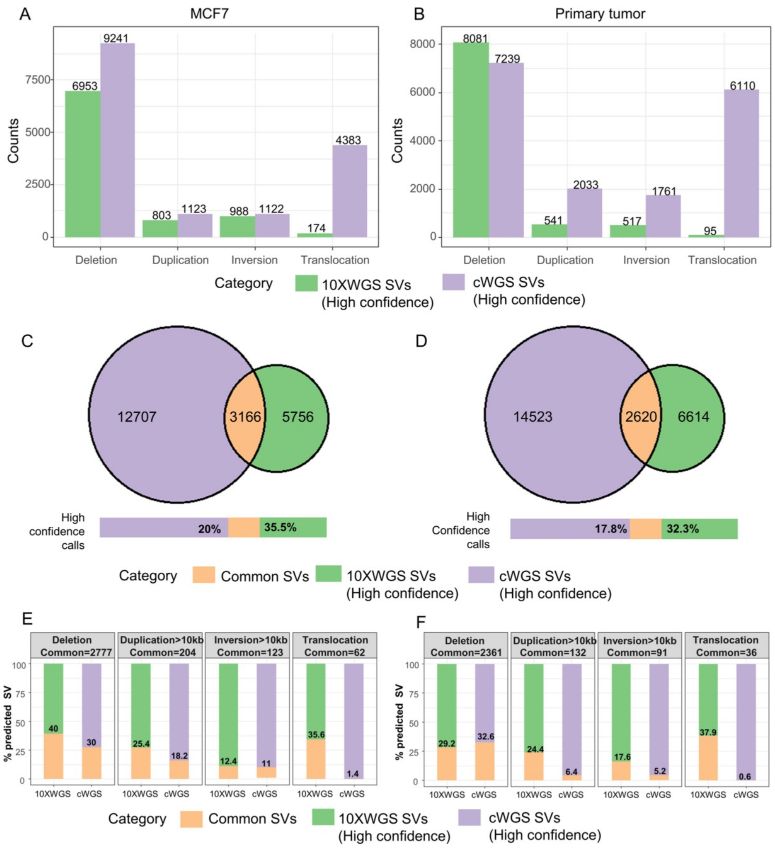


Fig 1. cWGS and 10XWGS predict a variable number of SVs with low proportion of common predictions. (A and B) Number of different types of SVs predicted with high confidence by cWGS and 10XWGS pipelines for (A) MCF7 and (B) primary breast tumor. (C and D) Number of high confidence SVs commonly predicted by both technologies for (C) MCF7 and (D) primary breast tumor. (E and F) Percentages of the indicated high confidence SVs commonly predicted by the two approaches for (E) MCF7 and (F) primary breast tumor.

https://doi.org/10.1371/journal.pcbi.1008397.g001

However, ~99% of inversions in the 10XWGS pipeline are predicted as low quality calls from Long Ranger that lie in the size range of 10–100 kb. This seems to be an attribute of Long Ranger methodology as other tools (NAIBR and GROC-SV) did not predicted such high number of inversion (S4, S5C and S6C Figs). The 10XWGS pipeline detected 100–200 fold fewer SVs with size >100 kb compared to the cWGS pipeline (S4E and S4F Fig). Since the 10XWGS pipeline generates long-range information from short-reads, it should be able to minimize alignment-based artefacts and therefore have a specificity advantage especially for those larger events.

The most remarkable difference in numbers was observed for translocations (Figs 1A, 1B, S4A and S4B). The cWGS pipeline predicted a much higher number of translocation in comparison to the 10XWGS pipeline. Majority of these translocations in the cWGS pipeline are contributed by low quality calls from SvABA and Delly (S5D and S6D Figs), which can be result of imprecise breakpoints, low mapping quality of reads, lower support in terms of discordant read-pars or split read etc. Moreover, as for other large SVs >100 kb from the 10XWGS pipeline, long-range information and low false discovery rate (FDR) translated into more precise number of translocations. Overall, the order of magnitude of predicted SVs is comparable between the cell line and the primary tumor sample, but the overlap is low.

Debarcoded and downsampled MCF7 SVs

Since the average genomic coverage of cWGS MCF7 sample was higher than 10XWGS MCF7, we tested SV prediction pipeline on downsampled cWGS reads (downsampled MCF7, equivalent genomic coverage as 10XWGS). We also tested a strategy to use cWGS tools with 10XWGS linked-reads. For this, barcodes in 10XWGS linked-reads were trimmed and the reads were processed in cWGS pipeline (debarcoded 10XWGS MCF7). It was observed in S7 Fig, the overall number of predicted SVs is reduced in the downsampled and debarcoded samples. This was especially true for the only cWGS predicted SVs (drops to ~50% and 70% respectively), while the number of common remained stable (~99.1% for debarcoded and 85.3% for downsampled samples). It is also evident from the debarcoded sample that allows analysis of exactly the same reads without linkage information in cWGS pipeline. However, the cWGS pipeline with debarcoded reads predicted very high number of small size SVs (size <1 kb, as seen in S7A Fig). This can be a ripple effect of reads from a different technology processed by algorithms designed for alternate technology. For further analysis, we decided to stick with the sequenced cWGS data sets whose genomic coverage matches physical coverage of the 10XWGS data.

A small fraction of predicted SVs is common to both cWGS and 10XWGS pipelines

We compared the calls between both technologies according to the breakpoints (within a window of ±500 bp), type and orientation of SVs: Fig 1C and 1D depicted the rather small overlap between both technologies for high confidence calls. This overlap was even smaller when low confidence calls were also considered in S4C and S4D Fig. Since we pool SV calls from multiple tools in both cWGS and 10XWGS pipelines, it is expected to have a high number of false positive predictions but higher true positive as well. However, this aggregation of the cWGS calls should result in high sensitivity and have rather higher overlap with 10XWGS calls. Contrastingly, the majority of high confidence 10XWGS calls do not overlap and only 35.5% and 32.3% of 10XWGS-predicted SVs were also predicted by the cWGS pipeline for MCF7 and the primary tumor, respectively. This raises the question of whether 10XWGS predicts SVs inaccessible by cWGS technology or whether the 10XWGS suffers from a high FDR. Or, vice versa, cWGS technology is more sensitive than 10XWGS, which misses many SVs.

There were differences with respect to different types of SVs (Fig 1E and 1F). Nearly 35.6% and 37.9% of high confidence translocations as predicted by 10XWGS were also predicted by cWGS from MCF7 and primary tumor respectively. The overlap increased slightly to 48.2% and 53.2% for MCF7 and primary tumor respectively, when low confidence calls were also considered (S4G and S4H Fig). Conversely, the percentage of common translocations by cWGS was extremely small (1.4% for MCF7 and 0.6% for primary tumor) due to the much higher number of predicted events. This implies that the cWGS pipeline is possibly sensitive, but has a very high FDR especially for translocations.

Additionally, we investigated whether high confidence calls by either pipeline are enriched among the common SVs. As depicted in S8A Fig, 41.1% and 35.1% of high confidence 10XWGS calls in MCF7 and primary tumor, respectively, were common between both the technologies. And, only 1.6% and 1.3% of low confidence 10XWGS calls were common in MCF7 and primary tumor, respectively. Comparatively, 20.4% and 15.5% of high confidence cWGS calls in MCF7 and primary tumor, respectively, were common between both the technologies. But, only 0.18% and 0.11% of low confidence cWGS calls were common in MCF7 and primary tumor, respectively. This indicates that common calls are high confidence calls from respective technologies. Moreover, 38.4% and 54.9% of calls predicted by all three tools in the cWGS dataset for MCF7 and the primary tumor (S8C and S8D Fig) were also predicted by 10XWGS. Comparably, all the calls predicted by all three tools in 10XWGS were predicted by cWGS pipeline. However, as depicted in S5, S6 and S8E Figs, very few calls were commonly predicted by all three tools in the 10XWGS pipeline. This is exemplified by the fact that 50% of common calls were predicted by all three tools in cWGS pipeline, while only 1.2% of common calls were predicted by all tools of the 10XWGS pipeline for MCF7.

Common SVs have higher read and GEM coverage

Since junction reads (JR), spanning pairs (SP) from both the technologies and unique barcodes (GEM) from linked-reads sequencing are the main cues for true SVs, we quantified them by a common computational approach for all identified SVs (Eqs 1, 2, 3 and 4). This allowed us to investigate differences in different categories of SVs: calls predicted by both the technologies (common SVs), calls predicted only by cWGS technology (only cWGS SVs) and calls predicted only by 10XWGS technology (only 10XWGS SVs). Common SVs had a significantly higher median count for JRS (median = 1.9) and GEM (median = 1.73) in comparison to only cWGS SVs (JRS: median = 0, GEM: median = 0) and only 10XWGS SVs (JRS: median = 0, GEM: median = 0) (Fig 2A and 2B). This inference was also drawn when different types of SVs were considered separately (Figs 2C, S9 and S10). Furthermore, since there might be differences in the libraries of the two technologies, we also calculated requantification support using 10XWGS reads. As depicted in S9–S12 Figs, we can draw same inference irrespective of the source of reads (cWGS or 10XWGS). Conclusively, regardless of the used technology and the used metric (JRS or GEM), common SVs were in all situations better supported.

Overall 63.5% of common SVs were supported by at least two JRS from cWGS data for MCF7. While 9.9% of only cWGS SVs and 6.6% of only 10XWGS SVs had at least a JRS of two from the respective technology. When high confidence calls were considered from the respective pipelines, 31.7% of only cWGS SVs and 14.6% of only 10XWGS SVs had at least a JRS support of two from their respective technology. It is surprising to note that the only cWGS SVs also had support from 10XWGS linked-reads: 30.4% of only cWGS high confidence calls were also supported with at least a JRS of two calculated from 10XWGS linked-reads. Comparatively, only 10.8% of only 10XWGS high confidence calls had at least a JRS of two from cWGS data. It is somehow expected that each technology gives overall higher support to the SVs identified by them. However, we observed that a higher fraction of high confidence SVs only predicted by cWGS still had higher requantification support in comparison to the ones predicted only by 10XWGS. This implies that many of the SVs predicted only by the cWGS pipeline do have evidence in the 10XWGS sequenced data (overlapping GEMs, JRs and SPs) but the 10XWGS tools did not identify them (Figs 2A, 2B and S11). Vice versa, high confidence SVs predicted only by 10XWGS have overall lower support from both the technologies. The same observations that are described here for MCF7 were also made for the primary tumor sample (S10 and S12 Figs). This data indicated that common events are most likely enriched for true positive events. Nevertheless, additional true positive events are contained in only cWGS SVs while only 10XWGS SVs contributes a lower number of true SVs.

To further characterize differences between both sequencing technologies, we annotated each breakpoint of the SVs for repetitive regions and ambiguous mappability regions. It is well established that short-reads originating from repetitive regions are often misaligned [3]. Considering the breakpoints of high confidence SVs from both pipelines in Fig 2D, it was observed that breakpoints of 57.2% common SVs and 54.3% only 10XWGS SVs are inside a repetitive region with majority being in SINE and LINE (S13B Fig). However, for only cWGS SVs, 71.8% of the breakpoints were inside repeats where satellite and simple repeats contributed

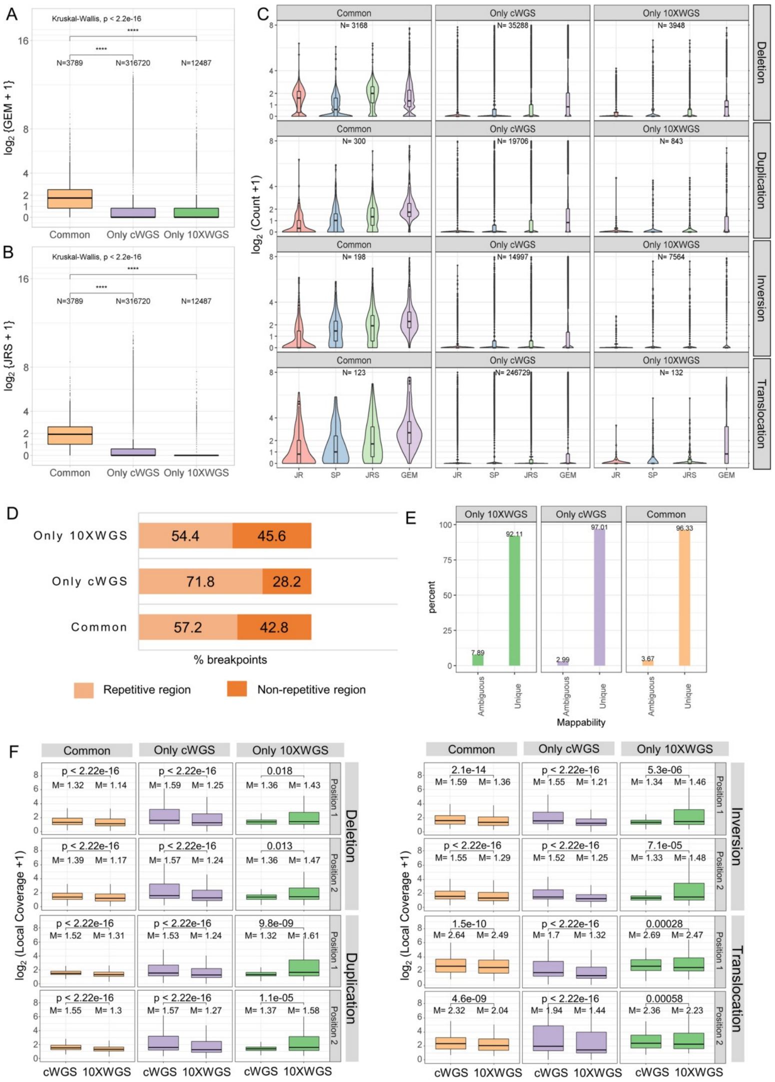


Fig 2. Requantification support and GEM coverage for SVs common between cWGS and 10XWGS is higher than that predicted by a single technology. (A) Distribution of GEMs containing SVs that were predicted by both the technologies (common) or only by one technology (only cWGS or only 10XWGS) for MCF7. (B) Shown is the combined requantification support (JRS) as the sum of junction and spanning reads from cWGS data for common SVs and SVs predicted only by cWGS or 10XWGS for MCF7. p-values were calculated using Kruskal-wallis test and pairwise Wilcoxon rank sum test. **** represents a p-value <0.0001. (C) Comparison of requantification support (Junction reads-JR, Spanning pairs-SP, JRS = JR+SP) and GEMs for different type of SVs that are common between technologies and only predicted by 10XWGS or cWGS for MCF7. The black lines in the boxes represent median (centre line), upper quartile (upper line) and lower quartile (lower line), respectively. The area of violin plots is scaled to the number of observations. (D) Percentage of breakpoints of high confidence SVs from two technologies covered by repetitive regions. (E) Percentage of breakpoints of high confidence SVs from two technologies covered by unique mappability regions. (F) Distribution of normalized local coverage around the positions of high confidence SVs (size >10 kb), calculated from cWGS and 10XWGS aligned reads respectively. p-values were calculated by pairwise Wilcoxon rank sum test and 'M' is median of normalized local coverage.

https://doi.org/10.1371/journal.pcbi.1008397.g002

towards 49% of the breakpoints. This indicates that a high fraction of these calls may be false positive calls due to misalignment. Secondly, when considered all the SVs (both high and low confidence ones), the percentage of breakpoints in ambiguous mappability regions were higher for only cWGS SVs than only 10XWGS SVs (S13C Fig). When only high confidence calls were considered in Fig 2E, more than 90% of breakpoints were in unique mappability regions. Overall, cWGS and 10XWGS technology contributed fewer SVs with breakpoints in low complexity and LTR regions, while SVs with breakpoints in SINE and LINE elements were common in both.

The 10XWGS technology links short-reads to their larger size DNA fragment and is assumed to improve local physical coverage of SV breakpoints. Thus, we compared the normalized local coverage derived from both cWGS and 10XWGS aligned reads for all SVs greater than 10kb. When we considered all SVs (both high and low confidence calls), we did not observe a significant difference in local coverage for 10XWGS only calls between the two technologies (except in inversions) (S13D Fig). However, in Fig 2F we considered only the high confidence calls and had shown that the local coverage in only 10XWGS SVs is higher when 10XWGS aligned reads were considered (except in translocations). Moreover, common and only cWGS calls had higher local coverage from cWGS aligned reads. This indicates that prediction of additional SVs from 10XWGS might indeed be the result of improved coverage, with these SVs missed by cWGS sequencing.

PCR confirms high specificity of common SVs

We validated a comparable number of randomly selected common and uncommon SVs from the three categories: 135 common SVs, 118 only cWGS SVs and 102 only 10XWGS SVs (S2 Table). The orthogonal validation was performed with PCR and Sanger sequencing of SVs from MCF7. Fig 3A exemplifies the PCR validation results for seven SVs: Five SVs led to amplification of a product of expected size and were therefore determined as positive. Additionally we selected a subset of positive amplicons for Sanger sequencing for confirmation of the sequence across the breakpoint, as depicted in Fig 3A. In total, we confirmed 36 out of 42 amplicons by Sanger sequencing. The remaining six amplicons had poor quality sequence traces and could not be analysed.

The pie charts in Fig 3B illustrated the confirmation rate for SVs from the respective categories. 89% of common SVs were confirmed by PCR. This indicated that the combined approach of 10XWGS and cWGS is highly specific for the prediction of SVs. Only 15 common SVs were not confirmed by PCR. We followed these up in detail by manual inspection of the sequence alignment from cWGS data. Here, we observed that either the breakpoints were in repetitive

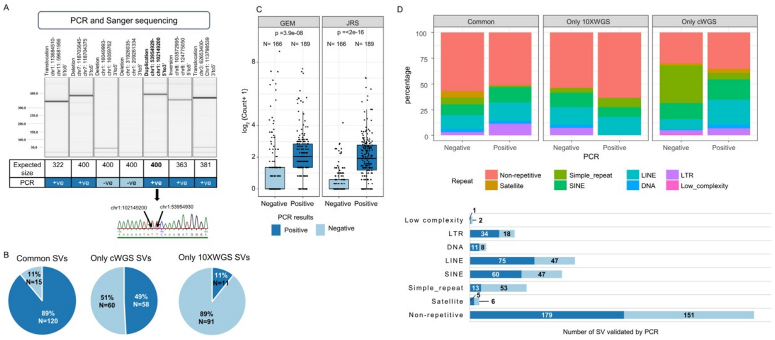


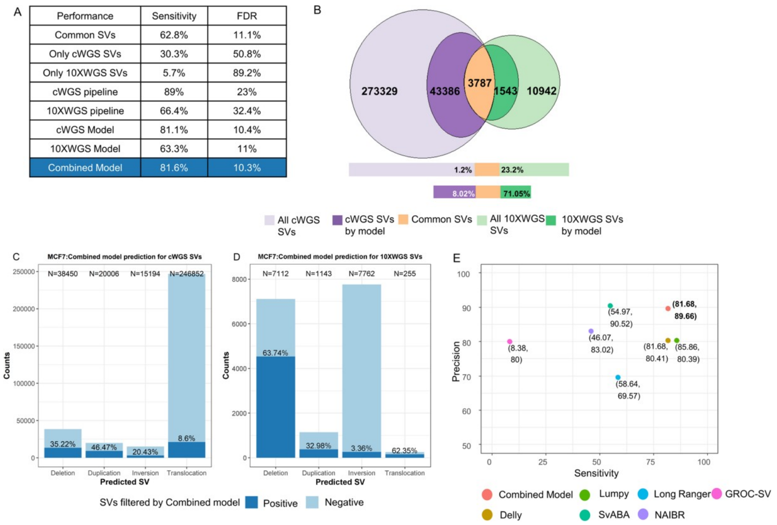
Fig 3. Orthogonal validation of SVs using PCR and Sanger sequencing. (A) SVs within the MCF7 dataset were selected for validation by PCR and Sanger sequencing. From the PCR-amplified products, a subset was further confirmed by Sanger sequencing. Shown are representative results involving seven SVs. (B) Number and percentage of PCR-validated SVs for the three categories: SVs common between cWGS and 10XWGS (common SVs), SVs only predicted by cWGS pipeline (only cWGS SVs) and SVs only predicted by 10XWGS pipeline (only 10XWGS SVs) are shown. (C) The difference in normalized counts of combined requantification support (JRS from cWGS reads) and GEM for PCR-validated SVs is shown. Each data point represents counts for PCR tested SVs and box-and-whisker plots represent lower quartile, median and upper quartile. p-values were derived from Wilcoxon rank sum test. (D) Percentage and number of repetitive element classes in PCR validated SVs for three categories: common, only cWGS and only 10XWGS SVs.

https://doi.org/10.1371/journal.pcbi.1008397.g003

regions, SVs lacked proper read support, reference genome region was not annotated or the SV events shared the same breakpoint i.e. they were complex in nature (S14 Fig). In contrast, the confirmation rate for SVs only predicted by cWGS and 10XWGS dropped to 49% and 11% respectively. This confirms that the 10XWGS pipeline is prone to prediction of false positive SVs. We further investigated the PCR validation rate for SVs that are an overlap between tools from respective pipelines. S15 Fig shows that cWGS SVs predicted by the consensus of all tools have a maximum PCR confirmation rate (i.e. 84.4%). This is in agreement with the popular approach of considering consensus SV calls from multiple tools to reduce false positive calls by cWGS technology. Similarly, consensus predictions from the 10XWGS pipeline had 84% confirmation rate. The confirmation rate for consensus deletions and duplications by 10XWGS was 100% and 60% respectively. However, confirmation rate for duplications predicted by two tools of the 10XWGS pipeline was higher at 84.2%. A similar trend of most confirmation rates for calls predicted by all three tools of the 10XWGS pipeline was followed for inversions (75%) and translocations (100%) and also by all the SV types in cWGS pipeline.

In order to confirm that requantification support and GEM counts can serve as a metric to filter out true positive SVs, we plotted their counts for PCR-tested SVs in Fig 3C. The PCR-positive SVs had significantly higher requantification support (JRS) and GEM coverage in comparison to ones that are tested PCR-negative. This was also true for requantification support calculated using 10XWGS reads (S16 Fig). Moreover, we compared the confirmation rate for PCR validated SVs with respect to the repeat class of breakpoints in Fig 3D. It was observed that validation rate for SVs in simple repeats was lower, while differences in validation rates for other classes could not be derived. Moreover, a higher percentage of SVs only predicted by cWGS in simple repeats could not be confirmed by PCR. As expected, this indicates that cWGS pipeline cannot resolve SVs in simple repeats.

For a direct comparison of these two technologies, we calculated the sensitivity and FDR using PCR-tested SVs in Fig 4A. The SVs predicted by both technologies had 62.8% sensitivity



Performance	Sensitivity	FDR
Common SVs	62.8%	11.1%
Only cWGS SVs	30.3%	50.8%
Only 10XWGS SVs	5.7%	89.2%
cWGS pipeline	89%	23%
10XWGS pipeline	66.4%	32.4%
cWGS Model	81.1%	10.4%
10XWGS Model	63.3%	11%
Combined Model	81.6%	10.3%

Fig 4. Prediction of SVs by trained models for the cWGS and 10XWGS technology. Two logistic regression models were trained on PCR tested SVs from the respective technologies. (A) The table depicts the performance of different categories of SVs or technologies derived from PCR tested SVs. (B) Numbers and percentage of SVs common between the technologies before (lighter shades) and after (darker shades) applying the respective trained models. (C) Number of SVs predicted by the cWGS technology within the MCF7, and percentage predicted positive by the combined models. (D) Number of SVs predicted by the 10XWGS technology within the MCF7, and percentage predicted positive by the combined models. (E) Plot for performance of combined model and all other tools on internally validated SVs.

https://doi.org/10.1371/journal.pcbi.1008397.g004

with a very low FDR of 11.1%. However, SVs only predicted by one of the technologies had much higher FDR. Overall the cWGS pipeline had high sensitivity (89%) but with a high FDR of 23%. Comparatively, the 10XWGS pipeline had lower sensitivity (66.4%) with an even higher FDR of 32.4%. This indicated that even the 10XWGS pipeline is prone to high FDR and requires more stringent filtering criteria to further enrich true positive SVs.

Enrichment of true positive SV calls using requantification support and GEM count

The data indicates that for a highly specific and sensitive prediction of SVs a combined approach using cWGS and 10XWGS prediction data might be advisable. Nonetheless, we created prediction models for both the technologies independently to improve the prediction as much as possible for situations when only data from one of the technologies is available. Additionally, we combined all predictions into a unified approach to offer best sensitivity and FDR when both analyses are available. Initially, we also tested a simple filtering approach based on the number of supporting reads to enrich for true positive events, but observed poor sensitivity as there is no clear separation between PCR positive and negative SVs (Fig 3C).

To this end, we generated two logistic regression models using PCR validated data, one for the 10XWGS data and a second one for the cWGS data. In Fig 4A, we measured sensitivity and FDR of both the models based on PCR tested SVs. It was evident that FDR reduces drastically after applying the trained models. Predictions from the cWGS model and the 10XWGS model show a reduced FDR from 23% to 10.4% and from 32.4% to 11%, respectively. However, this came at the cost of reduced sensitivity, which decreased from 89% to 81.1% for the cWGS model and from 66.4% to 63.3% for the 10XWGS model. Moreover, SVs predicted by both technologies had a very low FDR but with sensitivity lower than for the overall cWGS pipeline (as shown by PCR). Application of both models increased the percentage of SVs common between both the technologies from 1.2% to 8.02% for cWGS and 23.2% to 71.05% for 10XWGS in MCF7 (Fig 4B). This is another evidence for the decrease in FDR achieved by both the models. A similar increase in overlap was also seen in an independent primary tumor sample (S18 Fig).

All three approaches (common SVs, cWGS model and 10XWGS model) aim to enrich different subset of true positive SVs. We therefore considered all these calls in a combined model for best sensitivity and low FDR and tested its performance on PCR validated SVs. To this end, we observed a reduced FDR to 10.3% and a high sensitivity of 81.6% similar to the cWGS model (Fig 4A). Application of the integrated approach made a dramatic difference on the overall landscape of predicted SVs from cWGS and 10XWGS (Figs 4C, 4D and S18): The combined model filtered out 85.3% and 86.9% of total calls in MCF7 and primary tumor respectively. Moreover, the most significant reduction in MCF7 was observed for translocations from cWGS where we observed a reduction to 8.6% of total calls. In case of the 10XWGS technology, we observed a maximum reduction of inversions to 3.36%.

Overall, the combined model gathered good sensitivity and precision for overall performance against the other tools (Fig 4E), for internal PCR validated SVs. The combined model achieved 81.68% sensitivity and 89.66% precision on the full MCF7 sample. Comparatively, only Delly and Lumpy had comparable sensitivity of 81.68% and 85.85% respectively. However, their precision was around 9% lower than for the combined model. SvABA had shown slightly superior precision with 90.52%, but at the cost of much lower sensitivity (54.97%). Therefore, the combined model offered best overall performance tradeoff in terms of sensitivity and precision. Compared to the 10XWGS tools the advantage was even more apparent. The combined model also greatly reduced cWGS only calls predicted in simple repeat and satellite regions (compare S19 to S13 Fig). Therefore, 10XWGS only calls contained a higher fraction of SVs in simple repeat regions. This is in-line with the notion that 10XWGS offers superior performance in these low complexity regions due to use of long range information. Of note, even when only cWGS or 10XWGS data is used, our established models can still compete well with the other tools of the respective technology. Moreover, we also compared the performance of the combined model against other tools, when the reads were downsampled or debarcoded. As depicted in S20 Fig, results on downsampled and debarcoded datasets had shown decreased sensitivity for all tools, but are otherwise very comparable.

Benchmarking combined model

We tested the performance of combined logistic regression model also on previously validated SVs in MCF7. A list was gathered from Li *et. al* [28] (external study 1) that included 183 SVs of size greater than 500 kb. These calls were detected by the tool Weaver and confirmed with optical mapping. Another set of 70 validated SVs was collected from Hillmer *et. al* [29] (external study 2) that was detected by a long-span paired-end-tag sequencing approach and was validated by PCR. We also benchmarked the model with germline SV calls as available in gnomAD study to confirm shared germline events present in MCF7 [30].

On the external study 1 data set, the combined model achieved sensitivity of 76.69% which was lower than Delly (94.54%) and Lumpy (95.06%) (S20 Fig). However, it was superior in terms of sensitivity to SvABA (72.13%), Long Ranger (26.23%), NAIBR (34.97%) and GROC-SV (10.93%). Here, the results differ from our own data, but this study only contains large structural variants and therefore offers insights into this subset of SVs only. For the external study 2 calls, the combined model achieved a sensitivity of 84.29%, which is comparable to Delly (85.71%) and Lumpy (84.29%). However, it was superior in terms of sensitivity to SvABA (58.57%), Long Ranger (62.86%), NAIBR (70%) and GROC-SV (11.43%). For this data set we observed similar sensitivities to our data set. When considering the germline SVs from gnomAD study as another set of validation, the calculated sensitivity was very small as only a small subset of known germ line SVs is expected in in MCF7 cell line (S20A, S20B and S20C Fig). Nevertheless, the combined model achieved better sensitivity in comparison to all other tools. When considering all gnomAD germline SVs present in MCF7, the combined model maintains good sensitivity compared to all unfiltered predictions (2629/3076 ~ 85.47%; S20F Fig). When we look at SV predictions with downsampled and debarcoded reads, then the combined model consistently performed better than all the tools (S20B, S20C, S20D and S20E Fig). This shows the robustness of the combined model for even lower genomic coverage samples. When calculating precision based on these external datasets, we observed artificially poor values for our combined model (S3 Table). However, these datasets only partially reflect the entire range of SVs (e.g. limited size range, only germ line SVs). Therefore, any general approach towards SV prediction will perform poor in such an analysis.

Taken together, the here presented logistic regression model provides a sensitive and accurate filter to predict true positive SVs. The model can also be utilized for reads from only one technology (cWGS or 10XWGS), but of course, at the cost of reduced sensitivity.

Discussion

Structural variations can have diverse functional impacts in humans; therefore, when performing genomic analysis of any disease state, it is imperative to find true positive SVs that might be associated with a certain phenotype. A popular approach to identify SVs is the cWGS technology, which suffers from high FDR (up to 85%) and varying sensitivity (30–70%) [31–33]. Here, we aimed to boost sensitivity for SV detection by integration of multiple bioinformatics tools, which is a common practise utilized in many studies [19,33,34]. Typically this comes at the cost of high FDRs. In order to reduce the FDR, many studies consider only the consensus from multiple bioinformatics tools [19,33,34]. In our analysis, we could show that the focus on SVs that are found by multiple tools can indeed achieve low FDR, but at the cost of much reduced sensitivity. This shows that better approaches are needed to enrich true positive SVs in such scenarios.

More recently, the development of 10XWGS technology seem to offer an elegant solution by taking into account long-range mapping information for the prediction of SVs. Our validation data had shown a relatively high FDR of 10XWGS for SVs which is improved when only high confidence calls are considered. However, compared to cWGS sequencing, 10XWGS had lower sensitivity when considering all types of SVs. This is in line with previous studies that reported varying sensitivity of 35–88.4% and moderate FDR of 50% for the 10XWGS technology [10,35]. Since 10XWGS is the latest technology, there are currently fewer algorithms available for the analysis of data. Nevertheless, we compared the performance of set of those algorithms against cWGS tools here. Contrary to previous studies, where performance metrics were derived from publically available datasets that are limited in type and size of SVs and are derived from diploid genomes, we presented a comprehensive analysis of all types of SVs in a cancer cell line and a tumor sample. Of note, sensitivity was here analysed with regard to all identified and confirmed SVs. However, true sensitivity may be lower, because additional SVs might exist that are neither detected by cWGS nor 10XWGS sequencing.

The reduced sensitivity in 10XWGS data raised a question whether it was a limitation of the analysis pipeline (ensemble of 10XWGS tools) or the technology did not cover the affected genomic regions. Interestingly, we observed that SVs, which were not identified by 10XWGS tools, did have support in the aligned linked-reads (i.e. overlapping GEM, JR and SP). We further analysed this by removal of barcodes in linked-reads and processed it with classical cWGS prediction tools. With the debarcoded sample, we were able to identify additional SVs that were missed by 10XWGS specific tools. This indicates that additional information is present in the raw 10XWGS sequencing data that is not fully utilized by currently available tools. Although the existing 10XWGS tools use similar category of evidence as cWGS tools (discordant read-pairs, split molecules, de novo assembly) apart from GEM coverage, they, however, seem to miss many true calls.

Previously, studies have shown that 10XWGS technology was especially useful in identifying complex genomic rearrangements or chained SVs [10]. Here we did not specifically address this subset of SVs, as we were interested in the overall performance of SV prediction. Nonetheless, the added benefit of 10XWGS sequencing becomes visible when looking at large SVs and translocations. This class of SVs is particularly difficult to resolve by the cWGS technology and suffers from high FDRs [36]. Utilization of long-range information by the 10XWGS pipeline should be powerful in resolving them. This was demonstrated by the fact that the 10XWGS pipeline reported a much lower and much more plausible number of translocations in comparison to the cWGS pipeline. We also observed for translocations the highest overlap (~48–53%) of the 10XWGS predictions with the cWGS pipeline that were all confirmed by PCR. However, only 65% of all high confidence translocations from the 10XWGS pipeline were confirmed by PCR. This suggests that not all translocations predicted by the 10XWGS pipeline are true events or are chained SVs. On the other hand, we were also able to confirm translocations reported only by the cWGS pipeline that were missed by the 10XWGS pipeline. Nevertheless, the 10XWGS pipeline was superior in predicting translocations in comparison to the cWGS pipeline.

The performance of cWGS technology suffers from high FDRs in low mappability and low complexity regions, such as simple repeats and LTRs [18], while the performance has previously been shown to be unaffected by SINE, LINE and DNA elements in the genome. In line with that, we identified a higher fraction of SVs in repetitive regions for cWGS technology compared to 10XWGS, especially in microsatellite, simple repeat and SINE elements. Furthermore, we observed a lower validation success rate for these SVs, demonstrating that a high fraction of predicted SVs in those regions are potentially false positive. Utilization of the long range information provided by 10XWGS seems to be able to greatly reduce these false positive predictions as indicated by a much smaller fraction of predicted SVs in those regions.

For both technologies we identified only a small fraction of SVs in regions with an ambiguous mapping of reads. Nonetheless, the fraction of SVs only identified with 10XWGS in such regions was more than double in comparison to cWGS. Moreover, 10XWGS technology did improve local coverage around breakpoints for SVs that were missed by cWGS pipeline. With the exception of translocations, all other type of large size SVs (size >10 kb) that were only identified by 10XWGS had significantly higher median local coverage around breakpoints from 10XWGS technology than cWGS. This indicated that the long range information utilized by 10XWGS allows improved mapping and coverage to those regions and improved subsequent identification of SVs.

Taken together, 10XWGS enabled more accurate detection of translocations and of SVs in low complexity regions. However, when all predicted SVs were considered, an improved detection on this subset does not translate into an overall improved FDR or sensitivity. This is also corroborated by other studies [10,33]. Our data had shown that this is largely due to methodology issues, demonstrating that the relatively new 10XWGS technology needs to catch up with methodological advancements from cWGS prediction tools.

Previous studies have also used a combination of cWGS and 10XWGS to predict SVs where 10XWGS data was often used as an orthogonal validation set. Confirming SVs predicted from cWGS technology with 10XWGS technology would lead to highly specific SVs, as we could confirm here by PCR. However, this comes also at the cost of missing a considerable fraction of true events.

Here we proposed an integrated statistical approach using both the technologies to achieve optimized FDR and sensitivity for all types of SVs. We tested the combined model on an exhaustive set of internally validated SVs and two externally validated data sets. We observed lower FDRs in comparison to FDRs of both technologies, however at the cost of minimal loss in sensitivity. The model efficiently combined different features as requantification, GEM support, type and size of SVs and local coverage around breakpoints. However, one limitation of this model would be for application in detection of chained SVs. Those events would have partial or no support from requantification pipeline. Nevertheless, it outperforms other tools for simple SVs and even a simple heuristic filter for the read support. We could also show the robustness of model with downsampled and debarcoded reads.

Another limitation of such an integrated approach is the requirement to run two sequencing experiments for each sample. Therefore, we generated models based on 10XWGS and cWGS pipeline independently. The overall performance of model was superior compared to the individual tools for the respective technologies. The individual models for cWGS and 10XWGS enables their usage when only one technology is available. This is of particular relevance for the 10XWGS data as our model provides a very prominent improvement in performance compared to the three tested 10XWGS tools. However, without cWGS data, a gap in sensitivity is evident. The debarcoding of 10XWGS data and its subsequent analysis with cWGS pipeline could provide an opportunity to boost sensitivity to almost the same level.

We also investigated shared germline SVs present in the gnomAD database. The fraction of MCF7 SVs present in gnomAD was low. However, individual or low frequency germline SVs of the respective samples are not covered by this analysis. Only the analysis of a matched sample would enable clear separation of germline and somatic SVs. Nonetheless we observed best sensitivity for known germline SVs with the combined model, indicating that these can be predicted with similar high sensitivity.

The sensitivities observed in our internally validated data set and existing datasets confirms this claim. Convincingly, the hereby used logistic regression approach with unique set of features opens up a broader application of the model.

Conclusively, our analysis for true SV events could show that specific and sensitive prediction of SVs is possible, but requires an integrative approach for best results. We could show that 10XWGS predicted SVs could be used for orthogonal validation but considering only those calls would miss many true events. Our combined model approach takes into account all the available data points to maintain high sensitivity and low FDR. Sensitive identification of SVs is necessary to get a complete picture of the mutational landscape in cancer and gain a better understanding of the disease. Additionally, the complex nature of many hereditary and genetic diseases could be resolved with reliable and sensitive prediction of SVs. Thus, we believe that the presented integrated prediction approach is a valuable tool that may identify novel targets for disease treatment.

Supporting information

S1 Fig. Workflow for calculation of requantification support with short-reads and GEM coverage for SVs. (A) Workflow to requantify supporting short-reads for SV. The reference genome sequence around the breakpoints A and B are extracted and fused according to the type and orientation of SVs. The short-reads are aligned to this fused genomic template. Junction reads (JR) and Spanning pairs (SP) are counted as requantification support. (B) Workflow to quantify unique GEMs or barcodes containing read-pairs that support a particular type and orientation of SV. First, discordant read pairs or split reads are retrieved from the 10XWGS pipeline generated alignment file. Then, unique GEMs are counted that support a particular SV type and orientation with breakpoints in window wi.
(TIF)

S2 Fig. Construction of synthetic genomic template from the reference genome for calculation of requantification support. Illustration of the procedure to extract the reference genome sequence around the SV breakpoints that are fused to generate 1kb genomic templates. The fusion of genomic sequence around the breakpoints of SVs is performed according to the type of SV and the respective orientation (deletion-3'to5', duplication-5'to3', inversion fusion1-3'to3', inversion fusion2-5'to5'). The same strategy is followed for translocation with the difference that the regions extracted belong to different chromosomes.
(TIF)

S3 Fig. PCR primer design for different types of SVs.The left primer (LP) and right primer (RP) were designed at least 100bp up- and downstream the predicted breakpoints and were designed based upon the amplicon template formed according to the structural variation (deletion, duplication, inversion, translocation) and its orientation (3'to5', 5'to3', 3'to3' and 5'to5').
(TIF)

S4 Fig. SV type, sizes distribution of SVs predicted by cWGS and 10XWGS technology and percentage of common SVs amongst them. (A), (B) Number of different type of SVs predicted by two technologies in MCF7 and Primary tumor respectively. (C), (D) Percentage of high and low confidence calls overlapping between technologies for MCF7 and Primary tumor respectively. (E), (F) Distribution of size of different type of SVs from both the technologies in

MCF7 Primary tumor respectively. (G), (H) Percentage of different SV types predicted by both the technologies in MCF7 and Primary tumor respectively.
(TIF)

S5 Fig. Distribution of all SV calls from all cWGS and 10XWGS tools and their overlap, in MCF7 sequenced sample. (A) SV calls for all deletion, (B) duplication, (C) inversion, and (D) translocation. Low confidence calls are marked by "LowQ" and high confidence calls are marked by "PASS".
(TIF)

S6 Fig. Distribution of all SV calls from all cWGS and 10XWGS tools and their overlap, in Primary tumor. (A) SV calls for all deletions, (B) duplications, (C) inversions, and (D) translocations. Low confidence calls are marked by "LowQ" and high confidence calls are marked by "PASS".
(TIF)

S7 Fig. Distribution of SV calls (both high and low confidence) from cWGS sequenced MCF7, 10XWGS sequenced MCF7, downsampled cWGS reads in MCF7 to equivalent coverage as 10XWGS MCF7 (downsampled cWGS), removal of barcodes in 10XWGS linked-reads and processing them through cWGS tools (debarcoded 10XWGS). (A) The size distribution of different SV types for all mentioned samples, (B) Number of calls commonly predicted by 10XWGS, sequenced cWGS and downsampled cWGS; and number of calls commonly predicted by 10XWGS, sequenced cWGS and debarcoded 10XWGS (NOTE: Some of the SV calls from sequenced cWGS overlaps with multiple debarcoded 10XWGS and downsampled cWGS calls), (C) Number of SV calls processed from all mentioned samples (considering all SVs except duplications and inversions of size>10kb).
(TIF)

S8 Fig. Only a small fraction of SVs overlap between the 10XWGS and cWGS predictions. (A), (B) Percentage of high and low confidence SVs from cWGS and 10XWGS pipeline that are common between technologies, in MCF7 and Primary tumor respectively. (C), (D) Percentage of 1 tool, 2 tools, 3 tools SVs from cWGS and 10XWGS pipeline common between the technologies, in MCF7 and Primay tumor respectively. (E) Number and percentage of common SV between two technologies that are predicted by 1 tool, 2 tools and 3 tools.
(TIF)

S9 Fig. Common SVs have significantly higher support in terms of requantification (Sample = MCF7). Different requantification support (junction reads-JR, spanning pairs-SP, JR+SP = JRS) and GEM count plotted for common SVs, only cWGS SVs and only 10XWGS SVs. The requantification support was calculated from two sources of reads (cWGS and 10XWGS). p-value calculated with Kruskal-wallis test for comparison of three categories and pairwise Wilcoxon rank sum test. **** represents p-value <0.0001.
(TIF)

S10 Fig. Common SVs have significantly higher support in terms of requantification (Sample = Primary tumor). Different requantification support (junction reads-JR, spanning pairs-SP, JR+SP = JRS) and GEM count plotted for common SVs, only cWGS SVs and only 10XWGS SVs. The requantification support was calculated from two sources of reads (cWGS and 10XWGS). p-value calculated with Kruskal-wallis test for comparison of three categories and pairwise Wilcoxon rank sum test. **** represents p-value <0.0001.
(TIF)

S11 Fig. Requantification support and GEM count is higher for common SVs for different types of SVs (Sample = MCF7) for all calls or only high-confidence calls. The plot of three categories of SVs (common, only cWGS and only 10XWGS) and different type of SVs with respect to requantification support and GEM count. Requantification count was calculated from cWGS reads and 10XWGS reads separately. Junction reads (JR), spanning pairs (SP), combined support (JRS = JR+SP). 'N' represents the total number of SVs in the particular category.
(TIF)

S12 Fig. Requantification support and GEM count is higher for common SVs for different types of SVs (Sample = Primary Tumor) for all calls or only high-confidence calls. The plot of three categories of SVs (common, only cWGS and only 10XWGS) and different type of SVs with respect to requantification support and GEM count. Requantification count was calculated from cWGS reads and 10XWGS reads separately. Junction reads (JR), spanning pairs (SP), combined support (JRS = JR+SP). 'N' represents the total number of SVs in the particular category.
(TIF)

S13 Fig. Annotation of breakpoints of SVs shared between technologies indicate the advantage of each technology. (A) Breakpoints of all SVs (both high and low confidence) annotated with repetitive regions and their percentage across categories of SVs: common, only cWGS and only 10XWGS SVs. (B) Breakpoints of only high confidence SVs annotated with repetitive regions and their percentage across common SVs, only cWGS SVs and only 10XWGS SVs. (C) Breakpoints of all the SVs (both high and low confidence calls) annotated with unique mappability regions. (D) Normalized local coverage across two positions of each SV event in cWGS and 10XWGS aligned reads. All these figures depict annotation of breakpoints in MCF7 sample.
(TIF)

S14 Fig. Some of the SVs common between technologies were not validated by PCR as their breakpoints lie in repetitive region, poor mappability region or when the reference genome was not annotated (Sample = MCF7). The table describes the possible reason for common SV calls that were not validated by PCR. Alignment of cWGS reads against reference genome for some negatively validated common SVs are shown in the form of IGV images.
(TIF)

S15 Fig. Validation rate for SVs shared between all tools is higher for cWGS (A, B & C) and 10XWGS technology (D, E & F)-Sample MCF7. (A) Ratio of PCR validated SVs from the cWGS technology that were predicted by 1, 2 or 3 tools. (B) Ratio of different type of SVs from cWGS technology validated by PCR. (C) Ratio of different type of SVs validated by PCR with respect to prediction by 1, 2 or 3 tools for the cWGS technology. (D) Ratio of PCR validated SVs from the cWGS technology that were predicted by 1, 2 or 3 tools. (E) Ratio of different type of SVs by the 10XWGS technology validated by PCR. (F) Ratio of different type of SVs validated by PCR with respect to prediction by 1, 2 or 3 tools for the cWGS technology.
(TIF)

S16 Fig. PCR validated SVs have significantly higher GEM and requantification support. p-values were derived from Wilcoxon-rank sum test
(TIF)

S17 Fig. Training and testing logistic regression model for cWGS and 10XWGS on the test data set. (A) An unbalanced data set for training as number of PCR validated SVs are higher

than negative class from the cWGS technology. (B) Percentage importance of each feature used in the training of cWGS model calculated using varImp function of caret package. (C) Performance of the cWGS trained model on test data with different type of sampling for balancing the training data. (D) Percentage of relative feature importance calculated with dominance analysis using the complete set of PCR validated SVs trained with features derived from cWGS technology. The statistical significance was calculated using two-tailed test corresponding to z-ratio. (E) A balanced data set for training a model for 10XWGS technology. (F) Percentage importance of each feature used in the training of 10XWGS model calculated using varImp function of caret package. (G) Performance of the 10XWGS trained model on test data. (H) Percentage of relative feature importance calculated using dominance analysis with complete set of PCR validated SVs trained with features derived from 10XWGS technology. The statistical test was calculated using two-tailed test corresponding to z-ratio. The significance levels are: p-value<0.001 '***', p-value<0.01 '**', p-value<0.05 '*', p-value<0.1
(TIF)

S18 Fig. SVs predictions by trained combined model from cWGS and 10XWGS SVs in primary tumor. (A) Number of SVs prediction by the cWGS technology and percentage predicted by applying combined model. (B) Number of SVs prediction by the 10XWGS technology and percentage predicted by applying combined model. (C) Numbers and percentage of SVs common between technologies before (light colour) and after (dark colour) applying respective trained models.
(TIF)

S19 Fig. Majority of breakpoints of filtered SVs by model lie in Non-repetitive, SINE or LINE regions in MCF7. (A) The graph depicts percentage of breakpoints of SVs that lie in different repetitive regions. The SVs were filtered with the best trained combined model. (B) Breakpoints of SVs filtered by best trained combined model annotated with repetitive regions and their percentage across common, only cWGS and only 10XWGS SVs.
(TIF)

S20 Fig. The performance of combined model on internally validated SVs, two external data sets and gnomAD data set (Sample = MCF7). Sensitivity of combined model and other tools on the four data sets where SVs were predicted from (A) sequenced MCF7 sample. (B) downsampled cWGS MCF7 (equivalent coverage to 10XWGS MCF7 sample). (C) debarcoded 10XWGS linked-reads and processed with cWGS pipeline (for MCF7). (D) Overall performance of combined model on internally validated SVs with SVs predicted from downsampled cWGS reads. (E) Overall performance of combined model on internally validated SVs with SVs predicted from debarcoded 10XWGS linked-reads and processed with cWGS pipeline (for MCF7). (F) Number of gnomAD calls also present in SV calls filtered by the combined model in sequenced MCF7 sample.
(TIF)

S1 Table. Sequencing statistics for MCF7 and Primary tumor with both the technologies.
(XLSX)

S2 Table. PCR primers, PCR and Sanger sequencing results for SVs tested in MCF7.
(CSV)

S3 Table. Sensitivity and precision of combined model against other tools on external data set 1, 2 and gnomAD calls.
(XLSX)

Acknowledgments

We thank Barbara Schrörs and Karen Chu for editing the draft. We thank Stefania Gangi Maurici for PCR testing of SVs. In addition, we thank Patrick Sorn for the discussion on the development of requantification pipeline.

Author Contributions

Conceptualization: Ugur Sahin.

Data curation: Riccha Sethi, Martin Löwer, David Weber.

Formal analysis: Riccha Sethi, Martin Suchan.

Funding acquisition: Ugur Sahin.

Investigation: Riccha Sethi, Martin Löwer, Ugur Sahin, David Weber.

Methodology: Riccha Sethi, Julia Becker, Jos de Graaf, Martin Suchan.

Project administration: Martin Löwer, David Weber.

Resources: Riccha Sethi, Martin Löwer.

Software: Riccha Sethi.

Supervision: Martin Löwer, Ugur Sahin, David Weber.

Validation: Riccha Sethi, Martin Suchan.

Visualization: Riccha Sethi.

Writing – original draft: Riccha Sethi, David Weber.

Writing – review & editing: Riccha Sethi, Martin Löwer, David Weber.

References

1. Hurles ME, Dermitzakis ET and Tyler-Smith C. The functional impact of structural variation in humans. Trends Genet 2008; 24(5):238–45. https://doi.org/10.1016/j.tig.2008.03.001 PMID: 18378036
2. Nowell C. The minute chromosome (Ph1) in chronic granulocytic leukemia. Blut 1962; 8(2):65–6.
3. Treangen TJ SSL. Repetitive DNA and next-generation sequencing: Computational challenges and solutions. Nat Rev Genet 2011; 13(1):36–46. https://doi.org/10.1038/nrg3117 PMID: 22124482
4. Chaisson MJ, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL and Fan X. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun 2019; 10(1):1784. https://doi.org/10.1038/s41467-018-08148-z PMID: 30992455
5. Sedlazeck FJ, Lee H, Darby CA and Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nature reviews Genetics 2018; 19(6):329–46. https://doi.org/10.1038/s41576-018-0003-4 PMID: 29599501
6. Genome of the Netherlands C. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet 2014; 46(8):818–25. https://doi.org/10.1038/ng.3021 PMID: 24974849
7. Zheng GX, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM and Mudivarti PA. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat Biotechnol 2016; 34(3):303–11. https://doi.org/10.1038/nbt.3432 PMID: 26829319
8. Bell JM, Lau BT, Greer SU, Wood-Bouwens C, Xia LC, Connolly ID, Gephart MH and Ji HP. Chromosome-scale mega-haplotypes enable digital karyotyping of cancer aneuploidy. Nucleic acids research 2017; 45(19):e162. https://doi.org/10.1093/nar/gkx712 PMID: 28977555
9. Xia LC, Bell JM, Wood-Bouwens C, Chen JJ, Zhang NR and Ji HP. Identification of large rearrangements in cancer genomes with barcode linked reads. Nucleic acids research 2018; 46(4):e19. https://doi.org/10.1093/nar/gkx1193 PMID: 29186506

10. Eisfeldt J, Pettersson M, Vezzi F, Wincent J, Käller M, Gruselius J, Nilsson D, Lundberg ES, Carvalho CM and Lindstrand A. Comprehensive structural variation genome map of individuals carrying complex chromosomal rearrangements. PLoS genetics 2019; 15(2):e1007858. https://doi.org/10.1371/journal.pgen.1007858 PMID: 30735495

11. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N and Henaff E. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data 2016; 3:160025. https://doi.org/10.1038/sdata.2016.25 PMID: 27271295

12. Greer SU, Nadauld LD, Lau BT, Chen J, Wood-Bouwens C, Ford JM, Kuo CJ and Ji HP. Linked read sequencing resolves complex genomic rearrangements in gastric cancer metastases. Genome medicine 2017; 9(1):57. https://doi.org/10.1186/s13073-017-0447-8 PMID: 28629429

13. Viswanathan SR, Ha G, Hoff AM, Wala JA, Carrot-Zhang J, Whelan CW, Haradhvala NJ, Freeman SS, Reed SC, Rhoades J and Polak P et.al. Structural alterations driving castration-resistant prostate cancer revealed by linked-read genome sequencing. Cell 2018; 174(2):433–47. https://doi.org/10.1016/j.cell.2018.05.036 PMID: 29909985

14. Collins RL, Brand H, Redin CE, Hanscom C, Antolik C, Stone MR, Glessner JT, Mason T, Pregno G, Dorrani N and Mandrile G. Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. Genome biology 2017; 18(1):36. https://doi.org/10.1186/s13059-017-1158-6 PMID: 28260531

15. Li H and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009; 25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168

16. Faust GG HI. SAMBLASTER: Fast duplicate marking and structural variant read extraction. Bioinformatics 2014; 30(17):2503–5. https://doi.org/10.1093/bioinformatics/btu314 PMID: 24812344

17. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G and Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009; 25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

18. Cameron DL, Di Stefano L, Papenfuss AT. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software: Nat Commun 2019; 10; 3240. https://doi.org/10.1038/s41467-019-11146-4 PMID: 31324872

19. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M and Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome biology 2019; 20(1):117. https://doi.org/10.1186/s13059-019-1720-5 PMID: 31159850

20. Rausch T, Zichner T, Schlatt A, Stütz AM, Benes V and Korbel JO. DELLY: Structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 2012; 28(18):i333–i339. https://doi.org/10.1093/bioinformatics/bts378 PMID: 22962449

21. Layer RM, Chiang C, Quinlan AR and Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome biology 2014; 15(6):R84. https://doi.org/10.1186/gb-2014-15-6-r84 PMID: 24970577

22. Wala JA, Bandopadhayay P, Greenwald NF, O'Rourke R, Sharpe T, Stewart C, Schumacher S, Li Y, Weischenfeldt J, Yao X and Nusbaum C. SvABA: genome-wide detection of structural variants and indels by local assembly. Genome research 2018; 28(4):581–91. https://doi.org/10.1101/gr.221028.117 PMID: 29535149

23. Yang L, Luquette LJ, Gehlenborg N, Xi R, Haseley PS, Hsieh CH, Zhang C, Ren X, Protopopov A, Chin L and Kucherlapati R. Diverse mechanisms of somatic structural variations in human cancer genomes. Cell 2013; 153(4):919–29. https://doi.org/10.1016/j.cell.2013.04.010 PMID: 23663786

24. Elyanow R, Wu H-T, Raphael BJ. Identifying structural variants using linked-read sequencing data. Bioinformatics 2018; 34(2):353–60. https://doi.org/10.1093/bioinformatics/btx712 PMID: 29112732

25. Spies N, Weng Z, Bishara A, McDaniel J, Catoe D, Zook JM et al. Genome-wide reconstruction of complex structural variants using read clouds. Nat Methods 2017; 14(9):915–20. https://doi.org/10.1038/nmeth.4366 PMID: 28714986

26. Karimzadeh M, Ernst C, Kundaje A, Hoffman MM. Umap and Bismap: Quantifying genome and methylome mappability. Nucleic acids research 2018; 7:e30377. https://doi.org/10.1093/nar/gky677 PMID: 30169659

27. Azen R, Traxel N. Using Dominance Analysis to Determine Predictor Importance in Logistic Regression. Journal of Educational and Behavioral Statistics 2009; 34(3):319–47.

28. Li Y, Zhou S, Schwartz DC, Ma J. Allele-Specific Quantification of Structural Variations in Cancer Genomes. Cell Syst 2016; 3(1):21–34. https://doi.org/10.1016/j.cels.2016.05.007 PMID: 27453446

29. Hillmer AM, Yao F, Inaki K, Lee WH, Ariyaratne PN, Teo ASM et al. Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes. Genome research 2011; 21(5):665–75. https://doi.org/10.1101/gr.113555.110 PMID: 21467267

30. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC et al. A structural variation reference for medical and population genetics. Nature 2020; 581(7809):444–51. https://doi.org/10.1038/s41586-020-2287-8 PMID: 32461652

31. English AC, Salerno WJ, Hampton OA, Gonzaga-Jauregui C, Ambreth S, Ritter DI, Beck CR, Davis CF, Dahdouli M, Ma S and Carroll A. Assessing structural variation in a personal genome-towards a human reference diploid genome. BMC Genomics 2015; 16:286. https://doi.org/10.1186/s12864-015-1479-3 PMID: 25886820

32. Pendleton M, Sebra R, Pang AWC, Ummat A, Franzen O, Rausch T, Stütz AM, Stedman W, Anantharaman T, Hastie A and Dai H. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods 2015; 12(8):780–6. https://doi.org/10.1038/nmeth.3454 PMID: 26121404

33. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MHY and Konkel MK. An integrated map of structural variation in 2,504 human genomes. Nature 2015; 526(7571):75–81. https://doi.org/10.1038/nature15394 PMID: 26432246

34. Zook JM, Hansen NF, Olson ND, Chapman LM, Mullikin JC, Xiao C, Sherry S, Koren S, Phillippy AM, Boutros PC and Sahraeian SME. A robust benchmark for germline structural variant detection. bioRxiv 2019.

35. Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, Bjornson K, Catalanotti C, Delaney J, Fehr A and Fiddes IT. Resolving the full spectrum of human genome variation using Linked-Reads. Genome research 2019; 29(4):635–45. https://doi.org/10.1101/gr.234443.118 PMID: 30894395

36. Abel HJ DEJ. Detection of structural DNA variation from next generation sequencing data: A review of informatic approaches. Cancer Genet 2013; 206(12):432–40. https://doi.org/10.1016/j.cancergen.2013.11.002 PMID: 24405614

3 Reliable predictions of SVs from Illumina short-reads sequencing

3.1 Introduction

The structural variations (SVs) are a source of genetic variation and evolution that are responsible for diseases like cancer. The rearrangement of genomic space with SVs are well established for different sub-types of cancer (68). Such rearrangements could be simpler like deletion, duplication, inversion of genomic section on the same chromosome and rearrangement of multiple chromosomes with translocation. Apart from the simple SV events, various types of complex genomic rearrangements (CGR) like chromothripsis, chromoplexy, chains of template insertions etc. are also prevalent in cancer (4,68). Reliable and accurate pipeline for their detection would be of utmost importance in the cancer treatment.

Several new technologies like 10X Genomics linked-reads sequencing (16), long reads from Pacific Biosciences single molecule real time, Oxford nanopore sequencing (69,70) and HiC (71) etc. have emerged for the detection of SVs. However, they are expensive and demanding in terms of the input sample material and computational resources. Comparatively, the Illumina's short-reads sequencing technology is more accessible and stable with advanced bioinformatics tools for analysis. But this technology suffers from high false discovery rate (FDR). This fact was also reiterated in the benchmarking study that compared Illumina's short-reads sequencing with 10X Genomics linked-reads sequencing technology (72). Despite of high FDR, the Illumina's short-reads sequencing technology was sensitive for the prediction of SVs. It was the right ensemble of state-of-the-art bioinformatics tools that improved FDR for the prediction of SVs with short-reads.

Different strategies and combination of bioinformatics tools in an ensemble pipeline have been discussed in (73–75). A perfect ensemble of tools amalgamates the SV calls predicted with all the possible signals extracted from short-reads aligned to the reference genome. These include discordant read-pairs (RP) as pair of reads aligned with an unexpected distance or orientation to the reference genome, split reads (SR) as reads whose sequence aligns to two different locations on the reference genome that captures the exact breakpoints of SV, read depth (RD) as quantification of reads mapping in the affected genomic section covering the SV, and local *de novo* assembly (LA) where the reads around breakpoints are reassembled. Apart from different signals, an ensemble pipeline uses different filtering criteria to enlist possible SVs. The traditional approach involves filtering SV calls predicted by more than one tool as true SVs. This approach reduces FDR but also reduces sensitivity. A more recent approach considers all the calls from tools and uses algorithms from data mining to resolve the breakpoints of SVs predicted by multiple tools (76). This approach (as implemented in FusorSV) can improve sensitivity and reduce FDR of Illumina short-reads technology. However, it is limited to the prediction of simple SVs like deletion and duplication while SVs from CGR are not considered. The SVs arising from CGR in the cancer complicate the resolution of breakpoints, their type and thus, impact both sensitivity and FDR. There are specialized algorithms for predicting CGR related SVs (77–79). Henceforth, SVs related to CGR needs special treatment in an ensemble pipeline that has not been addressed previously.

Generally, the SVs are formed while repairing a nick in a DNA chromosome. There are two DNA repair mechanisms, namely template and ligate and break and ligate. The template and ligate mechanisms require a template of homologous DNA to fix the broken DNA. Based on the length of template sequence, processing and ligation involved in fixing the nicked DNA, this mechanism is further classified into homologous recombination (HR), non-allelic homologous recombination (NAHR), break induced replication (BIR) and microhomology-mediated break induced replication (MMBIR), also known as fork stalling and template switching (FoSTeS) (17,18). On the other hand, the break and ligate mechanisms directly ligates the nicked DNA without a homologous template. This mechanism includes the classical non-homologous end joining (NHEJ), alternate end-joining (alt-EJ) and single strand annealing (SSA) that differs based on initiation, processing and ligation proteins for the DNA repair. Several studies have linked the length of sequence homology and presence and length of small insertion-deletion (INDEL) around the breakpoints of SVs with the active DNA repair mechanism responsible for somatic SVs in a pan-cancer analysis (17,68). They inferred dominance of non-homology and microhomology based DNA repair mechanisms active in cancer cells. However, none of the studies have explored these criterions for filtering true somatic SV calls.

Here, the proposed FuseSV pipeline uses Illumina short-reads based SV calls from the combination of bioinformatics tools for maximum sensitivity and implements a random forest classifier for maximum precision. The classifier explores a novel set of features grouped into basic, homology and cluster features. The basic features comprise of discriminating features of SVs that can be predicted from the sequencing reads aligned to the reference genome. The homology features use the length of sequence homology and small INDELs around the breakpoint that are associated with different mechanisms of DNA repair that introduces SVs in the cancer cell. And the cluster features are derived from clustering of multiple breakpoints in a CGR region. The classifier is trained with validated SVs curated from several published and in-house studies that contain both simple and nested SVs with back-to-back calls like inverted duplication. The FuseSV trained classifier is applied to a liposarcoma cohort, a cancer type established with low somatic SNP burden and complex karyotypes (58). With this, the landscape of somatic SVs and prominent mechanisms of SVs are explored in liposarcoma for novel insights into treatment and diagnosis of this cancer type.

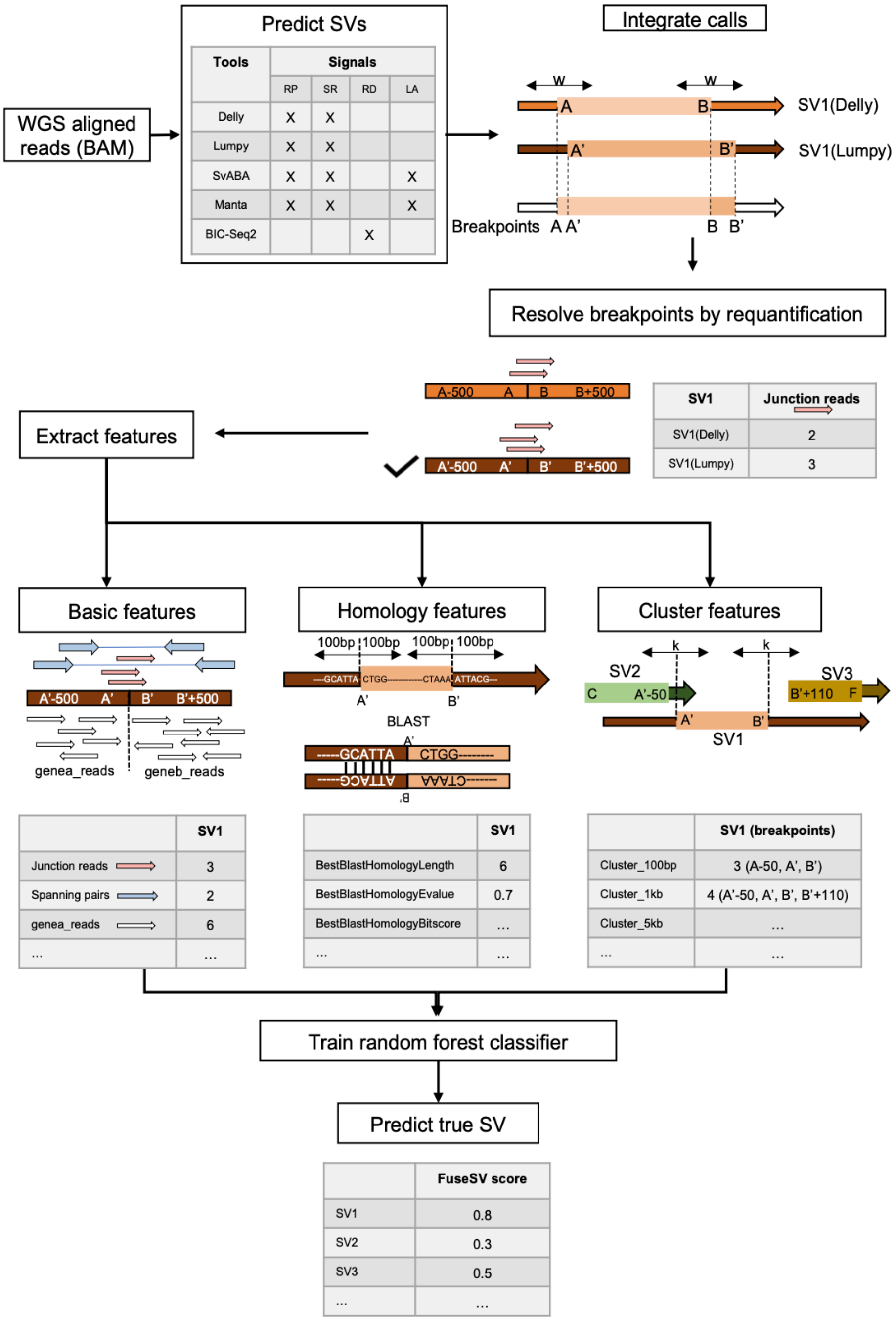
3.2 Results

3.2.1 Overview of FuseSV

A reliable prediction of somatic SVs in a clinical setting is essential for understanding tumor-specific mutations driving the disease. To address this issue, the FuseSV pipeline aligns WGS short-reads from Illumina technology from tumor and its paired normal to the reference genome and predicts somatic SVs with the five state-of-the-art bioinformatics tools (Delly (80), Lumpy (81), SvABA (82), Manta (83) and BIC-Seq2 (84)). These tools were selected due to different combination of signals they use for prediction (Figure 3-1) and their all-round performance for prediction of different types of SVs (74,75). The SV calls from multiple tools are integrated in a window of ‘w’ bp according to the type of SVs (deletion, duplication, inversion and translocation) and their orientation (3’to5’, 5’to3’, 3’to3’ and 5’to5’). The SVs from BIC-Seq2 are not included in this integration as it reports large size copy number variations (CNV) using read depth information that is not accurate in terms of breakpoints. The breakpoints of overlapping calls are resolved using the requantification process. As also explained in (72), the requantification process first generates a synthetic template of 1kb size that contains 500bp of reference genome’s sequences around the breakpoints of SVs. The WGS reads are aligned to this template and supporting reads at the junction of merged segments (junction reads) and read-pairs that span the merged segments (spanning read-pairs) are quantified. The presence of junction reads indicates correct breakpoints of respective SVs, therefore, breakpoints with maximum number of junction reads is selected for the overlapping calls (Figure 3-1 and Supplementary figure 3-1A). With the SV calls containing resolved breakpoints, a novel set of features are extracted for training a random forest classifier that are explained in detail in Supplementary table 3-2. These include:

Basic features

They are the discriminating features of a true SV some of which were established in the benchmarking studies (74,75). These include features like number of tools predicting the same SV, type and size of SV, local coverage around the breakpoints, presence of INDEL around the breakpoints and mappability of predicted breakpoints. It also includes features derived from the requantification process like junction reads, spanning read-pairs and number of reads mapping in the individual segments of merged genomic templates. Additionally, the features indicating presence of CNV are also captured in this set.



Tools	Signals			
	RP	SR	RD	LA
Delly	X	X		
Lumpy	X	X		
SvABA	X	X		X
Manta	X	X		X
BIC-Seq2			X	

SV1	Junction reads
SV1(Delly)	2
SV1(Lumpy)	3

	SV1
Junction reads	3
Spanning pairs	2
genea_reads	6
…	…

	SV1
BestBlastHomologyLength	6
BestBlastHomologyEvalue	0.7
BestBlastHomologyBitscore	…
…	…

	SV1 (breakpoints)
Cluster_100bp	3 (A-50, A', B')
Cluster_1kb	4 (A'-50, A', B', B'+110)
Cluster_5kb	…
…	…

	FuseSV score
SV1	0.8
SV2	0.3
SV3	0.5
…	…

Figure 3-1: FuseSV integrates calls and use machine learning for the accurate predictions of SVs. The SV predictions are made by Delly, Lumpy, SvABA, Manta and BIC-Seq2 using aligned WGS reads. All calls are combined based on their breakpoint and type of SV. The breakpoints of overlapping SV calls are resolved by requantification. This involves creation of a synthetic genomic template with neighboring sequences around the breakpoints in a window of 500bp and alignment of reads to it. The breakpoints of SV with maximum reads mapping to the junction of merged genomic segments (junction reads) is selected. For each SV call, features are extracted for random forest classifier. These include basic, homology and cluster features. The basic feature like junction reads, spanning pairs, reads mapping to each section of merged synthetic genomic template (genea_reads, geneb_reads) are extracted from the requantification process. The homology feature is derived from basic local alignment of sequences around the breakpoints of the SV using BLASTN. The cluster feature is quantified with the number of breakpoints of SVs clustering in variable range of window (100bp, 1kb, 5kb, 10kb). The trained classifier with these features is used to predict probability of true SV calls (FuseSV score). RP: Discordant read-pairs, SR: Split reads, RD: Read depth, LA: de-novo local assembly.

Homology features

These set of features are inspired from the fact that several error-prone DNA repair mechanisms that introduces SV in the genome such as NHEJ, NAHR, alt-EJ, SSA, MMBIR relies on small to higher degree of sequence homology around their breakpoints. The length of sequence identical around the breakpoints and presence and length of INDEL indicates prominent mechanism of DNA repair active in the cell. This indirectly refers to the prominent mechanisms by which an SV was created. Such features were derived by basic local alignment of sequences around the breakpoints of SV using BLASTN (85) (Supplementary figure 3-1B). The length of identical sequences, percentage identity, expectation value of BLASTN hits and their respective bit scores were the key features explored in this set.

Cluster features

This set of features aimed to serve two purposes. It is established that CGR are prevalent in cancer genome (68). With Illumina's short-reads sequencing technology, discovery and resolution of CGR is skewed that requires specialized set of tools. These tools use advanced technologies like linked-reads from 10X Genomics and long reads from Pacific Biosciences or Oxford Nanopore. However, the breakpoints of such rearrangements in aligned short-reads can be predicted with some sensitivity that requires incorporation of multiple SV events together for complete resolution (86). In FuseSV we aim to learn difference in features between simple SVs and SVs in CGR rather than resolve the breakpoints. The other purpose is to account for the SVs arising due to alignment-based artefacts which are often located in low complexity regions like simple repeats. Henceforth, this includes calculation of number of breakpoints clustering together in the variable set of windows around the breakpoints (100bp, 1kb, 5kb and 10kb) (Figure 3-1 and Supplementary figure 3-1C).

With the collection of mentioned features, a random forest classifier is trained on carefully curated set of SVs from various cancer cell lines and tumor genomes that have been additionally validated with PCR (Supplementary table 3-1).

3.2.2 FuseSV predicts SVs with high sensitivity and precision

Four random forest classifiers were trained with different combination of the features included in three main sets i.e., basic, homology and cluster features (Basic, Basic+Cluster, Basic+Homology, and, Basic+Homology+Cluster features). These classifiers were trained with 1138 SVs (varying sizes of deletions, duplications, inversions, translocations and nested SVs like Supplementary figure 3-2) that were collated from several studies and validated with PCR (Supplementary table 3-1). Since a high number of features were included in three sets of features, they were screened based on their relative importance in order to prevent overfitting. The features with 5% or higher relative feature importance was selected for training of the final models (Supplementary figure 3-3). Overall, the basic set of features had relatively higher feature importance in comparison to the homology and cluster set of features (Figure 3-2A). Amongst the basic features, the number of tools predicting a SV (NumberTools), spanning reads (Span_reads) and junction reads (Junc_reads) from the requantification pipeline achieved 97.8%, 87.4% and 52.1% of relative feature importance. Moreover, the homology features with the expectation value from BLASTN local alignment of sequences around the breakpoints (BestBlastHomologyEvalue), their bit scores (BestBlastHomologyBitscore) and the length of sequence similar around the breakpoints of SVs (BestBlastHomologyLength) achieved relative feature importance of 24.6%, 24.5% and 20.5% respectively. Amidst the cluster features, the number of breakpoints clustered in 10kb (Cluster_10kb) and 5kb (Cluster_5kb) window around a SV's breakpoints achieved 10.9% and 8.1% of relative feature importance.

The performance of trained classifiers was tested with two approaches (Supplementary table 3-1). They were analysed with 5-fold cross-validation repeated 10 times and two test data sets. The cross-validation analysis for the trained classifiers revealed the classifier with basic, homology and cluster features with maximum area under curve (AUC) of 0.962 in the receiver operating curve (ROC) (Figure 3-2B). However, the performance of other trained classifiers was comparable with AUC as 0.956 (Basic), 0.958 (Basic+Cluster) and 0.962

(Basic+Homology). Nevertheless, each of the trained classifier had better performance than the individual (Delly: 0.82, Lumpy: 0.797, SvABA: 0.819, Manta: 0.884) and the calls predicted by all the tools (Consensus: 0.827).

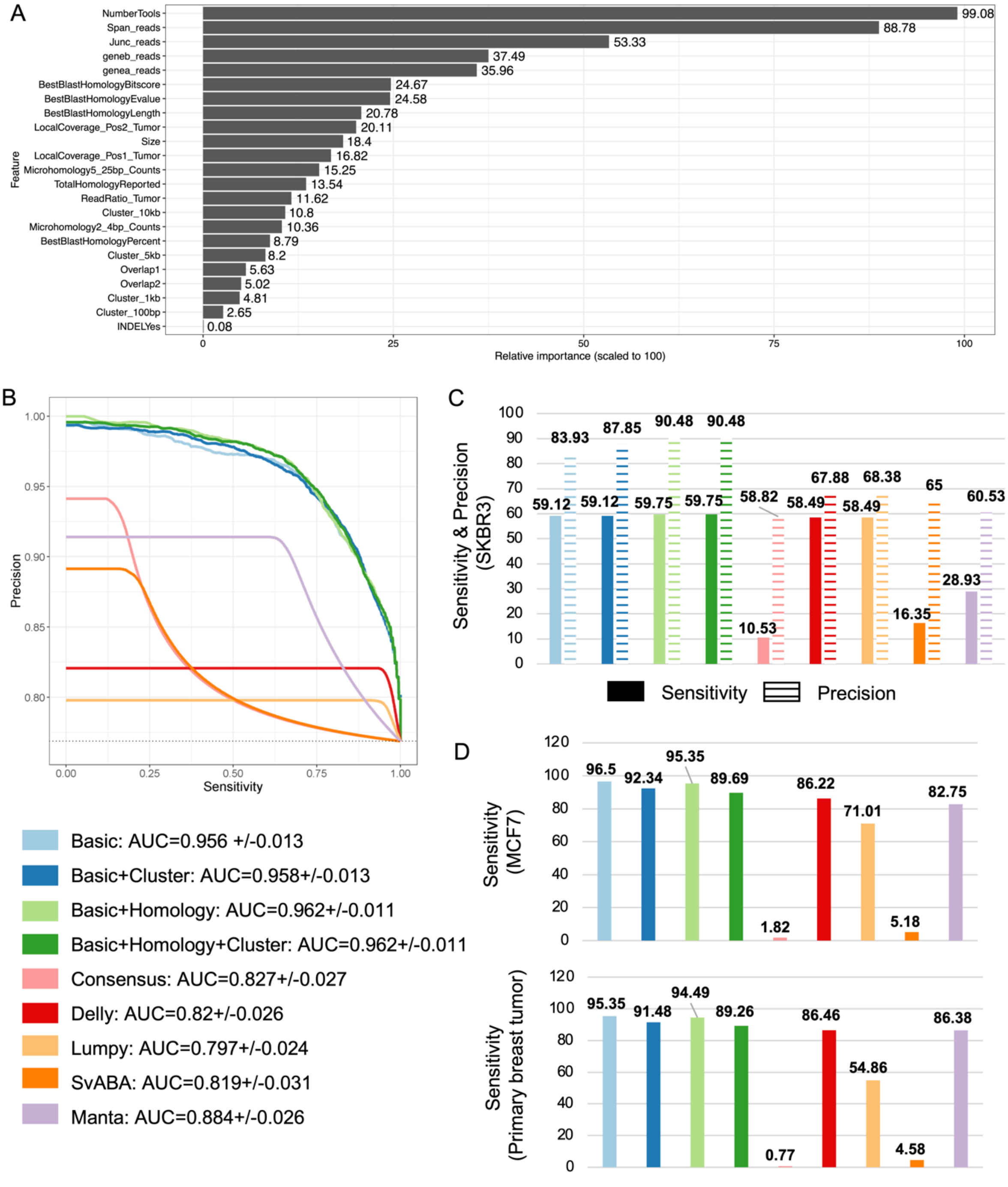


Figure 3-2: The ensemble of SV prediction tools and the random forest (RF) classifier in FuseSV generates higher sensitivity and precision. A) Percentage relative importance of features used in the classifier trained with all the set of features. This data was derived from the average values obtained with 5-fold cross-validation repeated 10 times. B) Area under curve (AUC) in receiver operating classifier (ROC) curve for FuseSV with different combination of extracted features (Basic, Basic+Cluster, Basic+Homology and Basic+Homology+Cluster) is higher than the individual tools and SV calls predicted by all the tools (Consensus). The ROC curve was generated with 5-fold cross-validation repeated 10 times. C) Performance of FuseSV (different combination of features) and individual tools is higher on the test data. This test data was derived

from SKBR3 SVs predicted by Pacific Biosciences long reads and validated by PCR and Sanger sequencing. D) Sensitivity of FuseSV (different combination of features in trained classifiers) and individual tools on test data generated with 10X Genomics linked-reads sequencing validated calls in the MCF7 cell line and a primary breast tumor.

The performance of classifiers was further tested on two test data sets. The first data set included SVs in SKBR3 predicted by Pacific Biosciences long reads that were validated by PCR and Sanger sequencing (87). This test data set included calls that were both validated and not validated by Sanger sequencing and thus, we compared sensitivity and precision for each classifier and individual tools. As depicted in Figure 3-2C, the classifier trained with all the set of features had maximum sensitivity and precision of 59.75% and 90.48% respectively. The improvement in precision by inclusion of homology and cluster features is clearly evident when precision increased from 83.83% in the classifier trained with only basic features to 90.48% in the classifier trained with all the features (Basic+Homology+Cluster). Comparatively, the consensus SVs as calls predicted by all the tools had minimum sensitivity of 10.53% and surprisingly, lower precision of 58.82%. However, sensitivity of Delly (58.49%) and Lumpy (58.49%) was comparable to the trained classifiers while their precision was lower (Delly: 67.88%, Lumpy: 68.38%). The performance of both SvABA (sensitivity: 16.38%, precision: 65%) and Manta (sensitivity: 28.93%, precision: 60.53%) was lower in this test data set. The second data set included SVs in the MCF7 cell line and a primary breast tumor that was validated with 10X Genomics linked-reads sequencing (72). As depicted in Figure 3-2D, the sensitivity of the classifiers trained with all three sets of features was slightly lower than the classifier trained with only basic features (Basic+Cluster: 92.34% in MCF7 and 91.48% in a primary breast tumor, Basic+Homology: 95.35% in MCF7 and 94.49% in a primary breast tumor, Basic+Homology+Cluster: 89.69% in MCF7 and 89.26% in primary breast tumor). Nevertheless, it was better in comparison to the other tools (Delly: 86.22% in MCF7 and 86.46% in primary breast tumor, Lumpy: 71.01% in MCF7 and 54.86% in primary breast tumor, SvABA: 5.18% in MCF7 and 4.58% in primary breast tumor, Manta: 82.75% in MCF7 and 86.38% in primary breast tumor, Consensus: 1.82% in MCF7 and 0.77% in primary breast tumor).

In conclusion, our trained model achieved maximum performance with higher sensitivity and precision in comparison to other tools and calls predicted by all the tools (Consensus). Since there is always a trade-off between sensitivity and precision, with minimal drop in sensitivity the trained classifier with all the set of features attained better precision. The performance of trained classifiers over different test data sets indicates the robustness of FuseSV.

3.2.3 Application of FuseSV to liposarcoma cohort

31 liposarcoma samples (DDLS: 21 samples, MLS: 8 samples, WDLS: 2 samples) with average median coverage of 74.5X in DDLS samples, 77.4X in MLS samples and 74X in WDLS samples with 2X151bp (paired-end reads) of read length were analysed with the FuseSV pipeline.

A total of 116624 of high confidence somatic SVs (with FuseSV score>=0.7) were identified in the liposarcoma samples. Amongst all the sub-types of liposarcoma analysed in this study, the DDLS sample had significantly higher average number of SVs (4029 SVs per sample) in comparison to MLS (3075 SVs per sample) (Figure 3-3A). The WDLS sub-type had an average of 3708 SVs per sample. However, the significance of difference in numbers between this sub-type with others was not established as only two samples were classified with WDLS liposarcoma.

The CNV profiles of the samples of different sub-types of liposarcoma were investigated in Figure 3-3B. Around 90% of the DDLS samples were characterised with the amplification in chromosome 12q arm that was also observed in the two WDLS samples. Moreover, the WDLS samples also had amplification on chromosome 6q arm. Contrastingly, these amplifications were absent in the MLS samples that were characterised with an amplification in chromosome

8q arm in nearly 50% of the samples. Overall, DDLS samples had a more complex CNV profile with many amplifications and deep deletions on different chromosomes.

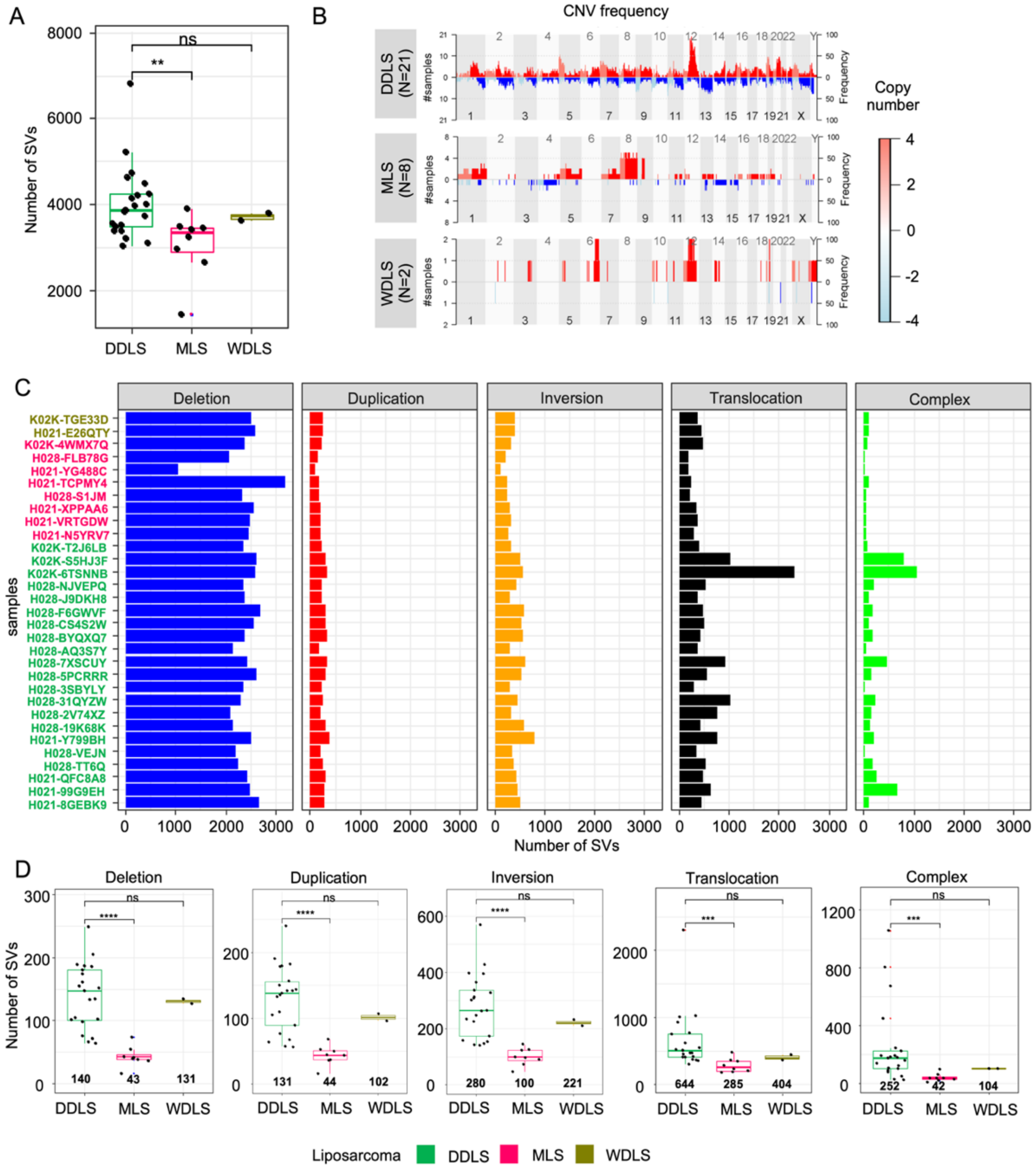


Figure 3-3: DDLS type of sarcoma has significantly higher number of SV. A) Predicted number of SVs amongst different sarcoma type. B) The copy number variation (CNV) and its frequency in sarcoma samples. C) The number of large sized SV (size>100kb) of different types (deletion, duplication, inversion, translocation and complex) predicted across different liposarcoma samples. The p-values in all these plots is derived with Kruskal-Wallis test with following significance levels: ns-not significant, *-value <0.05, **-value <0.01, *-value <0.001, ****-value <0.0001. The mean value of number of SV for sample type in graph is mentioned at bottom in bold.**

The distribution of different types of somatic SVs (deletion, duplication, inversion, translocation and complex SVs that contained more than 10 breakpoints in the cluster of 5kb window) was investigated in Figure 3-3C and Supplementary table 3-3. The deletions were most prevalent in all the liposarcoma samples. The DDLS samples had an average of 61% deletions, 6.7% duplications, 11.4% inversions, 15% translocations and 5.5% complex SVs per sample. While the MLS samples had an average of 75% deletions, 5.85% duplications, 9.45% inversions, 9.45% translocations and 1.2% complex SVs per sample. The average SVs frequency in the

WDLS samples were like the DDLS samples (68.8% deletions, 6.7% duplications, 10.7% inversions, 10.8% translocations and 2.8% complex SVs). Furthermore, the frequency of different types and sizes of somatic SVs were explored in Figure 3-3D, Supplementary figure 3-4 and Supplementary table 3-3). The major difference in the number of SVs between DDLS and MLS samples were attributed to significantly higher number of large-sized SVs (size>100kb or translocations and complex SVs) present in the DDLS samples. The mean number of large-sized SVs per sample in the DDLS were 140 deletions, 131 duplications, 280 inversions, 644 translocations and 252 complex events. Relatively, the MLS samples had 43 deletions, 41 duplications, 100 inversions, 285 translocations and 45 complex SVs. The mean number of large-sized SVs observed in WDLS were 131 deletions, 102 duplications, 221 inversions, 404 translocations and 104 complex SVs. These numbers were closer to the umbers reported in DDLS samples.

3.2.4 Chromosome shattering in DDLS and WDLS sub-types of liposarcoma

The chromosome 12 was amplified in 90% of the DDLS samples as visualized in (Figure 3-3B). On further analysis, it was observed that nearly 60% and 70% of complex SVs had breakpoint on chromosome 12 in DDLS and WDLS cohort respectively (Figure 3-4A). In contrast, only 9.5% of complex SVs in the MLS samples had breakpoints on chromosome 12. These results indicate immense shattering and rearrangement of chromosome 12 in the DDLS and WDLS samples.

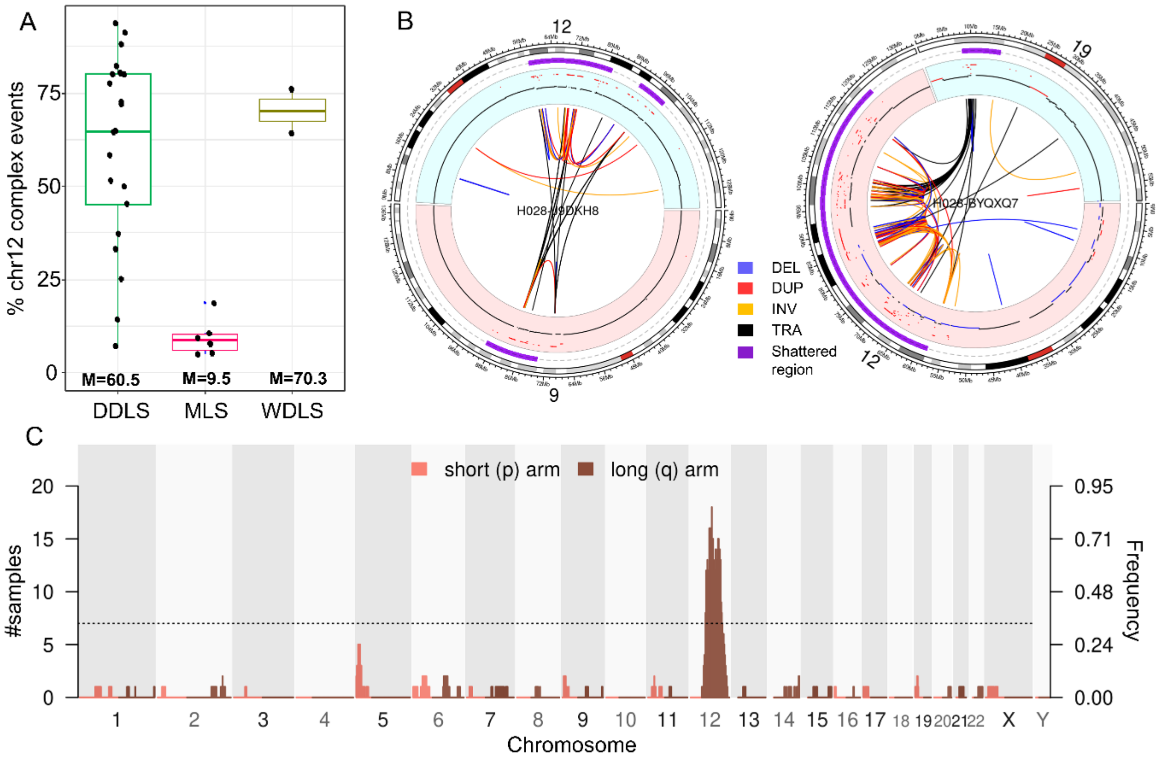


Figure 3-4: Chromosome 12 in majority of the DDLS samples is shattered and rearranged. A) Percentage of complex SVs on chromosome 12 within different type of liposarcoma samples (DDLS, MLS and WDLS). The mean (M) proportion of complex SV on chromosome 12 is depicted in bold. B) Circos plot of two DDLS samples representing different type of SVs (deletion-DEL, duplication-DUP, inversion-INV, translocation-TRA) and rearrangement of the shattered chromosome 12. C) Proportion of DDLS samples with the shattered chromosome 12q arm. Figure B and C were generated with svpluscnv (77) package in R.

The CGR in the liposarcoma samples were further investigated with svpluscnv package (77). This package integrates the CNV and SV information to find common regions shattered amongst the samples. Amongst the DDLS samples, 19 out of 21 (except H028-VEJN and

H028-3SBYLY) samples had chromosome 12 102012374-112012374 coordinates severely rearranged (as seen in circos plot of two DDLS samples in Figure 3-4B). This region lies on chromosome 12q arm that is significantly shattered in 90% of the DDLS samples (Figure 3-4C and Supplementary figure 3-5). In the DDLS samples, chromosome 12q shattering resembled chromothripsis where oscillation between different copy numbers and SVs was observed. Moreover, the events resembling breakage-fusion-bridges that affect telomeres with fold back inversions were also observed in 9 out of 21 DDLS samples. Some of the samples like K02K-S5HJ3F and K02K-6TSNNB had severely rearranged chromosomes with multiple events oscillating between chromosome 6, 12, 2 and 5. The shattering of chromosome 12q was also observed in the two WDLS samples with two hotspots of shattered regions were: 68012374-78012374 and 90012374-100012374 (Supplementary figure 3-6). Even though MLS samples included complex SV events, a common genomic region shattered between the samples was not seen. When all the large-sized SVs (size>100kb, translocation and complex events) were considered, shattering was observed only in one sample of MLS cohort (K02K-4WMX7Q). This caused rearrangement of chromosome 1 (76049148-86049148, 156049148-166049148, 172049148-182049148), chromosome 8 (30071162-40071162), chromosome 11 (100192287-110192287) and chromosome 13 (22180033-32180033, 44180033-54180033) as seen in Supplementary figure 3-7. Such pattern of shattering resembled chromoplexy with chains of rearrangement in a closed loop.

3.2.5 Mechanisms of SVs across liposarcoma

The formation of SVs via different DNA-repair mechanisms like NAHR, NHEJ, alt-EJ, FoSTeS/MMBIR, SSA etc. are explored in this section. The length of homologous sequence around the breakpoints, presence of INDEL at the merged genomic sections and its length can indicate the prevalent DNA-repair mechanisms for the SV formation. For example, NAHR is known to be prominent mechanism when the neighboring region around the breakpoints have longer homology. Whereas NHEJ is active when no to very small microhomology (0-4bp) along with small INDELs at the breakpoints are observed. Different alt-EJ mechanisms like MMEJ, SD-MMEJ and TMEJ are associated with small microhomology of 1-8bp around the breakpoints. SSA is prominent mechanism when 15-70bp of homologous sequences are found around the breakpoints whereas FoSTeS/MMBIR mechanism are known to form complex SVs which are often found with INDELs and microhomology at the merged genomic section. Here

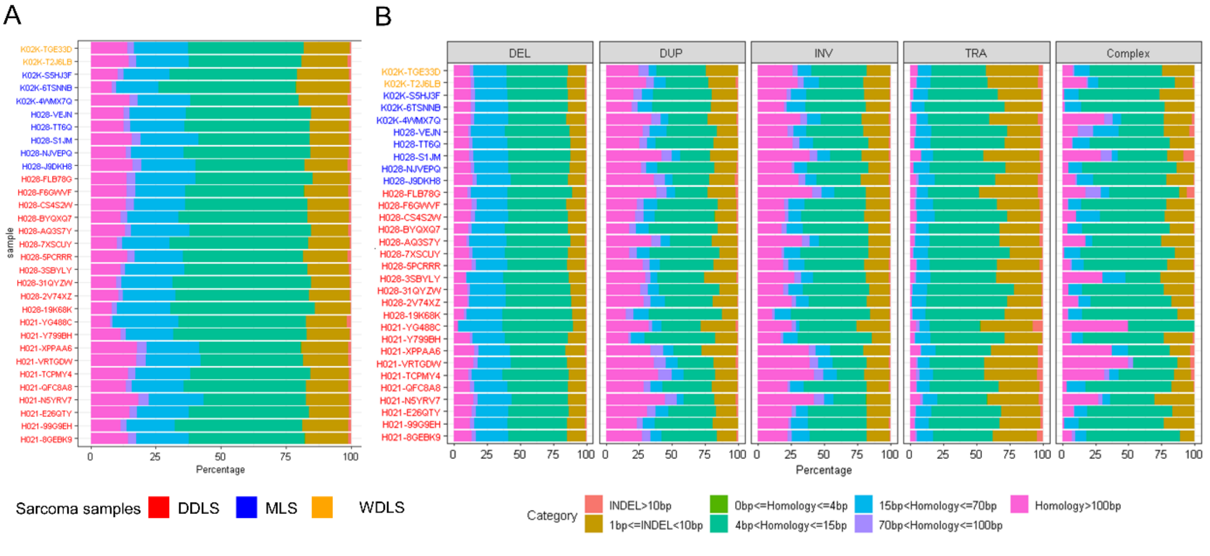


Figure 3-5: Majority of SVs across liposarcoma samples have 4-15bp homology around the breakpoints. A) The plot depicts percentage of SVs in each sample that have variable length of homology around the breakpoints and INDEL at the merged genomic section. B) The plot depicts percentage of different type of SVs (DEL-deletion, DUP-duplication, INV-inversion, TRA-translocation and Complex) in each sample with variable length of homology around the breakpoints and INDEL at the merged genomic section.

we analyse these parameters with BestBlastHomologyLength and INDEL length features that were calculated in the FuseSV pipeline applied to the liposarcoma sample.

As depicted in Figure 3-5A, 47% of SVs had microhomology between 4-15bp around the breakpoints amongst all the samples. This indicates alt-EJ pathways is the most active mechanisms of DNA repair in liposarcoma. Furthermore, 20% SVs had 15-70bp homology around the breakpoints, 16.3% SVs had INDEL of length 1-10bp at the merged genomic section, 13% SVs had homology greater than 100bp, 2.6% SVs had 70-100bp homology, 0.8% SVs had INDEL of length greater than 10bp. This indicates that apart from alt-EJ, the SSA repair mechanism that is characterized with presence of 15-70bp of homology around the breakpoints of SVs is the next most active DNA repair mechanisms in liposarcoma.

Amongst the different type of SVs in Figure 3-5B, 46.6% and 25.6% of deletions contained 4-15bp and 15-70bp of homology around the breakpoints respectively. While 37.4% and 28.4% of duplications had 4-15bp and greater than 100bp of homology around the breakpoints respectively. Like duplications, inversions had 40.3% and 26.6% with 4-15bp and greater than 100bp homology around the breakpoints respectively. However, amidst translocations, 50% and 32.2% had 4-15bp homology around the breakpoints and INDELs of 1-10bp length at the merged genomic sections respectively. The SVs characterized as complex followed a pattern like translocations i.e., 54.4% and 18.4% had 4-15bp homology around the breakpoints and INDELs of 1-10bp length at the merged genomic section respectively. It can be estimated from the distribution that deletions are predominantly formed via alt-EJ and SSA pathways, duplications and inversions with alt-EJ and NAHR pathways, and translocations and complex SVs with alt-EJ and FoSTeS/MMBIR pathways. Since this cohort has very high number of small-sized deletions (size<=1kb), we investigated the prominent mechanisms active for the formation of large-sized SVs (size>100kb, translocation or complex SVs) in Supplementary figure 3-8. It was observed that majority of deletions, duplications and inversions had 4-15bp and greater than 100bp homology around the breakpoints. This indicates prevalence of alt-EJ and NAHR repair mechanisms that leads to formation of large sized deletions, duplications and inversions.

Since chromosome 12 in DDLS and WDLS samples was significantly rearranged, we investigated the prominent repair mechanism observed with the SVs having breakpoints on chromosome 12. As observed in Figure 3-6A, two most dominant mechanism in the DDLS samples had homology of 4-15bp around the breakpoints (60.3% of chromosome 12 SVs against 45.5% in SVs on other chromosomes) and INDELs of 1-10bp length at the merged genomic section (17.8% of chromosome 12 SVs against 15.8% in SVs on other chromosomes). Moreover, SVs with breakpoints on chromosome 12 had significantly lower calls with homology greater than 100bp and significantly higher calls with homology of 4-15bp around the breakpoints against SVs on other chromosomes in the DDLS samples (Figure 3-6B). This indicates prevalence of alt-EJ and MMBIR/FoSTeS and downregulation of NAHR repair mechanisms for the formation of SVs on chromosome 12. This significant difference in

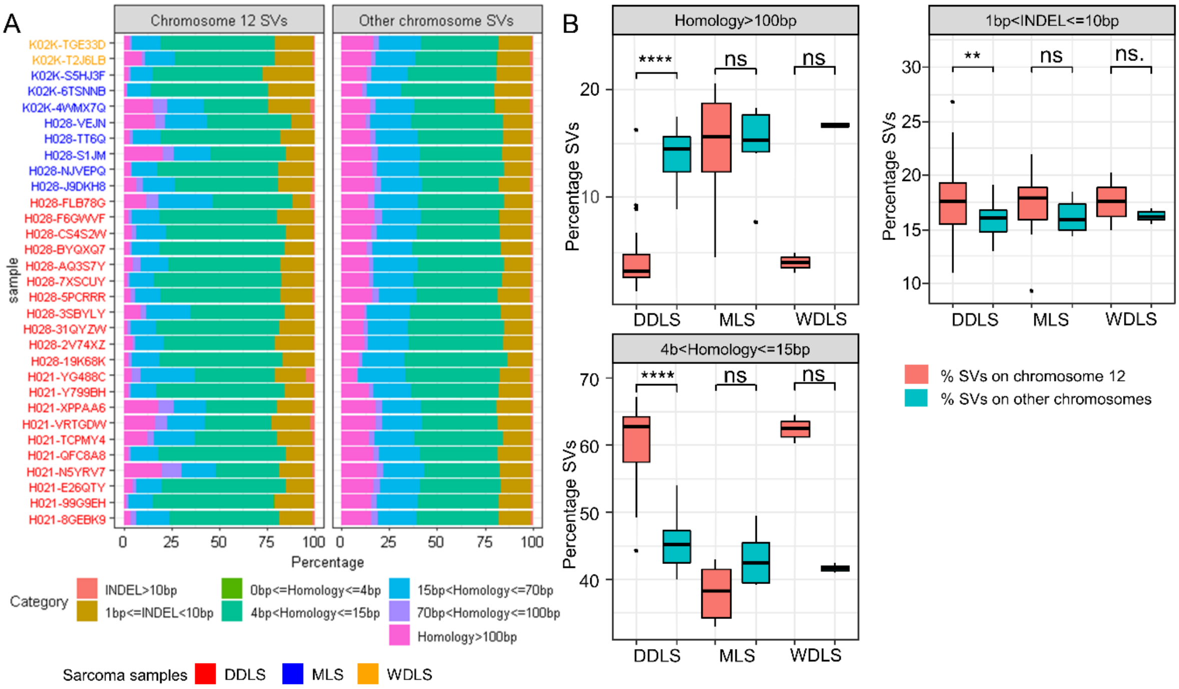


Figure 3-6: SVs on chromosome 12 of the DDLS samples are created with mechanism that utilize homology of length 4-15bp around the breakpoints. A) Different liposarcoma samples plotted against percentage of SVs containing variable length of homology and INDELs around the breakpoints. B) Percentage of SVs on chromosome 12 against SVs on other chromosomes with different homology length for different liposarcoma sub-types. The significant difference in proportion of SVs with different homology length on chromosome 12 against another chromosome's SV was derived with Wilcoxon paired test. The p-values are mentioned as: ns-not significant, *-value <0.05, **-value <0.01, *-value <0.001, ****-value <0.0001.**

percentages was not established between MLS and WDLS samples. Nevertheless, WDLS samples followed a similar pattern as DDLS samples.

3.3 Discussion

Using an ensemble of SV prediction tools comes with advantages and disadvantages. The advantage includes predicting SVs of various types and sizes with higher sensitivity and precision. The proposed FuseSV pipeline presents these advantages while overcoming the challenges in resolution of calls from multiple tools along with consideration of SVs arising in the CGRs that are prevalent in complex diseases like cancer. Moreover, we also explored novel features like homology around the breakpoints of somatic SVs that further improved precision without much compromise in the sensitivity. With the unique combination of features in a machine learning approach, FuseSV was robust in prediction of all types of SVs. However, there are few limitations to the FuseSV pipeline. Currently, FuseSV accepts SVs calls from Delly, Lumpy, SvABA, Manta and BIC-Seq2, and the future update would include calls from other bioinformatics tools as well. Secondly, the random forest classifiers were trained with SVs predicted by the Illumina short-reads sequencing technology that has its inherent limitations. One of the future updates would include SV calls from other sequencing technologies for training. Nevertheless, FuseSV classifiers were trained with SVs from the cancer cell lines and primary tumors that imitates cancer model more closely in comparison to other tools that were trained with SVs from normal human genome (example FusorSV (76)).

Adult soft tissue sarcoma is one of the many highly aggressive cancer types with variable type of karyotypes from simple to more complex ones and low mutational burden in terms of SNV. In this study, we analysed samples from a specific class of sarcoma, liposarcoma, with three pathologically classified sub-types i.e., DDLS, MLS and WDLS. Complementary to the findings by Abeshouse et. al. (58), we also observed the amplification of chromosome 12q arms in the DDLS and WDLS sub-types of liposarcoma. The DDLS sub-type is known to originate when

WDLS invades non-lipogenic regions and becomes more aggressive. This is substantiated with a similarity in the SV and CNV profiles amongst DDLS and WDLS samples. However, the DDLS samples had a much wider complex spectrum of copy number changes in comparison to WDLS and MLS samples. This was also demonstrated with significantly higher number of SVs in the DDLS samples that was associated with the higher frequency of large-sized SVs (size>100kb), translocations and complex SVs. The characterization of liposarcoma samples with higher number of complex SV events was also established in the PCAWG study (68). However, further classification of soft tissue liposarcoma to different sub-types is missing in the PCAWG study.

The genomic landscape of complex SVs in DDLS samples was specifically prominent on chromosome 12. The clustering of several breakpoints on this chromosome was observed in 90% of DDLS samples. This indicated severe shattering of chromosome 12q arm that might be under selection pressure in DDLS patients. The pattern of rearrangements on the chromosome 12 sometimes resemble chromothripsis and breakage-fusion-bridges, but often it is highly complex without any known classification of CGR. This was also reported in the PCAWG study of sarcoma samples (68), where 25% of soft tissue liposarcoma samples resembled chromothripsis while remaining events were other CGR types. Nevertheless, the chromothripsis is postulated to be an initiator event for rearrangements involved on chromosome 12 of DDLS and WDLS (88) that is followed by multiple rearrangements like formation of double minutes and neochromosomes. Interestingly, chromothripsis in tumors is associated with lower infiltration of cytolytic T cells, natural killer cells and tumor antigen presentation markers, and tumor aneuploidy (88). Additionally, tumor aneuploidy is associated with reduced response to the immune checkpoint-based immunotherapy (89). An indirect inference from these studies would suggest lower efficacy of classical immunotherapy for the treatment of DDLS sub-type of liposarcoma. Henceforth, personalized vaccines developed from neoantigens would be efficient in managing this sub-type of cancer. However, further research and investigation is required in this direction.

Apart from characterization of CGR on chromosome 12 in the DDLS samples, we also investigated the prominent DNA repair mechanisms actively involved for somatic SVs on chromosome 12. There was a significant increase in SVs utilizing 4-15bp microhomology around the breakpoints and presence of INDELs of 1-10bp size in SVs on chromosome 12. This implies dominance of alt-EJ and MMBIR pathways for the DNA repair on this chromosome. SVs on chromosome 12 are part of CGR that to some extent resemble chromothripsis in our DDLS samples. Our findings are consistent with a study that hypothesis formation of chromothripsis like events with alt-EJ pathway (9). While there has also been evidence for the involvement of NHEJ and MMBIR pathways in chromothripsis events (90). Conclusively, alt-EJ pathways play an important role in the formation of SVs on chromosome 12 of DDLS and WDLS samples, while contribution by MMBIR and NHEJ pathways is inevitable.

Illumina short-reads sequencing technology is a popular approach for detecting SVs, however, third generation sequencing technologies like Pacific Biosciences and Oxford nanopore long reads sequencing, 10X Genomics linked-reads sequencing and HiC can detect more types of SVs that are not approachable with shorter reads. Going forward combination of different technologies for overall detection of SVs would shed light over novel insights in mechanisms of DNA repair used for formation of SVs. This would also expand the current landscape of rearrangements in any cancer genome and offer diagnosis and treatment related advice by the medical doctors.

3.4 Methods

3.4.1 Whole genome sequencing (WGS) samples and upstream processing

WGS samples were obtained from various sources as mentioned in Supplementary table 3-1. These included MCF7 breast cancer cell line, SKBR3 breast cancer cell line, MZ-GaBa-018 breast cancer cell line and 5 primary breast tumor (obtained from two sources). The fastq files containing reads of respective samples were aligned to the reference genome GRCh38 using

BWA-MEM (v0.7.17), duplicate reads were removed with Samblaster (v0.1.24-0) and aligned reads sorted by coordinated with Samtools sort (v1.3.1).

3.4.2 Prediction of SVs with an ensemble of bioinformatics tools

The short-reads aligned to the reference genome the .bam file served as input to the bioinformatics tools. Generally, bioinformatics tools use discordant read-pairs (RP), split reads (SR) and local de novo assembly (LA) to predict SVs. They can be used to identify different types of SVs like deletion, duplication, inversion and translocations. Apart from these strategies, read depth (RD) or variable number of reads against neighbouring sections can be used to identify CNV like deletions and duplications. An ensemble of tools was chosen to combine all these signals. This was inspired from (72) along with an additional tool utilizing RD information for detection of CNV. This ensemble included Delly (v0.7.6), Lumpy (v0.2.13), SvABA (v0.2.1), Manta (v1.6.0), and BIC-Seq2 (normalization v0.2.4 and segmentation v0.7.2).

Each SV call was classified into deletion, duplication, inversion, translocation and complex events. A deletion was defined as an event when section was deleted (associated with 3'to5' orientation). A duplication was defined as an event when section was duplicated (associated with 5'to3' orientation) and inserted back-to-back (this type was checked with requantification process as explained in the next section). However, duplication events with insertion somewhere else are also considered but they were not verified by requantification. Inversions were defined with two events where two different genomic segments merged (associated with 3'to3' and 5'to5' orientation). All the inter-chromosomal events were classified as translocation. The events that had more than 10 number of breakpoints clustering in a 5 kb window around the breakpoints were classified as complex SVs.

3.4.3 Integration of calls

Each SV call from tools (predicted with filter "PASS") were classified according to the orientation predicted by the tools. For example, an inversion can have two corresponding orientation 3'to3' and 5'to5'. Each of this orientation is counted as two calls. With this, calls from all the tools (except Bic-Seq2) were integrated as one when their breakpoints lie within a 500 bp window and have same orientation. The strategy described as requantification was inspired from (72) for the resolution of breakpoints of overlapping calls. The process involved creation of a synthetic genomic template of 1kb size with 500bp on each side of the breakpoint. This template was generated with the sequence from reference genome around the breakpoints. Further, the WGS reads of the sample were aligned to this template using BWA-aln. With the aligned reads on the template, junction reads as the reads mapping on the merged segments were retrieved for resolving the breakpoints. The breakpoints with maximum junction reads were considered as final breakpoints of that SV call (Supplementary figure 3-1A).

3.4.4 Feature extraction

Three set of features were extracted for each SV call for training random forest classifiers. They were categorized as basic, homology and cluster features. As listed in Supplementary table 3-2, the basic features included junction reads, spanning read-pairs, number of reads mapping to the first and second segment merged in synthetic genomic template of the requantification pipeline, number of tools predicting a SV, overlap of SV with CNV from BIC-Seq2 and vice-versa, presence of INDEL around the breakpoints, size and type of SV, local coverage around the breakpoints and ratio of reads mapping between breakpoints of SV.

The homology features involved the calculation of homologous sequence around the breakpoints of SVs and INDEL at the merged segments. These features were extracted from output results of BLASTN (v2.5.0) of 200 bp of genomic sequence around each breakpoint of each SV call (Supplementary table 3-2 and Supplementary figure 3-1B). This included the length of homologous sequence hits, its identity, bit score and expectation value as reported with BLASTN. The length of INDELs at SV's merged segments were calculated with the requantification pipeline while aligning reads to the synthetic genomic segment.

The cluster features were calculated as the number of SV breakpoints clustering in a variable size of windows. The sizes of window were 100bp, 1kb, 5kb and 10kb (Supplementary table 3-2 and Supplementary figure 3-1C).

3.4.5 Random forest classifier

A validated list of SVs was collated from several sources and samples (Supplementary table 3-1). The labels for training were derived from SVs that were validated by PCR that included 1138 data points with 875 positive calls and 263 negative calls (Supplementary table 3-4). Four random forest classifiers were trained with different sets of features. First was trained with basic features, second with basic and cluster, third with basic and homology features and fourth with all three sets of features. A random forest algorithm was selected for training because of its robustness over unbalanced data set and prevention of overfitting. All these models were first trained with all the features listed in Supplementary table 3-2. However, the relative feature importance of each feature in the fourth model trained with all the sets depicted lower to no importance for many of homology and basic features (Supplementary figure 3-3). Hence, we selected a smaller number of features from these three sets considering their relative feature importance was more than 5%. The final set of features used for training are mentioned in Figure 3-2A.

The four trained classifier were tested with two approaches. Firstly, 5-fold cross-validation repeated 10 times was used to test models and performance was measured with area under the curve (AUC) in ROC curve. Secondly, sensitivity or validation rate of models were compared with two independent test data sets (not used for training). The first test data included PCR validated SKBR3 SVs predicted from Pacific Biosciences long reads (87) and validated with PCR, and second test data included MCF7 SVs, and a primary breast tumor SVs validated with linked-reads sequencing (common SVs predicted with both Illumina short-reads and 10X Genomics linked-reads sequencing pipeline in (72)).

Each random forest classifier was trained with 500 trees and 10 nodesize with randomForest package in R. The probability of true SV was measured as $P(Y=\text{True SV} \mid X = x_i)$ where x_i is set of features for training. Calls with probability>0.5 were considered true and used for the measurement of performance.

3.4.6 Liposarcoma samples

31 WGS liposarcoma samples comprising of 21 DDLS, 8 MLS and 2 WDLS samples were obtained from University of Medical Centre, Mainz. The DNA from tumor (stored in FF or FFPE) was extracted and whole genome sequenced with Illumina technology with short insert size of 400-500bp and paired-end reads of 2X151bp length. The average coverage of tumor samples in DDLS cohort was 74.5X, 77.4X in MLS and 74X in WDLS samples. Since the paired normal tissue for each tumor sample was also sequenced, the somatic SV calls were predicted by FuseSV pipeline. The fastq files of each sample was processed as mentioned above and SVs with probability>0.7 were considered as true.

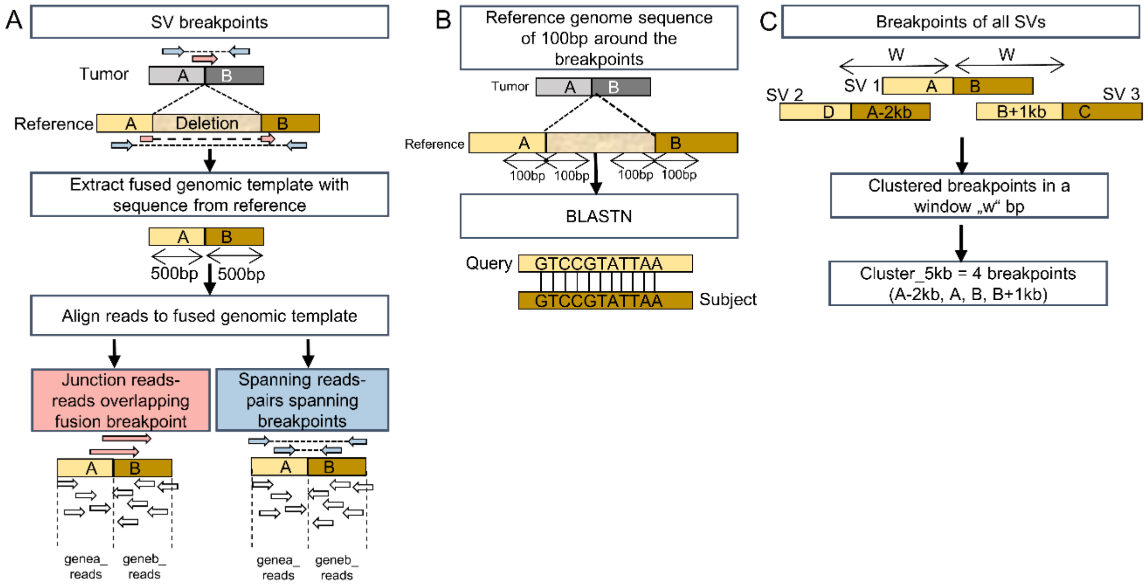
3.4.7 Shattering of chromosome

The CGR with complex SVs were calculated with svpluscnv (77) package in R. The frequency of CNV distribution (Figure 3-3B) in the liposarcoma samples, circos plot depicting shattered chromosomes (Figure 3-4B and Supplementary figure 3-6) and shattered chromosome frequency in DDLS samples (Figure 3-4C) was plotted with the same package.

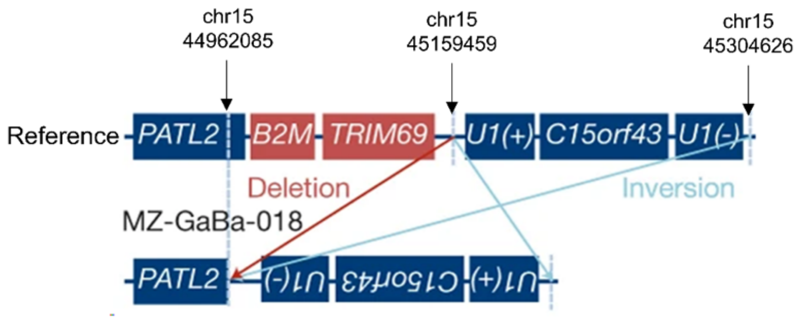
3.4.8 Statistical analysis and graphs

The statistical analysis and assessment of trained models was performed in Rv3.6. The graphs were plotted with ggplot2 (91) package in R.

3.5 Supplementary figures

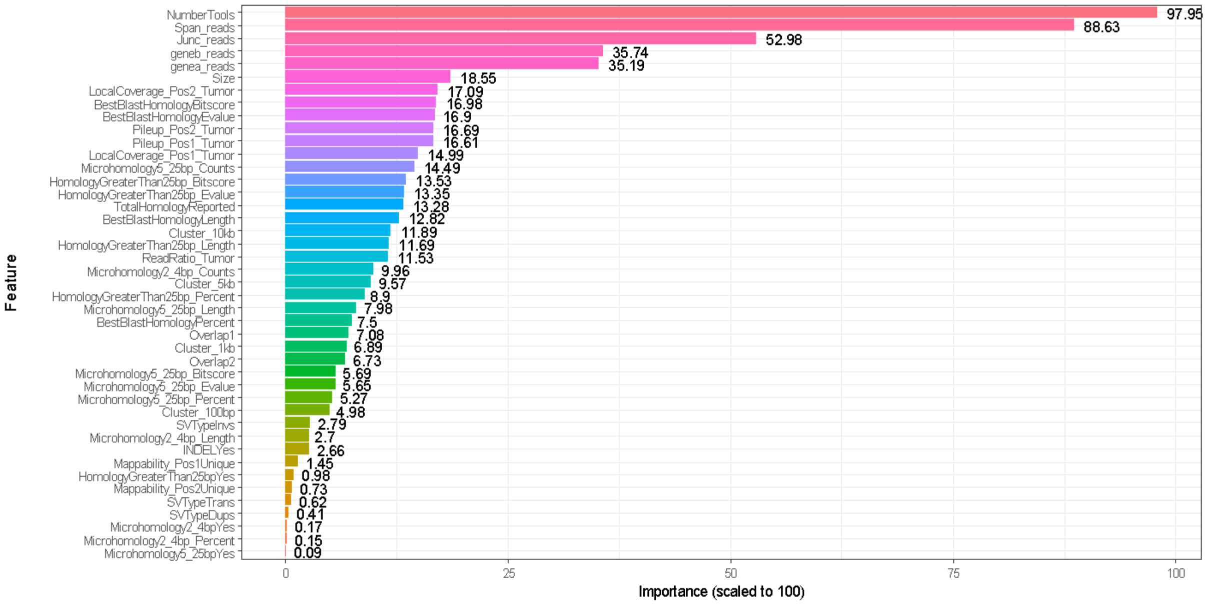

A
SV breakpoints
Tumor
A
B
Reference
A
Deletion
B
Extract fused genomic template with sequence from reference
A
B
500bp
500bp
Align reads to fused genomic template
Junction reads- reads overlapping fusion breakpoint
Spanning reads- pairs spanning breakpoints
A
B
A
B
genea_ reads
geneb_ reads
genea_ reads
geneb_ reads
B
Reference genome sequence of 100bp around the breakpoints
Tumor
A
B
Reference
A
B
100bp
100bp
100bp
100bp
BLASTN
Query
GTCCGTATTAA
GTCCGTATTAA
Subject
C
Breakpoints of all SVs
W
W
SV 1
A
B
SV 2
D
A-2kb
B+1kb
C
SV 3
Clustered breakpoints in a window „w" bp
Cluster_5kb = 4 breakpoints (A-2kb, A, B, B+1kb)

Supplementary figure 3-1: Representation of basic/requantification, homology and cluster features of FuseSV pipeline. A) The requantification pipeline for resolution of breakpoints of overlapping SV calls using junction reads. Other basic features derived from requantification includes spanning reads, genea_reads (number of reads mapping on segment 1) and geneb_reads (number of reads mapping on segment 2). B) Homology features derived from BLASTN of query sequence (left segment of merged genomic templates) and subject sequence (right segment of merged genomic templates). A 200bp of sequence around each breakpoint of each SV was utilized for local sequence alignment by BLASTN. C) Cluster features calculated by counting number of breakpoints of different SVs that lie in a certain window. The figure shows clustered breakpoints in a 5kb window and the derived Cluster_5kb feature.

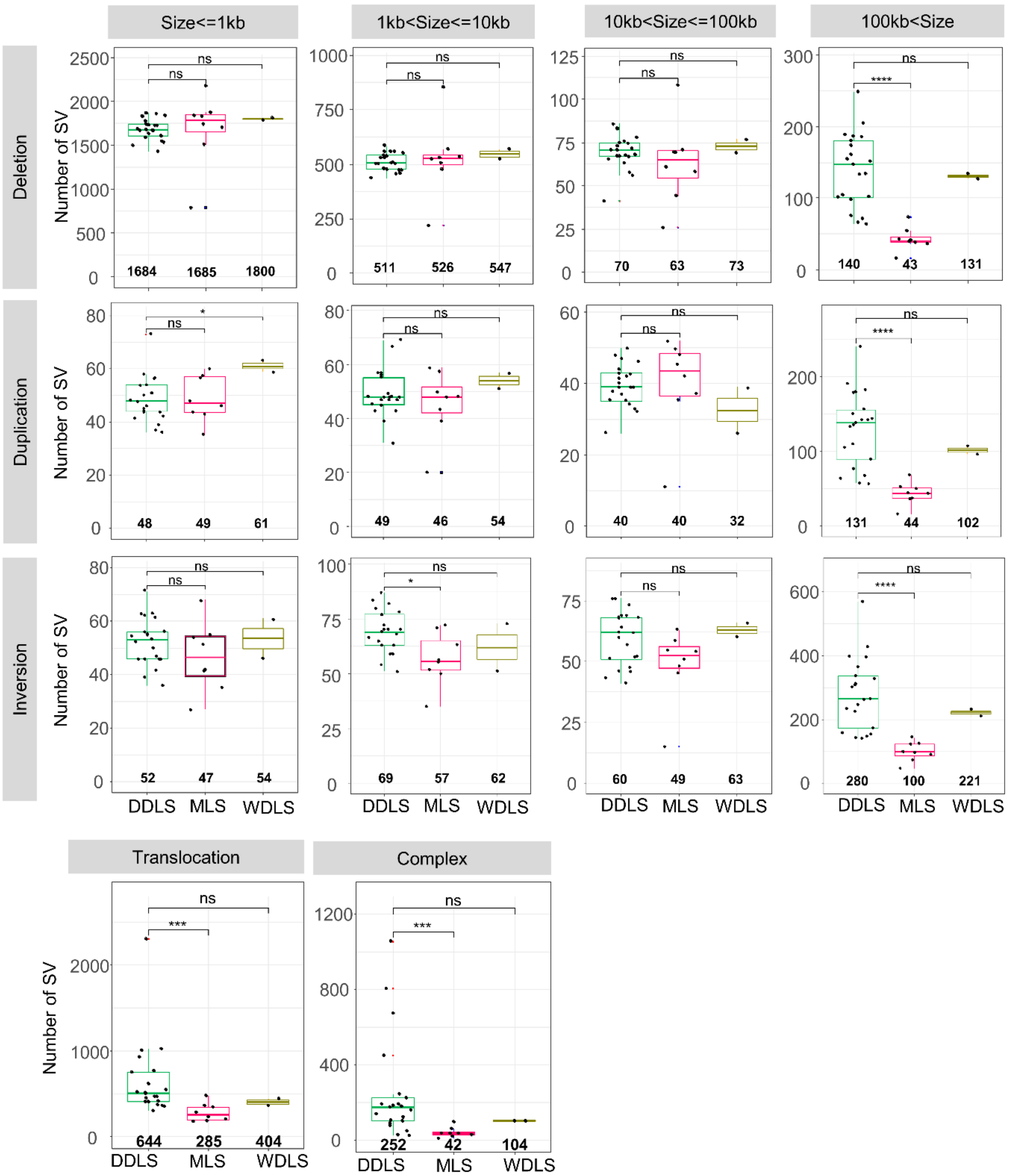


Chrom1	Pos1	Chrom2	Pos2	SVType	Delly	Lumpy	SvABA	Manta	BIC-Seq2
Chr15	44962085	Chr15	45304626	Invs (3to3)	Chr15:44962085-45304628: Invs (3to3)	Chr15:44962085-45304626: Invs (3to3)	No	No	Chr15:44962095-45159459: Dels

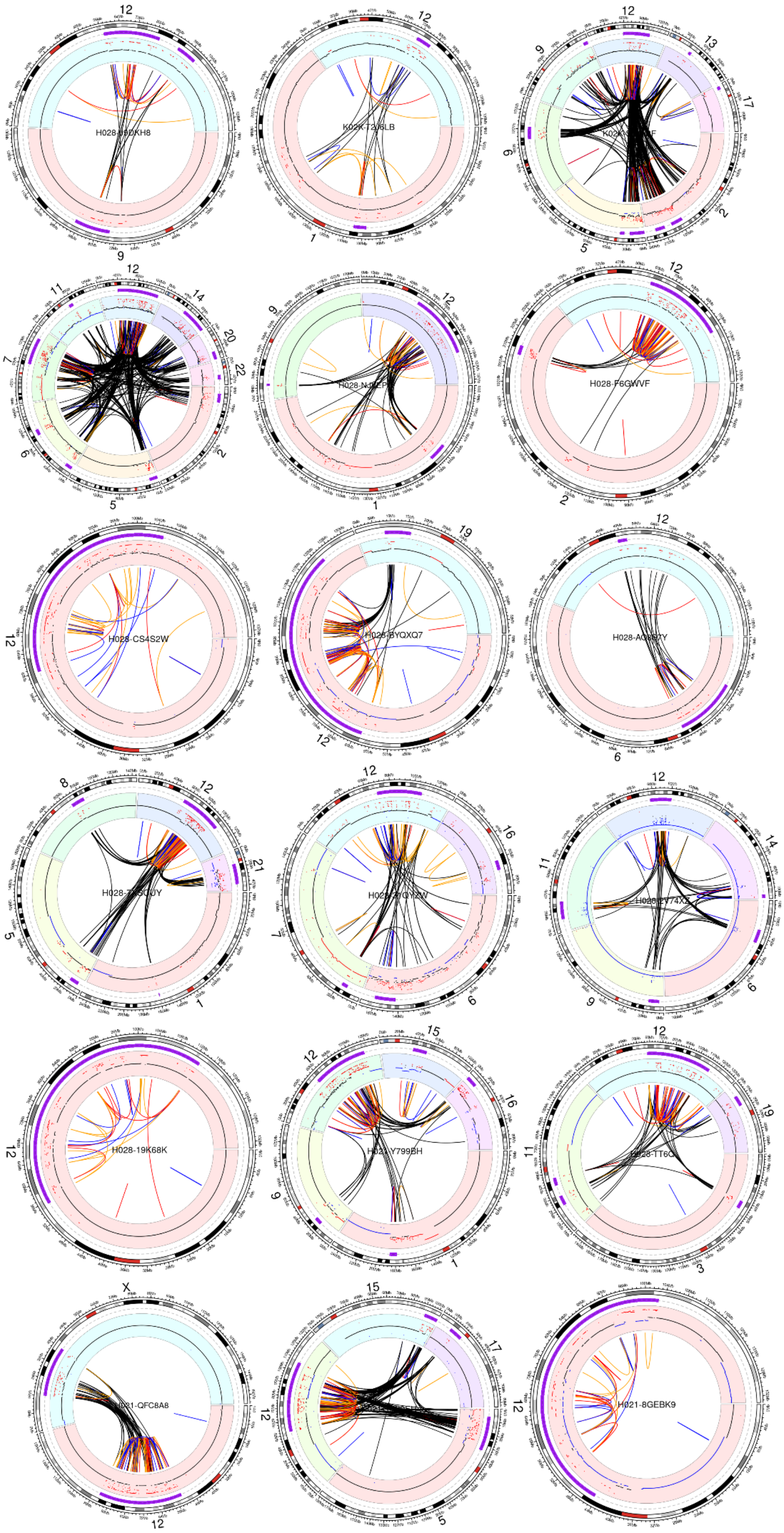
Supplementary figure 3-2: Nested SV in MZ-GaBa-018 cell line with back-to-back deletion and inversion event. The figure was adapted from (92). It depicts the genomic coordinates of deletion and inversion occurring together in the MZ-GaBa-018 cell line and the call predicted by ensemble of tools in FuseSV. Such nested SVs were also used for training the FuseSV random forest classifiers.



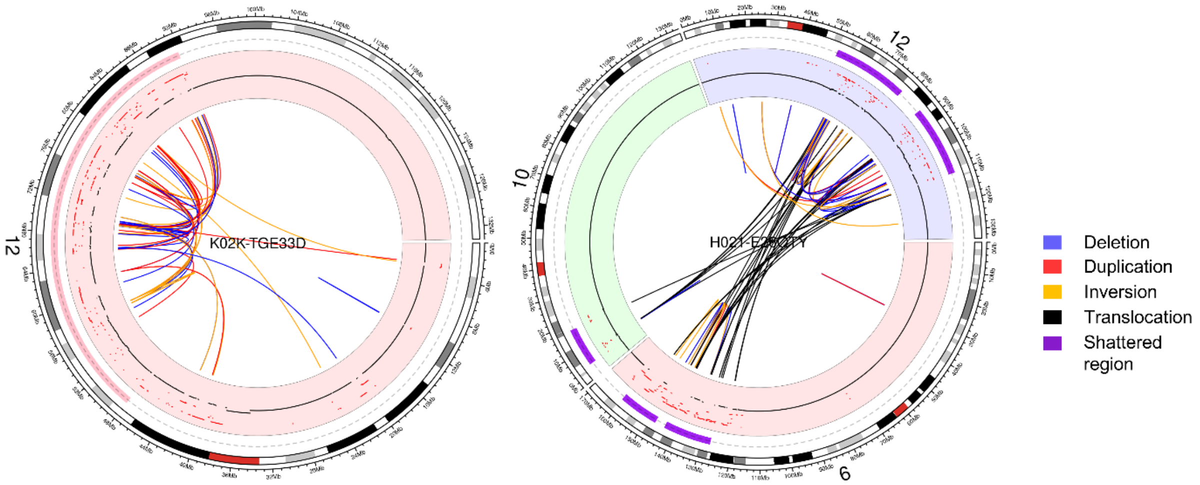
Supplementary figure 3-3: The relative importance of features in trained random forest classifier. These include relative importance of all tested features in three sets (basic, homology and cluster) in the trained classifier with all the features (Basic+Homology+Cluster). This was obtained with varImp function random forest in R.



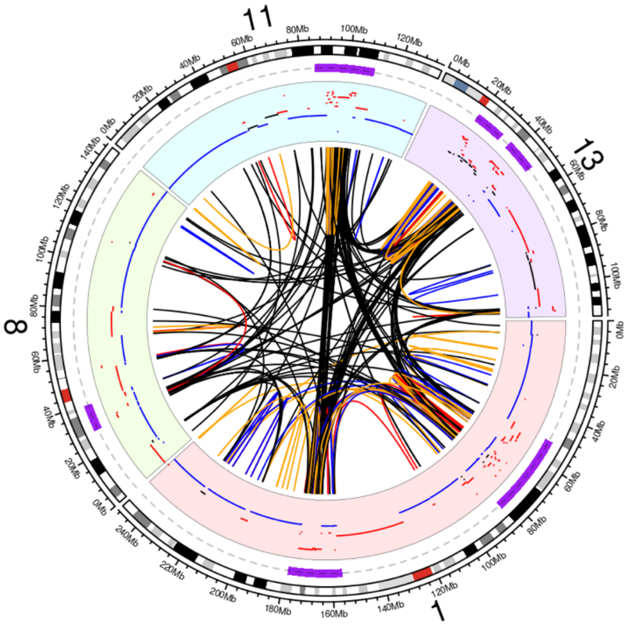
Supplementary figure 3-4: Landscape of different sizes of SVs in three sub-types of liposarcoma (DDLS, MLS and WDLS). The plot depicts number of SV of different types (Deletion, Duplication, Inversion, Translocation, Complex SV) with the mean number of SVs mentioned in bold at the zero coordinate. All the events with more than 10 SV breakpoints clustered in a 5kb window are categorized as complex. The statistics for difference in number amongst sub-types of liposarcoma were derived with Kruskal-Wallis test with p-values as: ns-not significant, *-value <0.05, **-value <0.01, ***-value <0.001, ****-value <0.0001.



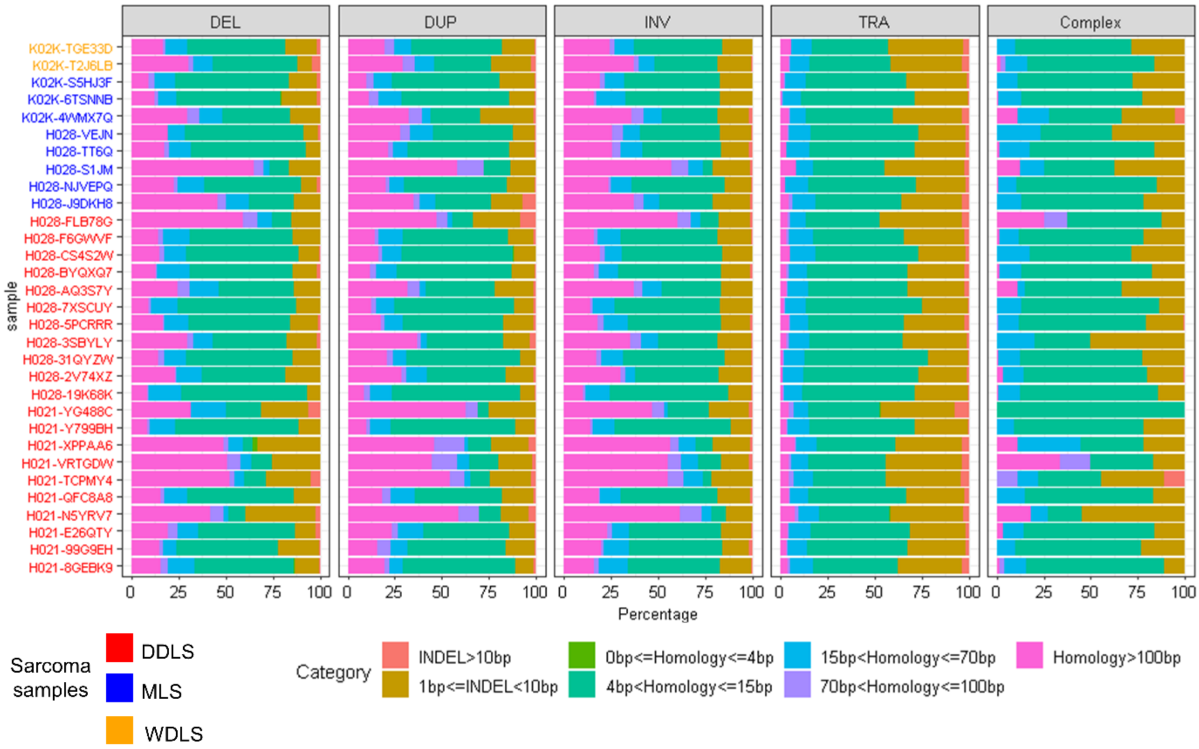
Supplementary figure 3-5: The shattering of chromosome 12 in DDLS samples. The circos plot generated with svpluscnv package in R depicts different SVs and chromosomes involved in complex SVs in DDLS samples. The figure shows plots for following samples: H028-J9DKH8, K02K-T2J6LB, K02K-S5HJ3F, K02K-6TSNNB, H028-NJVEPQ, H028-F6GWVF, H028-CS4S2W, H028-BYQXQ7, H028-AQ3S7Y, H028-7XSCUY, H028-31QYZW, H028-2V74XZ, H028-19K68K, H021-Y799BH, H028-TT6Q, H021-QFC8A8, H021-99G9EH and H021-8GEBK9.



Supplementary figure 3-6: The shattering of chromosome 12 in WDLS samples. The circos plot generated with svpluscnv package shows shattering of chromosome 12 in WDLS samples that included complex SVs. The peak region shattered in both these samples were: 68012374-78012374 and 90012374-100012374 on chromosome 12.



Supplementary figure 3-7: The shattering of chromosomes in K02K-4WMX7Q MLS sample. Only one sample in MLS cohort had shattering of chromosomes that involved chromosome 1, 8, 11 and 13.



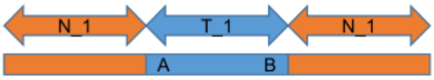
Supplementary figure 3-8: Majority of large sized SVs (size>100kb) are formed with DNA repair mechanisms utilizing long stretches of homology around the breakpoints. The plot depicts percentage of SVs with size greater than 100kb or translocation and complex events with variable length of homology and INDEL around the breakpoints amongst different type of SVs.

3.6 Supplementary tables

Supplementary table 3-1: The list of WGS samples and validated SVs used for training and testing FuseSV pipeline. The table mentions different WGS samples, the source of next generation sequencing reads, source of validated SVs in the respective samples, their validation techniques and the strategy the validated calls were used for testing FuseSV.

S.No.	Sample	WGS source	Validated SVs source	FuseSV testing	Validation technique
1	MCF7	doi:10.1371/journal.pcbi.1008397	doi:10.1371/journal.pcbi.1008397	Cross-validation	PCR
2	MCF7		doi:10.1371/journal.pcbi.1008397	Test data 2	10X Genomic linked-reads sequencing
3	MCF7		doi: 10.1016/j.cels.2016.05.007	Cross-validation	Optical mapping
4	MCF7		doi: 10.1101/gr.113555.110	Cross-validation	PCR
5	SKBR3	doi:10.1101/gr.231100.117	doi:10.1101/gr.231100.117	Test data 1	PacBio long reads and PCR
6	Primary breast tumor	doi:10.1371/journal.pcbi.1008397	doi:10.1371/journal.pcbi.1008397	Test data 2	10X Genomic linked-reads sequencing
7	Mz-GaBa-018	Sequenced with 33.6X coverage, 2X101 read length	In-house	Cross-validation	PCR
8	4 ICGC primary breast tumor (PD4088, PD4116, PD4107, PD4103)	https://doi.org/10.1038/nature17676 EGAS00001000161	https://doi.org/10.1038/nature17676	Cross-validation	PCR

Supplementary table 3-2: Explanation of different features included in three set of features (basic, homology and cluster) for training the random forest classifiers.

Basic features		
1.	NumberTools/Concordance	Number of tools predicting same SV (when breakpoints overlap within 500bp).
2.	Span_reads	Number of spanning read pairs derived from requantification.
3.	Junc_reads	Number of junction reads derived from requantification.
4.	genea_reads	Number of reads aligned to segment 1 of merged genomic template in requantification.
5.	geneb_reads	Number of reads aligned to segment 2 of merged genomic template in requantification.
6.	LocalCoverage_Pos1_Tumor	Average coverage in 200bp window around the position 1.
7.	LocalCoverage_Pos2_Tumor	Average coverage in 200bp window around the position 2.
8.	Pileup_Pos1_Tumor	Number of reads at the position 1 of SV (calculated using samtools pileup).
9.	Pileup_Pos2_Tumor	Number of reads at the position 1 of SV (calculated using samtools pileup).
10.	Size	Size of SV
11.	ReadRatio_Tumor	Average read coverage between breakpoints of SV in tumor sample: T_1/((N_1+N_2)/2) 
12.	Overlap1	Percentage overlap of SV with copy number variant predicted by BIC-Seq2.
13.	Overlap2	Percentage overlap of copy number variant predicted by BIC-Seq2 with SV.
14.	SVType	Type of SV (Dels, Dups, Invs, Trans)
15.	Mappability_Pos1	Unique or ambiguous mappability of position 1 of SV
16.	Mappability_Pos2	Unique or ambiguous mappability of position 2 of SV
17.	INDEL	Presence of INDEL around the breakpoints.
Homology features		

1.	BestBlastHomologyLength	Maximum length of query sequence (left segment) that aligns with subject sequence (right segment) around the breakpoints of SV.
2.	BestBlastHomologyPercent	Maximum percentage of query sequence (left segment) is identical to subject sequence (right segment) around the breakpoints of SV.
3.	BestBlastHomologyEvalue	e-value reported by BLAST for query sequence’s similarity to subject sequence that is not by chance.
4.	BestBlastHomologyBitscore	Maximum bit score reported by BLAST for all the possible alignments of query and subject sequence around the breakpoints.
5.	TotalHomologyReported	Total number of possible alignments of query and subject sequences reported by BLAST.
6.	HomologyGreaterThan25bp	Whether length of alignment between query and subject sequence is greater than 25bp (Yes or No).
7.	HomologyGreaterThan25bp_Length	The length of alignment between query and subject sequence around the breakpoints, if length is greater than 25bp.
8.	HomologyGreaterThan25bp_Percent	Maximum percentage of identity between query and subject sequence if length of similarity is greater than 25bp.
9.	HomologyGreaterThan25bp_Evalue	Best evalue reported by BLAST between query and subject sequence at the breakpoints if length is greater than 25bp.
10.	HomologyGreaterThan25bp_Bitscore	Best Bit score reported by BLAST between query and subject sequence at the breakpoints if length is greater than 25bp.
11.	Microhomology5_25bp_Counts	Number of alignments reported by BLAST when length of identical sequence between query and subject sequence is between 5 to 25bp.
12.	Microhomology5_25bp_Length	Maximum length of identical query and subject sequence that are identical, if the length is between 5-25bp.
13.	Microhomology5_25bp_Percent	Maximum percentage of sequence identical between query and subject sequence, if the length of this sequence is between 5-25bp.
14.	Microhomology5_25bp_Evalue	e-value of selected aligned query and subject sequence, if length is between 5-25bp
15.	Microhomology5_25bp_Bitscore	Bitscore of selected aligned query and subject sequence, if length is between 5-25bp.

16.	Microhomology2_4bp_Counts	Number of identical sequence in query and subject sequence around the breakpoints that are 2 to 4 bp in length.
17.	Microhomology2_4bp_Length	Maximum length of identical sequence in query and subject sequence around the breakpoints that are 2 to 4 bp in length.
18.	Microhomology2_4bp_Percent	Maximum percentage of identical sequence in query and subject sequence around the breakpoints that are 2 to 4 bp in length.
Cluster features		
1.	Cluster_100bp	Number of breakpoints clustered within 100bp of SV breakpoints.
2.	Cluster_1kb	Number of breakpoints clustered within 1kb of SV breakpoints.
3.	Cluster_5kb	Number of breakpoints clustered within 5kb of SV breakpoints.
4.	Cluster_10kb	Number of breakpoints clustered within 10kb of SV breakpoints.

Supplementary table 3-3: Number of different type and size of SVs in liposarcoma samples. The table tabulates number of SV events of different types (DEL: deletion, DUP: duplication, INV: inversion, TRA: translocation, COMPLEX: Complex SVs events with more than 10 breakpoints clustered in 5kb window) and different sizes (Bin 1: Size<=1kb, Bin 2: 1kb<Size<=10kb, Bin 3: 10kb<Size<=100kb, Bin 4: Size>100kb).

Sample	DEL				DUP				INV				TRA	COMPLEX
	Bin1	Bin 2	Bin 3	Bin 4	Bin 1	Bin 2	Bin 3	Bin 4	Bin 1	Bin 2	Bin 3	Bin 4		
H021-8GEBK9	1869	560	83	161	44	57	43	136	46	77	68	313	451	108
H021-99G9EH	1736	539	67	147	54	39	37	138	55	70	69	248	623	676
H021-QFC8A8	1728	509	68	134	51	48	50	142	50	59	57	263	472	245
H028-TT6Q	1588	479	71	98	50	47	44	110	39	54	46	225	516	175
H028-VEJN	1536	475	68	103	36	48	32	90	36	78	52	173	352	28
H021-Y799BH	1673	504	76	249	51	47	42	240	63	82	69	570	767	195
H028-19K68K	1432	457	58	205	37	31	39	183	47	63	46	429	417	126
H028-2V74XZ	1501	436	41	101	48	48	26	64	54	66	43	157	755	141
H028-31QYZW	1608	474	72	135	39	43	39	142	46	68	62	265	1022	224
H028-3SBYLY	1673	557	63	66	73	49	35	57	53	70	41	140	299	28
H028-5PCRRR	1837	552	78	151	42	69	46	144	56	70	68	327	552	161
H028-7XSCUY	1669	502	71	189	48	43	38	191	46	80	76	398	925	451
H028-AQ3S7Y	1551	459	56	64	44	47	33	57	46	51	52	153	363	51
H028-BYQXQ7	1639	476	75	187	46	48	47	180	55	72	64	364	410	185
H028-CS4S2W	1782	540	67	180	45	56	40	150	62	87	62	309	509	93
H028-F6GWVF	1858	560	74	185	56	48	34	156	61	59	73	394	470	181
H028-J9DKH8	1728	502	67	72	57	67	43	68	42	59	48	147	377	102
H028-NJVEPQ	1690	502	65	104	41	45	35	105	52	83	51	233	521	192

K02K-6TSNNB	1741	585	83	174	44	55	41	178	71	69	68	336	2305	1055
K02K-S5HJ3F	1832	529	86	155	54	57	48	133	63	65	76	301	1007	804
K02K-T2J6LB	1684	529	73	76	58	45	39	77	56	63	60	143	405	79
H021-N5YRV7	1840	522	61	43	44	59	52	53	54	52	55	100	283	38
H021-VRTGDW	1834	531	70	55	57	50	50	45	68	71	59	123	366	38
H021-XPPAA6	1881	570	71	39	60	48	42	50	55	72	51	126	343	40
H028-S1JM	1706	532	58	37	46	48	37	43	35	63	54	90	202	32
H021-TCPMY4	2181	854	108	42	43	39	48	37	42	50	48	96	231	100
H021-YG488C	786	219	26	16	48	20	11	16	27	35	15	46	184	9
H028-FLB78G	1514	478	44	39	35	43	35	36	41	56	45	73	188	20
K02K-4WMX7Q	1740	506	69	74	57	57	45	68	51	55	63	143	482	63
H021-E26QTY	1812	571	77	127	59	57	39	96	61	73	66	211	445	104
K02K-TGE33D	1789	523	69	135	63	51	26	107	46	51	60	231	362	105

Supplementary table 3-4: List of PCR validated data points used for training the random forest classifier. Available at gitlab:

https://gitlab.rlp.net/tron/FuseSV/-/blob/master/Supplementary/SupplementaryTable4.xlsx

4 Direct prediction of neo-antigens from somatic SVs

4.1 Introduction

The structural variations (SVs) causing rearrangements of the genome alter the order of functional sequences (like regulatory regions containing promoter, enhancer etc.) as well as the transcript elements (like exons, introns, untranslated regions etc.). This rearrangement of elements can affect expression of genes via modification of regulatory regions, create altered mRNA transcripts (AMTs) via deletion of splice sites, and generate chimeric mRNA transcripts (CMTs) via fusion of open reading frames of two genes. The AMTs and CMTs derived from somatic SVs are a source of neoantigens that can activate the immune system against tumor cells in a cancer patient (47) and can be targeted in the personalized cancer treatments of that patient (48,93). Moreover, the recurrent CMTs from fusion genes in a sub-type of cancer might be ideal drug targets. One of the established ones is the BCR-ABL1 fusion transcript that is targeted with the kinase inhibitors- dasatinib, imatinib and ponatinib in chronic myeloid leukaemia patients (94).

The integrated analysis of WGS and RNA sequencing (RNA-seq) in cancer cohorts have revealed a diverse landscape of the CMTs generated by various mechanisms from SV's breakpoints both within the gene and intergenic region (95). After the genomic rearrangement, an additional layer of diversity in CMTs is contributed by the process of RNA splicing of intron in pre-mRNA that relies on functional splicing sites located in the intron. While detection of CMTs from RNA-seq data allows the direct identification of all forms of expressed transcripts, their detection from SVs in WGS data requires application of transcription and splicing rules to infer correct order of transcript elements. In this study, we focus on identification of three forms of CMTs i.e., direct fusion transcripts, intron-retained (IR) and intergenic region retained (INR) transcripts (Figure 4-1). A direct fusion transcript comprises of CMTs formed due to fusion of two annotated genes/genomic elements where the breakpoints of SV can be in intergenic and/or within gene. This also covers the CMTs formed as a part of classical fusion genes where the breakpoint of SVs within the intron of two different genes can combine functional splicing sites of the respective genes and cause fusion of their exon boundaries (as depicted in direct fusion transcripts in Figure 4-1). However, when one of the breakpoints is within an exon while other is in intron, the nearest functional splicing site in intron is unavailable. In such cases, the splicing machinery can either utilize the next proximal splicing sites leading to exon skipping or inefficient splicing leading to IR fusion transcripts (Figure 4-1). It is also possible for the splicing machinery to utilize an alternate splice site in the intron sequence. While these events can occur within a gene giving rise to an AMTs, this study explores the creation and expression of CMTs with fusion of two genomic elements. In addition to direct and IR fusion transcripts, we explore CMTs derived from SVs with one breakpoint within an intergenic region while second breakpoint is within annotated gene. Similar to the scenario of intron-retention, fusion of gene with an intergenic region might prevent proper splicing and lead to a shorter AMT or a transcript containing sequence of an intergenic region. Such cases lead to formation of the CMTs with intergenic region (INR). However, this scenario an intergenic region could by chance harbour an element that serves as an alternate splice site, but we do not consider such cases in this study.

The bioinformatics tools identifying the classical fusion genes with WGS and RNA-seq achieve higher accuracy in comparison to the ones utilizing only RNA-seq data. Such tools integrate information from these two sequencing modalities in different settings. For example, INTEGRATE utilizes RNA-seq data for prediction of the fusion mRNA transcripts and subsequently find genomic breakpoints from paired WGS data to support that fusion (96). On the other hand, nFuse (97) identifies complex genomic rearrangements (CGR) from WGS and use RNA-seq data for support. These existing tools concentrate only on the classical fusion genes.

In order to explore three categories of above-mentioned CMTs (direct/classical fusion gene, intron-retained and intergenic region retained fusion transcript), we propose a computational pipeline called FUdGE (FUsion of GEnomic segments). The novelty of our approach is inference of the final transcript makeup with information from the detected SVs and annotated genomic elements. Subsequently, RNA-seq data is used to prove expression of predicted CMTs on mRNA level. Focusing on the somatic SVs predicted from WGS data, we apply our pipeline to a liposarcoma cohort containing dedifferentiated liposarcoma (DDLS), myxoid liposarcoma (MLS) and well-differentiated liposarcoma (WDLS) samples and explore the expression of CMTs with paired RNA-seq data.

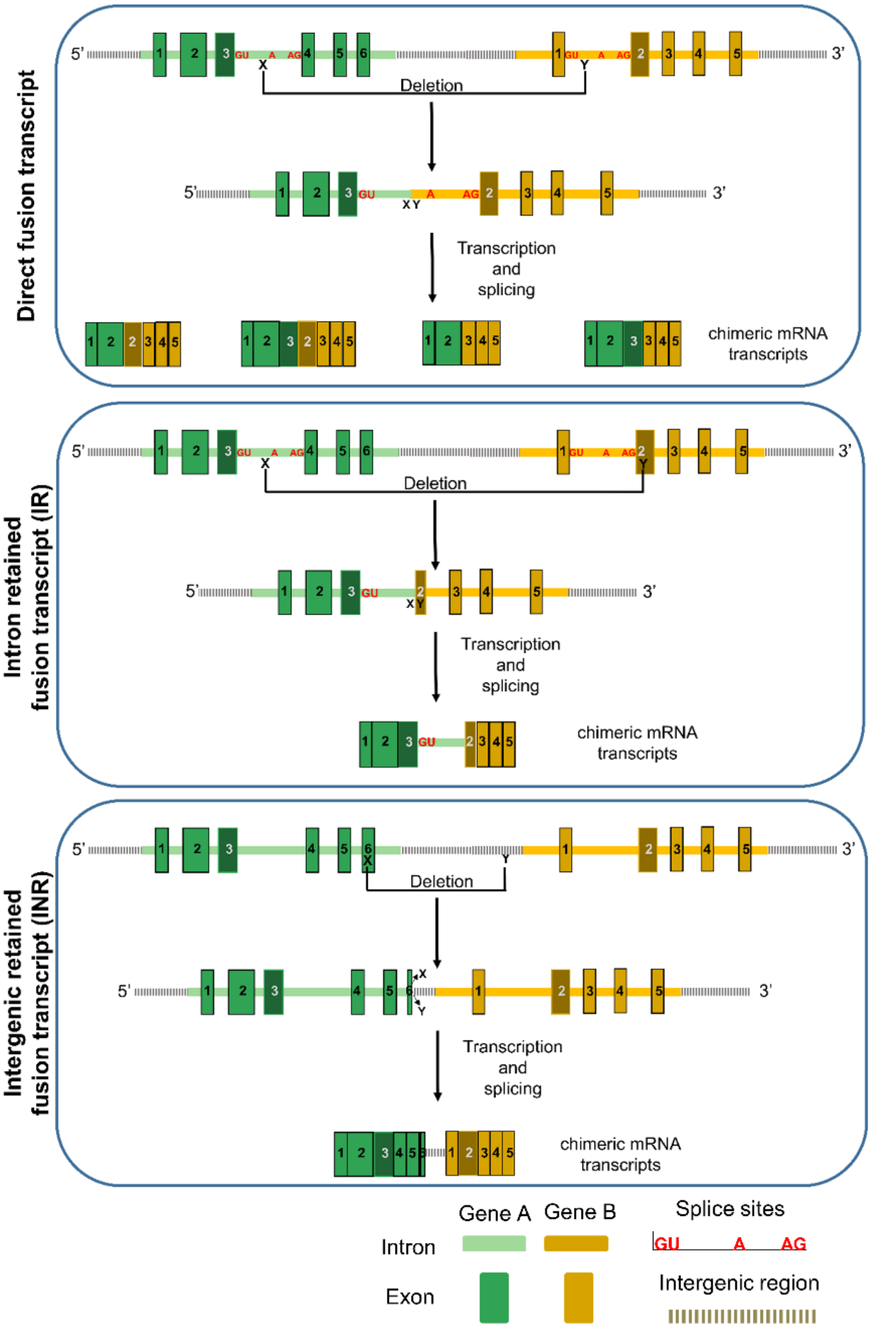
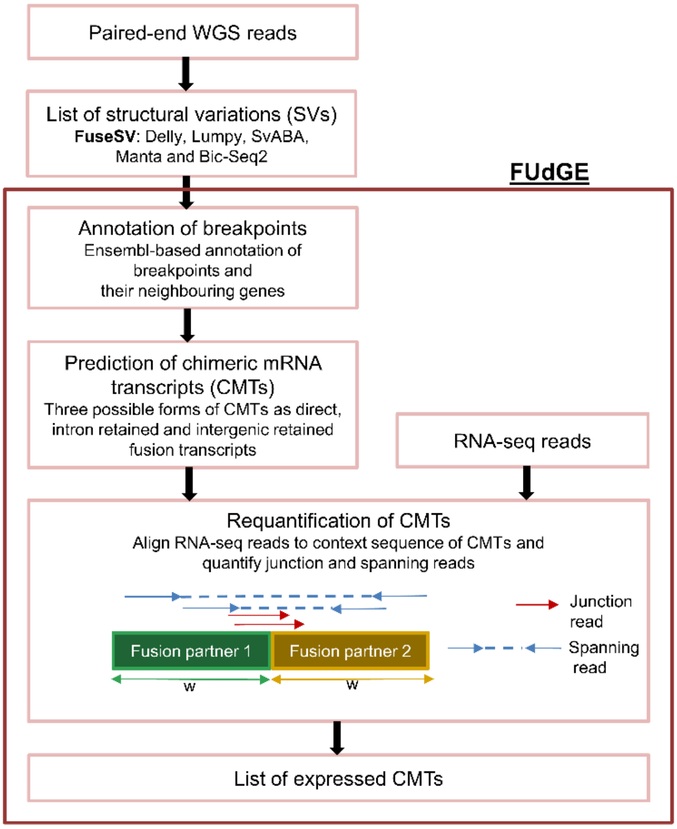

Direct fusion transcript
5'
3'
GU
A
AG
X
Y
Deletion
Transcription and splicing
chimeric mRNA transcripts
Intron retained fusion transcript (IR)
5'
3'
GU
A
AG
X
Y
Deletion
Transcription and splicing
chimeric mRNA transcripts
Intergenic retained fusion transcript (INR)
5'
3'
X
Y
Deletion
Transcription and splicing
chimeric mRNA transcripts
Gene A
Gene B
Splice sites
Intron
Exon
GU
A
AG
Intergenic region

Figure 4-1: Explored categories of the chimeric mRNA transcripts (CMTs) due to an SV event in this study. The direct fusion transcript includes the CMTs generated by an SV (deletion with breakpoints X and Y) causing fusion of annotated exons of two different genes. An SV event can also generate intron-retained CMTs due to the loss of functional splicing site and expression of annotated intron of one gene fused with exon of a different gene (intron retained fusion transcripts). The intergenic region retained fusion transcripts involves expression of unannotated intergenic region with annotated exon of a gene caused due to an SV event.

4.2 Results

4.2.1 FUdGE Scheme



Paired-end WGS reads
List of structural variations (SVs)
FuseSV: Delly, Lumpy, SvABA, Manta and Bic-Seq2
FUdGE
Annotation of breakpoints
Ensembl-based annotation of breakpoints and their neighbouring genes
Prediction of chimeric mRNA transcripts (CMTs)
Three possible forms of CMTs as direct, intron retained and intergenic retained fusion transcripts
RNA-seq reads
Requantification of CMTs
Align RNA-seq reads to context sequence of CMTs and quantify junction and spanning reads
Junction read
Spanning read
Fusion partner 1
Fusion partner 2
w
w
List of expressed CMTs

Figure 4-2: The scheme followed by FUdGE for the prediction of CMTs. First a list of SVs in a sample was collated with the paired-end WGS reads used in the FuseSV pipeline. This is the input to FUdGE where each breakpoint of SVs is labelled with ENSEMBL genome annotations along with exons of nearest annotated genomic sections. The mRNA structure of three different types of CMTs (direct, intron retained and intergenic retained) are predicted and a synthetic genomic template with the fused sequence around a window of 'w' bp in the CMTs is created. Next, the RNA-seq reads from same sample are aligned to this synthetic genomic template and supporting reads in terms of junction and spanning reads are calculated in the requantification step. The possible CMTs with supporting junction or spanning reads are returned as the expressed CMTs.

FUdGE predicts different possible CMTs formed due to the structural rearrangement of the genome. As depicted in Figure 4-2, it requires a list of SVs with its genomic breakpoints, orientation and type (deletion, duplication, inversion and translocation). In this study, we utilized paired-end WGS reads from the respective sample in the FuseSV pipeline for prediction of the SVs (Chapter 3). The FUdGE annotates each breakpoint of SVs with genome annotations from ENSEMBL along with their nearest neighbouring genes and their closest exons. The genome annotations from ENSEMBL include protein-coding genes, pseudogenes and different known non-coding RNA like long non-coding RNA, miRNA etc. and unannotated intergenic regions. With the breakpoints of each SV, the mRNA sequence for three possible CMTs types are predicted. The first type includes direct fusion transcripts where the two nearest neighbouring exons around the annotated genomic section are fused due to an SV event (Supplementary figure 4-1). The other categories predicted by FUdGE includes intron and intergenic regions retained with the annotated exons. The basis for creation of such CMTs is the localization of one breakpoint of SV within an exon of an annotated genomic section and other breakpoint in an intron or unannotated intergenic region to produce intron-retained (IR) CMTs or intergenic region retained CMT (INR) respectively. We suspect higher chances of a retained intron or intergenic sequence within a coding region in a mRNA transcript with such breakpoints. Subsequently, the matched RNA-seq reads are aligned to the predicted sequence for respective CMTs in the requantification pipeline. With this approach, we quantify number of reads specific for the predicted CMTs. Such reads mapping at the junction of fused sections are junction reads and the read-pairs spanning the fused sections are spanning reads (Figure 4-2).

There are multiple biological and technical reasons for FUdGE predicted CMTs to not be confirmed with RNA-seq data. The technical reason include false positive SV predicted by the SV prediction tool that leads to false positive CMTs predicted by FUdGE. Even though the predicted SVs are true, the FUdGE pipeline can also predict incorrect CMTs. From the biological point of view, there is a possibility of a correctly predicted CMT to not be expressed within the sample. Therefore, we cannot completely differentiate between non-expressed CMTs and false predicted CMTs. Nevertheless, CMTs that are formed due to an SV event and are also expressed can certainly give rise to altered proteins that can further be targeted by personalized immunotherapy related interventions. Henceforth, the term expressed or confirmed on RNA are used interchangeably.

4.2.2 Confirmation of FUdGE predicted CMTs by RNA-seq in breast cancer cell line and primary breast tumor sample

The SVs were predicted in the MCF7 breast cancer cell line and a primary breast tumor sample with the FuseSV pipeline (**Chapter 3**), and the FUdGE pipeline predicted three above-mentioned categories of CMTs.

Considering all the SV events from FuseSV, the FUdGE predicted 844469 direct, 770 IR and 1001 INR fusion transcripts in the MCF7 (Figure 4-3). Amongst the predicted direct fusion transcripts in MCF7, 89.2% of SVs had both breakpoints within intergenic region, 9.3% of SVs had one breakpoint within gene and other in intergenic region, and only 1.5% of SVs had both the breakpoints within gene (Supplementary table 4-1). Moreover, amongst all types of CMTs predicted, only 0.1% direct, 1.9% IR and 2.3% INR fusion transcripts were confirmed to be expressed on RNA (with at least one junction or spanning reads). Comparatively, there was higher number of CMTs predicted in the primary breast tumor sample: 2190781 direct, 2267 IR and 3666 INR fusion transcripts. Amongst the predicted direct fusion transcripts in the primary breast tumor, 85% of SVs had both breakpoints within intergenic region, 12.8% of SVs had one breakpoint within gene and other in intergenic region, and only 2.2% of SVs had both the breakpoints within gene (Supplementary table 4-1). Like the numbers expressed in MCF7, 0.1% direct and 2.7% INR fusion transcripts were expressed in the primary breast tumor sample. However, the percentage of expressed IR fusion transcripts (7.7%) was much higher in the primary breast tumor sample than in MCF7.

The distribution of genomic breakpoints of expressed direct fusion transcripts was investigated and is shown in Supplementary table 4-1. Most genomic breakpoints of underlying SVs leading to expressed direct fusion transcripts lie within gene (77.2% in MCF7 and 62.7% in the primary breast tumor sample), followed by cases with both the genomic breakpoints within intergenic regions (13.7% in MCF7 and 25.4% in the primary breast tumor sample) and least cases with one breakpoint within a gene and the other breakpoint in an intergenic region (9.12% in MCF7 and 11.9% in the primary breast tumor sample). Thus, SV with breakpoints in intergenic region can also leading to direct fusion transcripts.

Within different types of expressed CMTs, direct fusion transcripts had the highest confirmation rate by RNA (97% in MCF7 and 91.1% in the primary breast tumor sample). Relatively, a lower number of IR (1.2%) and INR transcripts (1.8%) were confirmed by MCF7 RNA-seq reads, while the primary breast tumor sample had a higher percentage of confirmed IR (5.6%) and INR (3.3%) fusion transcripts.

Next, we explored the probability score of SV events reported by FuseSV that generated expressed CMTs. Most expressed CMTs classes (except IR transcripts in MCF7) had significantly higher FuseSV score compared to the ones not expressed in MCF7 and the primary breast tumor sample (Figure 4-3).

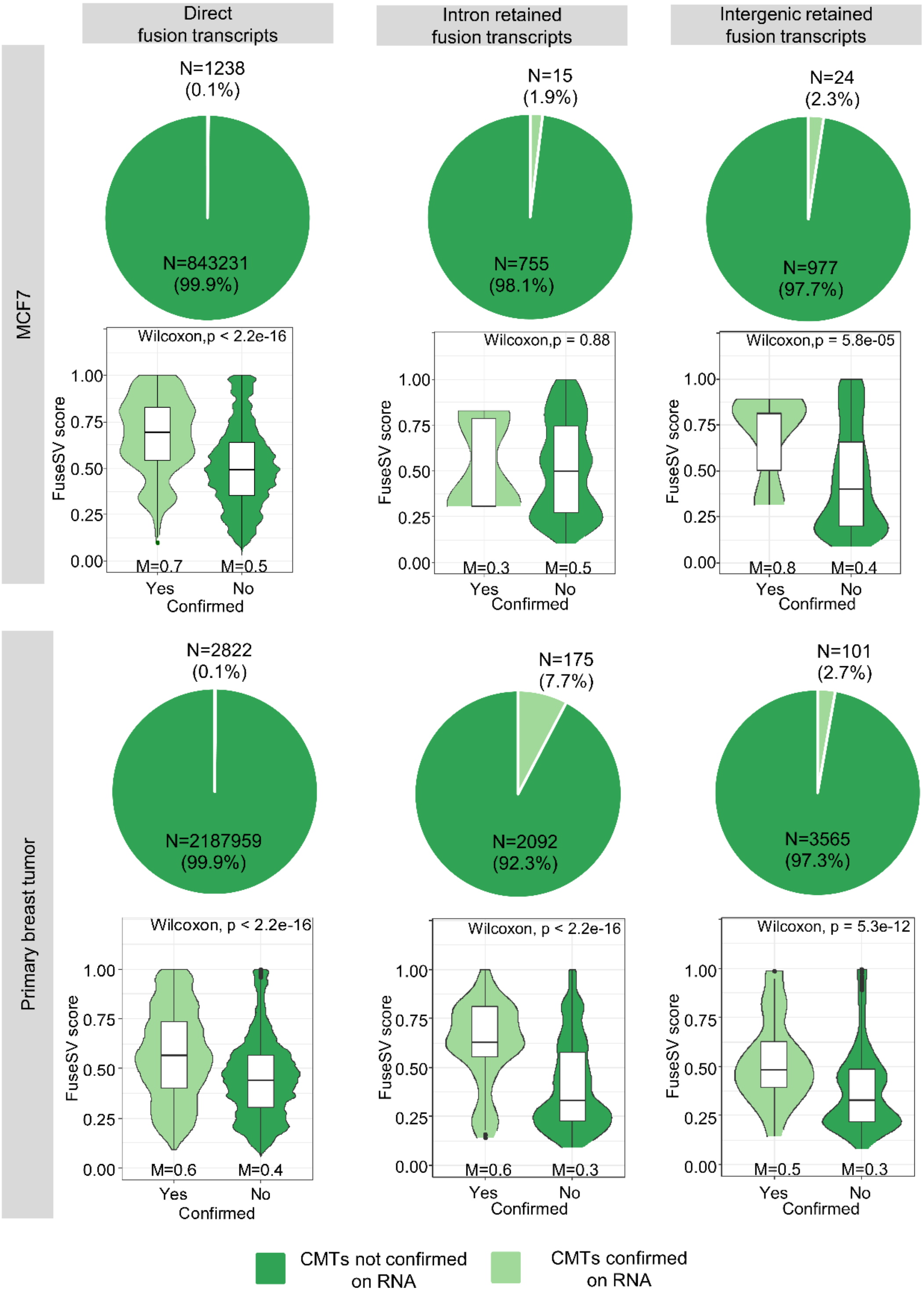


Figure 4-3: Distribution of possible and confirmed chimeric mRNA transcripts (CMTs) in MCF7 and the primary breast tumor sample. The pie plots depict the number (N) and percentage of the CMTs (direct, intron retained and intergenic retained) with an underlying SV predicted by FUdGE and confirmed by RNA-seq reads in the respective samples. The violin plots below the respective pie chart compares the probability score of confirmed CMTs to the not confirmed CMTs. The significance of difference in the median FuseSV score (M) between confirmed CMTs and not confirmed CMTs was estimated with Wilcoxon rank sum test. The p-values for each comparison are marked in the respective violin plots.

This indicates that SVs leading to expressed CMTs are enriched for true positive SVs that received a higher FuseSV prediction score due to confirmation of relevant sequence for CMTs (Supplementary figure 4-2) and sequence for rearranged genomic breakpoints (Supplementary figure 4-3). This was especially true for the direct fusion transcripts where presence of junction reads mapping to the predicted sequence of CMTs indicated that FUdGE predicted precise breakpoints of the fusion. However, the IR and INR fusion transcripts lacked support from junction RNA-seq reads (Supplementary figure 4-2) that merged boundaries of intron with exon or intergenic region merged with exon respectively. These two classes of CMTs rely more closely on the genomic breakpoints that could merge. In case of direct fusion transcript predictions, CMT sequence rely on the boundaries of nearest exon merged. Thus, if a breakpoint of SV is not precise for IR and INR fusion transcript, then a sequence with some difference in the merged breakpoint would be created. And this would not have junction reads mapping to the predicted sequence of CMT. This explains the reason for absence of junction reads for IR and INR fusion transcripts. Such cases were nonetheless considered expressed when a supporting spanning RNA-seq reads was calculated. Interestingly, such transcripts had both junction and spanning reads from WGS data supporting the underlying genomic breakpoints of the SV event (Supplementary figure 4-3).

4.2.3 Confirmation of FUdGE predicted CMTs by qRT-PCR in breast cancer cell line and primary breast tumor sample

Apart from confirming FUdGE predictions with RNA-seq reads, we also checked its performance with a list of qRT-PCR validated classical fusion genes collated from (48). This list of direct fusion transcripts arising from the classical fusion genes were first gathered from various published research studies. Subsequently, they were validated in-house with qRT-PCR/qPCR and published in (48).

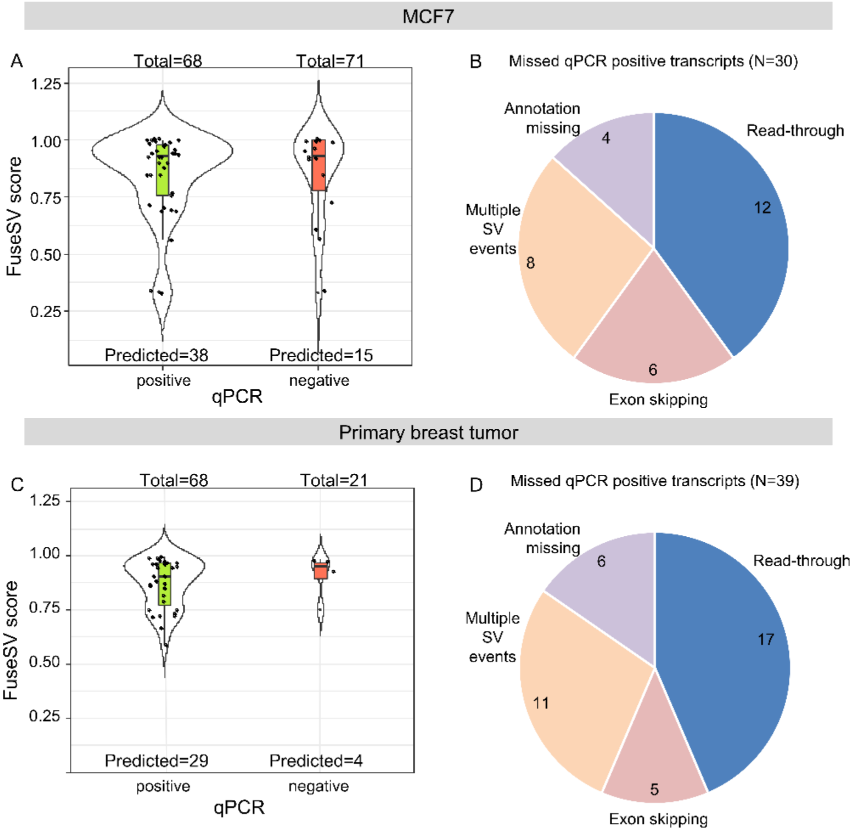


Figure 4-4: The performance of FUdGE on qPCR validated fusion transcripts from classical fusion genes. The plot represents the distribution of FuseSV score for underlying SV events of FUdGE predicted fusion transcripts (Predicted) amongst the ones tested by qPCR. Plot A and C represent the direct fusion transcripts tested in MCF7 and the primary breast tumor sample respectively. The pie charts represent number of qPCR positive fusion transcripts not predicted

with FUdGE. Plot B and D represent the possible reasons for the missing detection of direct fusion transcript by FUdGE in MCF7 and the primary breast tumor sample respectively.

As seen in Figure 4-4, FUdGE predicted ~56% (38 out of 68) and ~43% (29 out of 68) of qPCR positive fusion transcripts in MCF7 and the primary breast tumor sample respectively. These predictions were also confirmed by RNA-seq reads with at least one junction or spanning reads. Moreover, the median FuseSV score of underlying SV events giving rise to confirmed direct fusion transcripts was very high with 0.933 in MCF7 and 0.906 in the primary breast tumor sample. On the other hand, some qPCR negative fusion transcripts were also predicted by FUdGE: ~21% (15 out of 71) in MCF7 and ~19% (4 out of 21) in the primary breast tumor sample. They were also confirmed by RNA-seq data with at least one junction or spanning reads. Moreover, they also had a high FuseSV score (0.932 in MCF7 and 0.952 in the primary breast tumor sample). As seen in Supplementary figure 4-4, no significant difference in junction and spanning RNA-seq reads amongst positive and negative qPCR fusion transcripts was established (except for junction reads in MCF7). This indicates that these are most likely true candidates but were not confirmed by qPCR due to very low expression of transcripts (attributed by low CT value in qPCR). Since FUdGE predicts fusion transcripts directly with an underlying SV event, it is not dependant on expression levels.

Furthermore, we investigated the reasons for qPCR positive fusion transcripts missed by FUdGE. As seen in Figure 4-4B and D, many missed fusion transcripts (40% in MCF7 and 43.5% in the primary breast tumor sample) are fusions of the neighbouring genes. These can likely be attributed to read-through transcription of neighbouring genes that can occur without any underlying SVs. Such cases cannot be predicted by FUdGE due to absence of the genomic footprint. Since the read-through transcription events can also occur in the normal tissue, their exclusion by FUdGE which focus on somatic SV driven CMTs is seen as an advantage with our approach. Nevertheless, we cannot reject the possibility of an SV event missed in the input list of SVs from FuseSV. Nevertheless, the enrichment for neighbouring genes indicates a high rate of read-through transcripts among the missed qRT-PCR confirmed CMTs.

Furthermore, 20% and 12.8% of qRT-PCR confirmed fusion transcripts missed by FUdGE in MCF7 and the primary breast tumor sample respectively, corresponded to a different exon boundary of FUdGE predicted fusion genes. This can be attributed to CMTs with exon skipping. In the future this limitation could be resolved by including additional CMTs with additional neighbouring exon boundaries. The remaining missed qRT-PCR confirmed fusion transcripts included the ones generated from multiple SV events (~27% in MCF7 and 28.2% in the primary breast tumor sample) and difference in annotation of genome amongst various research studies from which the list of validated fusion transcripts was generated (~13% in MCF7 and 15.4% in the primary breast tumor sample). On ignoring the missed cases that lacked genomic footprints (like read-through transcription, alternative splicing and exon skipping) or mis-labelled annotations, FUdGE achieved a validation rate of 82.6% in MCF7 and 72.5% in the primary breast tumor sample.

4.2.4 Analysis of somatic CMTs in liposarcoma cohort

The FUdGE pipeline was applied to 26 liposarcoma samples (18 DDLS, 6 MLS and 2 WDLS) with both WGS and RNA-seq data available from a collaboration[1]. The predictions for somatic SVs were made with FuseSV with a probability score threshold of 0.7 or greater (a slightly stringent criteria). Next, different CMTs were predicted with FUdGE requiring at least 3 junction or spanning RNA-seq reads for confirmation by RNA-seq data.

On an average FUdGE predicted 410, 176 and 279 confirmed direct fusion transcripts in DDLS, MLS and WDLS samples, respectively (Figure 4-5A). The number of confirmed direct fusion transcripts was significantly higher in DDLS than MLS. Moreover, as evident in Figure 4-5B, most genomic breakpoints of underlying SV events for such transcripts were within known genes (on an average 70.8% in DDLS, 74% in MLS and 64.5% in WDLS). Nevertheless, there

---

[1] Collaboration with Prof. Thomas Kindler at University Center for Tumor Diseases, Mainz, Germany

were confirmed direct fusion transcripts with SV breakpoints in intergenic regions (one breakpoint of SV in intergenic region: 17.9% in DDLS, 11.5% in MLS, 19.3% in WDLS; and both breakpoints of SV in intergenic regions: 11.2% in DDLS, 14.5% in MLS and 16.1% in WDLS).

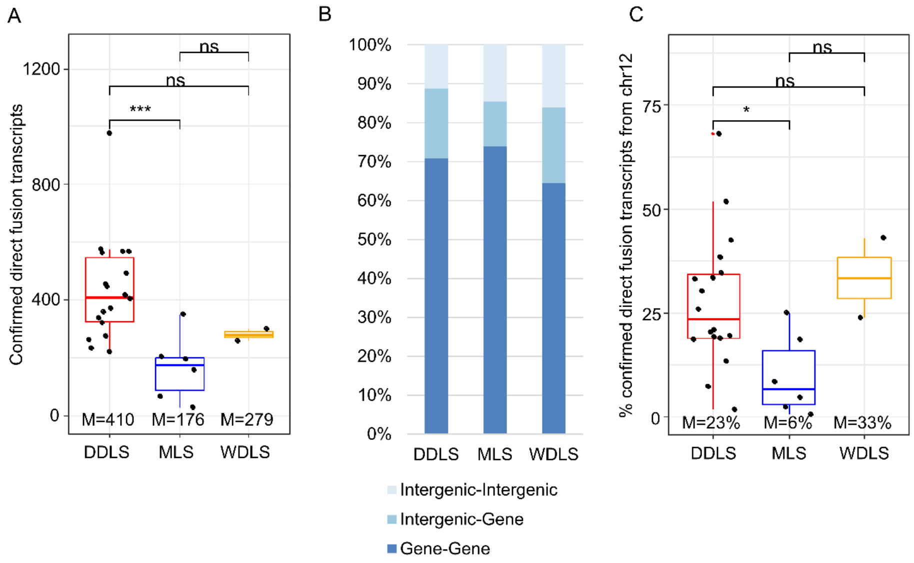


Figure 4-5: Landscape of direct fusion transcripts in the liposarcoma cohort. A) The bar plot depicts the number of expressed direct fusion transcripts within DDLS, MLS and WDLS samples. The median number (M) of expressed fusion transcripts within a sample type is shown. B) The percentage of confirmed direct fusion transcripts with underlying SV breakpoints within a gene or intergenic region in DDLS, MLS and WDLS samples that have breakpoints within a gene or in an intergenic region. C) The fraction of expressed direct fusion transcripts with breakpoints on chromosome 12 is significantly higher in DDLS samples in comparison to MLS and WDLS samples. The median percentage (M) of expressed direct fusion transcripts is shown. The p-values in all these plots are computed derived with the Wilcoxon rank sum test with following significance levels: ns-not significant, *-value <0.05, **-value <0.01, *-value <0.001, ****-value <0.0001.**

Since it is established that chromosome 12q arms in DDLS samples are highly rearranged (64), we investigated the percentage of confirmed direct fusion transcripts originating from chromosome 12. As expected, a significantly higher number of such fusion transcripts from chromosome 12 was reported in DDLS (23%) than in MLS (6%) samples (Figure 4-5C). Within WDLS samples, a high percentage of 33% confirmed fusions from chromosome 12 was observed, but the significance was not established due to a lower number of WDLS samples analysed in this study.

4.2.5 Limited expression of intron and intergenic-retained CMTs in liposarcoma

Somatic intron-retained (IR) and intergenic-region retained (INR) fusion transcripts detected by FUdGE were investigated in the liposarcoma sample cohort (Figure 4-6). The samples classified as DDLS and WDLS had a median number of 13 and 10 confirmed IR transcripts, respectively. Comparatively, MLS samples had a higher median number (M=18) of confirmed IR transcript but a significant difference between different liposarcoma cohorts was not established (Figure 4-6A). The number of confirmed INR transcripts was much lower in comparison to other CMTs. As seen in Figure 4-6B, the median number of events in this category of transcripts was 5, 4 and 3 in DDLS, MLS and WDLS respectively. Moreover, a significant difference in confirmed INR transcripts was established between DDLS and MLS samples.

The frequency of different CMTs confirmed by RNA-seq data in the sub-types of liposarcoma was investigated in Figure 4-6C. The majority of confirmed CMTs were direct fusion transcripts in all sub-types of liposarcoma (95.8% in DDLS, 88.8% in MLS and 95.5% in WDLS). Comparatively, a small percentage of IR transcripts were confirmed in DDLS (3%) and WDLS

(3.4%) samples. However, 9% of IR transcripts were confirmed within MLS samples. Amongst all categories of CMTs, INR fusion transcripts had the lowest confirmation rate by RNA-seq data (1% in DDLS, 2% in MLS and 1% in WDLS samples).

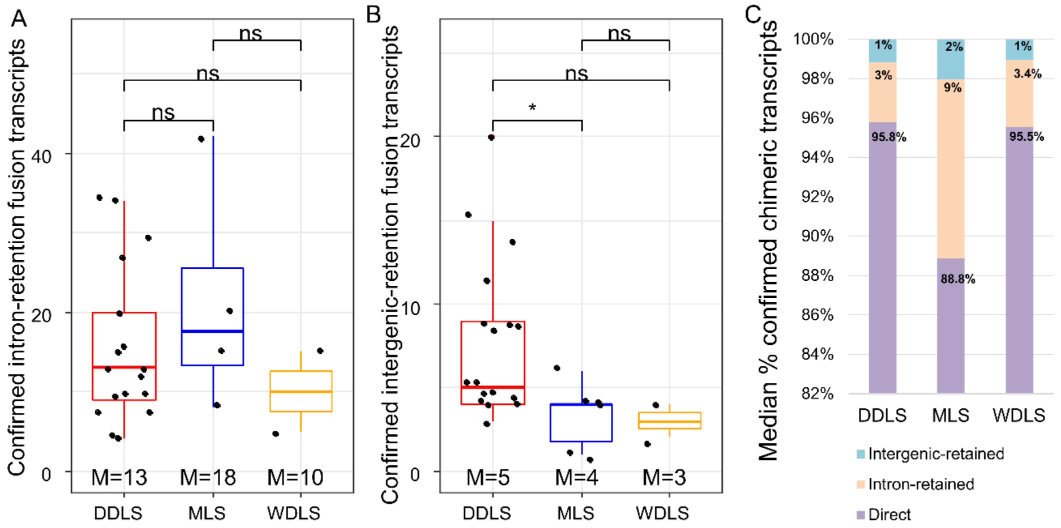


Figure 4-6: Landscape of somatic intron-retained and intergenic-retained fusion transcripts in the liposarcoma cohort. Plot A depicts the distribution of intron-retained fusion transcripts that were confirmed with RNA-seq data in different liposarcoma samples. Plot B depicts the distribution of intergenic-retained fusion transcripts confirmed with RNA-seq data in different liposarcoma samples. For each distribution in plot A and B, median number (M) of confirmed chimeric transcripts is shown and p-values are computed with the Wilcoxon rank sum test with following significance levels: ns-not significant, *-value <0.05, **-value <0.01, *-value <0.001, ****-value <0.0001. Plot C depicts the percentage of median CMTs that were confirmed with RNA-seq data in different liposarcoma sub-types.**

4.2.6 Recurrent CMTs in liposarcoma

Next, the recurrently occurring CMTs was investigated amongst the samples in the DDLS, MLS and WDLS sub-types of liposarcoma. As seen in Supplementary figure 4-5 and Supplementary table 4-2, there were very few recurrent classical fusion genes within the sub-types. Nevertheless, 88.3% of MLS (5 out of 6 samples) were characterized by the fusion gene transcript with *FUS* fused with *DDIT3*. This follows the published studies that report more than 90% of MLS samples with this fusion gene (98,99). The *FUS* gene encodes for multifunctional protein involved in various regulatory pathways like DNA repair, splicing and transcriptional regulation (100). On the other hand, *DDIT3* (also known as *CHOP*: C/EBP homologous protein) is a transcription factor with cellular function as stress sensor that is highly expressed under stress conditions like nutrient deprivation, DNA damage, growth stress etc. (101). The fusion of *FUS* and *DDIT3* produces an oncogenic protein that drives MLS.

Moreover, 61.1% DDLS samples and 100% WDLS samples contained a *CTSC-RAB38* fusion gene. *CTSC* is the Cathepsin C gene that encodes for lysosomal proteinase involved in activation of serine proteinases in the immune cells and *RAB38* is a member of *RAS* family of oncogenes. The *CTSC-RAB38* fusion gene is also characterised in the renal cell carcinoma (102) and brain arteriovenous malformations (103). However, these two genes are the neighbouring genes and other studies characterize it as read-through fusion gene. Here *CTSC-RAB38* was predicted by FUdGE due to a ~320bp deletion in the intergenic region between these two genes. The small somatic deletion event might increase the frequency of read-through transcription between these two gene in WDLS and DDLS samples leading to high recurrence. Nonetheless, we cannot exclude the possibility of the same event occurring in unmutated wild-type cells. Apart from this fusion gene, FUdGE also predicted fusion of paralogous genes like *ADGRE2-ADGRE5*, *ZNF* and the *HLA* related family of genes. This might be an artefact reported by FUdGE considering high sequence similarity (causing alignment artefacts) of the paralogous genes.

Apart from the classical fusion genes that involve fusion between protein-coding genes, we also explored recurrent direct fusion transcripts with annotated genomic sequences apart from protein-coding genes. ~88.9% (16 out of 18) DDLS samples were characterized with a *RP4-592A1.2-AK2* direct fusion gene. The protein-coding gene *AK2* encodes for an adenylate kinase involved in adenine composition while *RP4-592A1.2* is a processed pseudogene whose function is unknown. High expression of the *AK2* gene is implicated in lung cancer (104), potentially suggesting a functional role of this fusion transcript. The fusion between these two genomic segments was also present in 66.6% of MLS samples.

The recurrent IR and INR fusion transcripts are shown in Supplementary figure 4-6 and Supplementary figure 4-7. The IR transcripts with *ADGRE2* and *ADGRE5* fusions were present in 66.67%, 33.3% and 50% of DDLS, MLS and WDLS samples, respectively. Fusions between these two paralogous genes might be artificial and require experimental validation. In case of INR transcripts, a fusion of the *AK2* gene with an unannotated intergenic region was found in 83.3%, 66.6% and 50% of DDLS, MLS and WDLS samples, respectively.

Overall, the liposarcoma samples explored in this study had higher number of direct fusion transcripts than IR and INR fusion transcripts. Apart from *FUS-DDIT3* in MLS samples and CTSC-*RAB38* in DDLS and WDLS samples, there were very few fusion transcripts recurrently present in the sub-types or all the samples.

4.3 Discussion

Here, a novel computation pipeline as FUdGE was presented for prediction of somatic CMTs with an underlying chromosomal rearrangement event. It covers CMTs produced due to fusion of annotated segments of genome along with less studied intron-retained and intergenic-region retained mRNA transcripts. The pipeline used a top-down approach to first detect somatic SVs from WGS data with the FuseSV pipeline, predict different mRNA structures of chimeric transcripts and then check the expression with paired RNA-seq data. This approach demonstrates several advantages. First, the tumor specificity of somatic genomic rearrangements and the generated CMTs. Second, predictions from both WGS and RNA-seq data is more specific in comparison to predictions coming only from RNA-seq data that is confounded with coverage of the cancer transcriptome according to the expression levels. Moreover, this approach cannot account for CMTs generated at RNA level via trans-splicing, cis-splicing (read-throughs, alternative splicing, loss of splice site, exon-skipping etc.) that are not driven by genomic mutation are not necessarily tumor-specific (105). However, one of the disadvantages of this approach is that WGS of tumor and normal tissue with paired RNA-seq is required and analysis of the data is exhaustive in terms of computational resources. Even though FUdGE was successful in predicting the classical fusion genes, following enhancements would be required in future: a) Prediction of somatic CMTs generated due to multiple SV events because of a CGR. Currently FUdGE does not predict CMTs from such cases as the resolution and detection of SVs in CGR require special algorithms and filtering strategy, b) Reduction of false positive CMTs that are predicted due to high sequence similarity of the genomic regions involved in the fusion transcript. Nevertheless, this approach can be a boon for researchers targeting only tumor-specific fusion transcripts.

We also explored the landscape of liposarcoma cohort with above-mentioned classes of CMTs. In terms of distribution of expressed or CMTs confirmed with RNA-seq data in the liposarcoma samples, the highest percentage was represented by direct fusion transcripts (~88-95%) while IR and INR transcripts contributed ~3-9% and ~1-2% respectively. This is in concordance with a recent study that reported highest expression of chimeric transcripts from protein-coding genes in nuclear fraction of HeLa cells (106). Nevertheless, we suspect lower number of IR and INR fusion transcripts reported by FUdGE as the focus in this study was only on somatic SVs derived fusion transcripts while other studies looked at single nucleotide variants (SNVs) derived CMTs. Additionally, it is well known that the IR transcripts are known to be generated by alternative splicing, a factor predominant at RNA level in both tumor and wild-type cells. Thus, a lower number of tumor-specific IR transcripts was confirmed in our study against a published study that considered all IR transcripts predicted in tumor cells that is not necessarily tumor-specific (41). In spite of higher expression of direct fusion transcripts, only MLS samples had a recurrent classical fusion gene between *FUS* and *DDIT3.* Although WDLS and DDLS samples had a classical fusion gene (*CTSC-RAB38*), but we suspect it was formed due to the read-through transcription and might not be tumor specific. Overall, the DDLS samples had higher number of expressed direct fusion transcripts in comparison to the MLS and WDLS samples. Considering neoantigens derived from fusion genes have higher immunogenicity in comparison to SNVs or small insertion deletion derived neoantigens (107), the DDLS sub-type of liposarcoma might benefit from immune checkpoint blockade-based immunotherapy. Henceforth, the application of FUdGE for tumor specific CMTs in low SNVs mutation burden cancer sub-types can be particularly attractive. The prediction of CMTs in cancer transcriptome and associated neo-antigens can open immunotherapy related treatment opportunities for cancer patients.

Both IR and INR fusion transcript in cancer transcriptome is an actively evolving field. Previous studies have reported widespread expression of IR transcripts in cancer (41) and a source of neoantigens (108). The higher neoantigen load from IR fusion transcripts are also associated with poor survival in the multiple myeloma patients (109). Such studies signifies that the current repertoire of neoantigens can further be expanded with neoantigens derived from all possible tumor-specific mutations. However, the tumor-specific detection of CMTs from RNA-seq data remain a concern in the research community (110). The FUdGE pipeline can offer an advantage in such scenario. Moreover, the continuously evolving intergenic genomic space and the functional consequence of CMTs with transcribed long non-coding RNAs (111), pseudogenes (112) and upstream and downstream region of annotated genes (106) can provide key insights in transcriptional regulation of the genes in cancer cell. In order to treat cancer patients with classical immunotherapy or emerging immunotherapy approaches, FUdGE can further expand the repertoire of neoantigens that will entail higher and more specific immune response in the patients.

4.4 Methods

4.4.1 Chimeric fusion transcripts (CMT) predictions from SVs in FUdGE

Figure 4-2 describes the schematic pipeline of FUdGE for prediction of several chimeric mRNA fusion transcripts (CMTs). A list of SV events with the genomic coordinates, type and orientation of SV is given as input. An SV can be a deletion, duplication, inversion and translocation with 3to5, 5to3, 3to3 and 5to5 orientation as explained in (**Chapter 3**). Each SV's genomic coordinate is annotated using ENSEMBL 86 genome annotation file (.gtf) for GRCh38 genome. Each genomic breakpoint is annotated with two upstream and downstream exons of neighbouring genes on both the strands. Based on the location of breakpoints, type and orientation of SVs, the possible mRNA structure of CMTs is predicted as shown in Figure 4-1. Three possible categories of CMTs include direct, intron-retained (IR) and intergenic-retained (INR) fusion transcripts. The concept behind direct fusion transcripts involves merging of neighbouring exons of annotated genomic segment upstream and downstream with different possible combinations as outlined in Supplementary figure 4-1. For the case when the genomic coordinate of underlying SV is located within exon of two different annotated sections, the exact coordinate in exons is merged in the direct fusion transcript. Next, the plausible IR fusion transcripts are generated when at least one breakpoint is within intron of one annotated section and other is in exon of a different annotated section. Similarly, INR fusion transcripts cases are predicted when one breakpoint is within an unannotated intergenic region while other is within an exon of annotated section.

Each of the possible call is enlisted with the breakpoints or coordinates that would merge in the CMTs. Next, the sequence of 200bp around those merged breakpoints is retrieved from the GRCh38 reference genome and a synthetic genomic template is created. In the requantification step, the RNA-seq reads of respective tumor samples are aligned to this template using STAR (v2.6.1) (113). The supporting reads in terms of junction reads (reads mapping at the breakpoints with at least 10bp around the merged segments) and spanning reads (paired-end read pair with one read mapping to first segment and second read mapping to another merged segment) are calculated. The CMTs are considered expressed or confirmed when at least some number of junction or spanning reads support it.

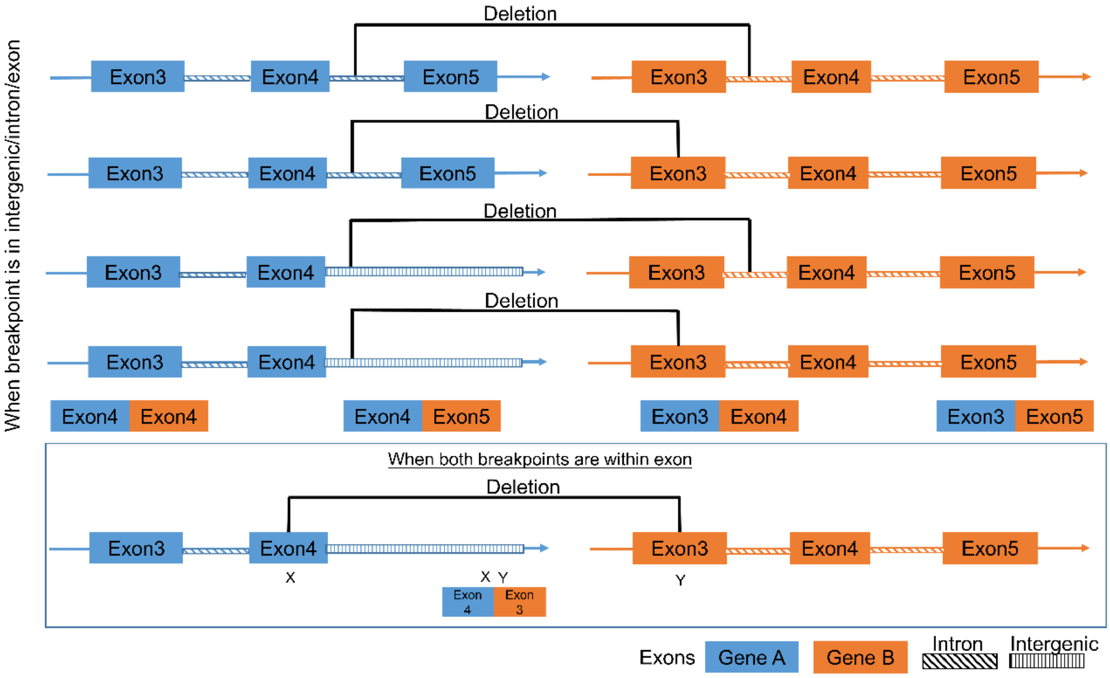
4.4.2 Application of FUdGE to MCF7 and primary breast tumor sample

The Illumina paired-end WGS reads for MCF7 and primary breast tumor were obtained from (72). A list of SVs was obtained with FuseSV pipeline (**Chapter 3**) and the probability score from FuseSV was used to benchmark performance of FUdGE with the qPCR validated direct fusion transcripts. The CMTs were confirmed by RNA-seq data with at least 1 junction or spanning reads in the requantification step.

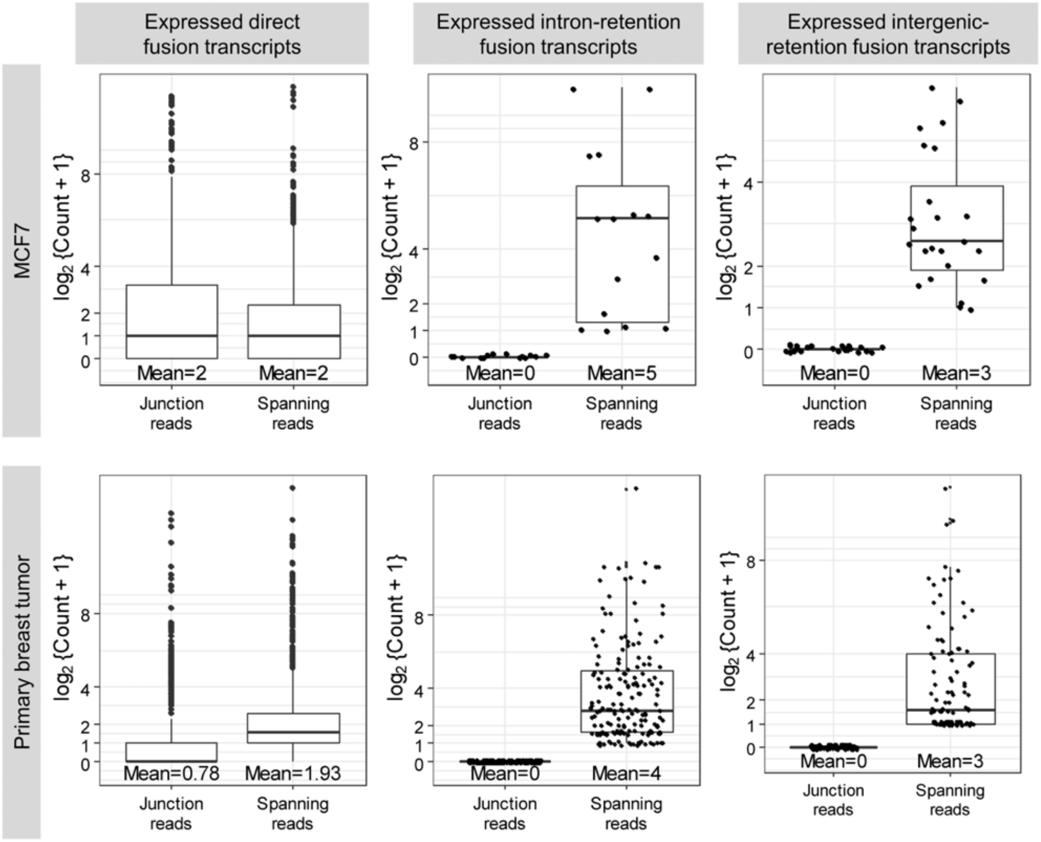
4.4.3 Application of FUdGE to liposarcoma samples

26 liposarcoma samples with paired tumor-normal Illumina WGS and RNA-seq of tumor sample was obtained from collaboration with Prof. Thomas Kindler at University Center of Tumor Diseases, Mainz. The WGS data of the liposarcoma samples was analysed as reported in Chapter 3. The paired RNA-seq data of the liposarcoma samples was analysed as described in the requantification step of FUdGE pipeline.

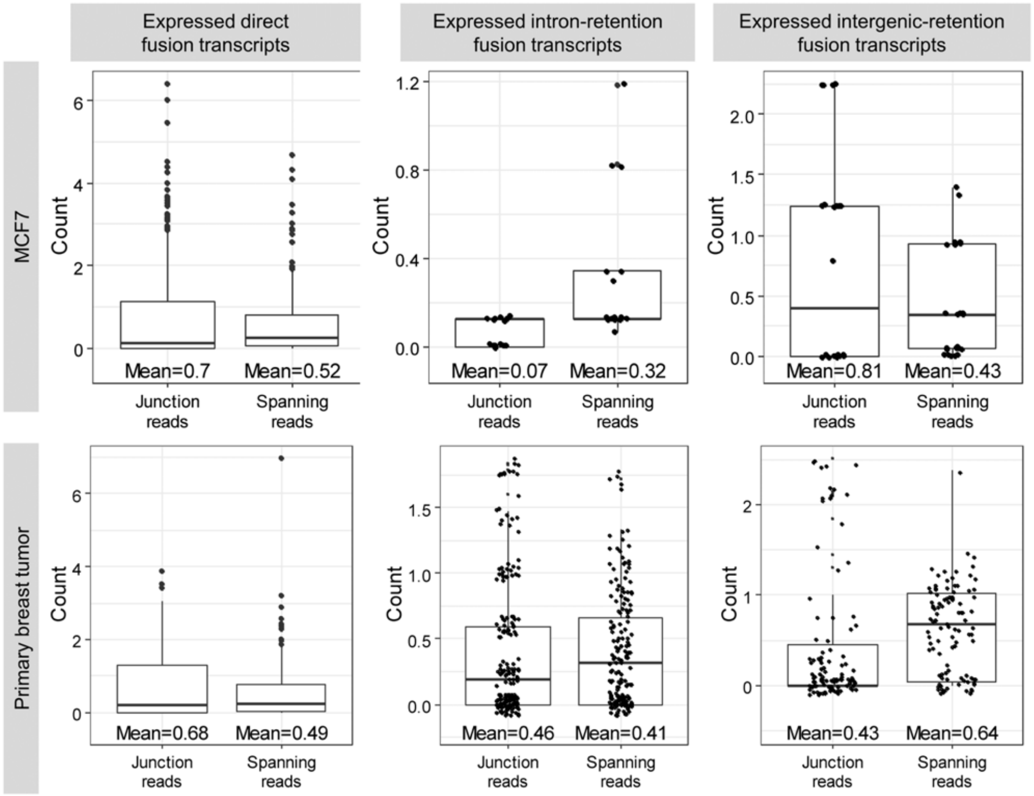
4.5 Supplementary figures



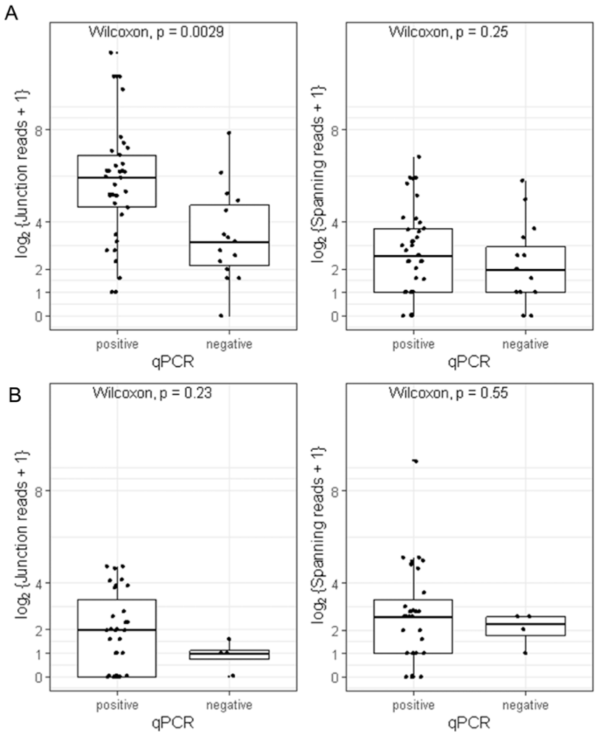
Supplementary figure 4-1: The schema for prediction of direct fusion transcripts by FUdGE. The scheme followed for prediction of different direct fusion transcripts with location of SV's breakpoints within an intergenic region or within a gene (in intron or exon) is shown.



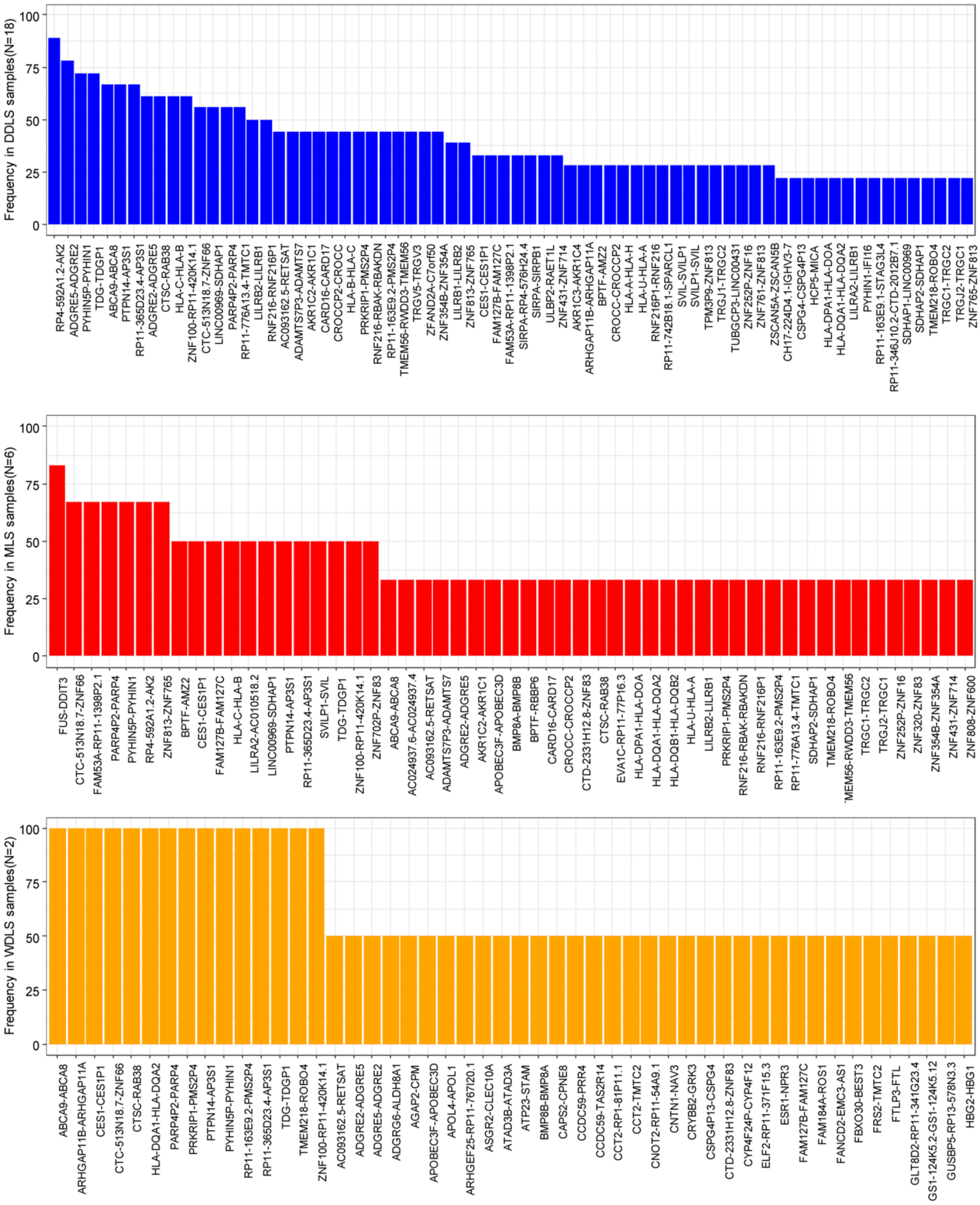
Supplementary figure 4-2: Various FUdGE predicted CMTs (direct, intron-retention, intergenic-retention) have support from RNA-seq data in MCF7 and the primary breast tumor sample. The plot depicts number of junction and spanning reads from the requantification of CMTs in RNA-seq data. The number of expressed direct fusion transcripts were 1238 and 2822 in MCF7 and the primary breast tumor sample respectively. The number of expressed intron-retained fusion transcripts were 15 and 175 in MCF7 and the primary breast tumor sample respectively. The number of expressed intergenic region retained fusion transcripts were 24 and 101 in MCF7 and the primary breast tumor sample respectively.



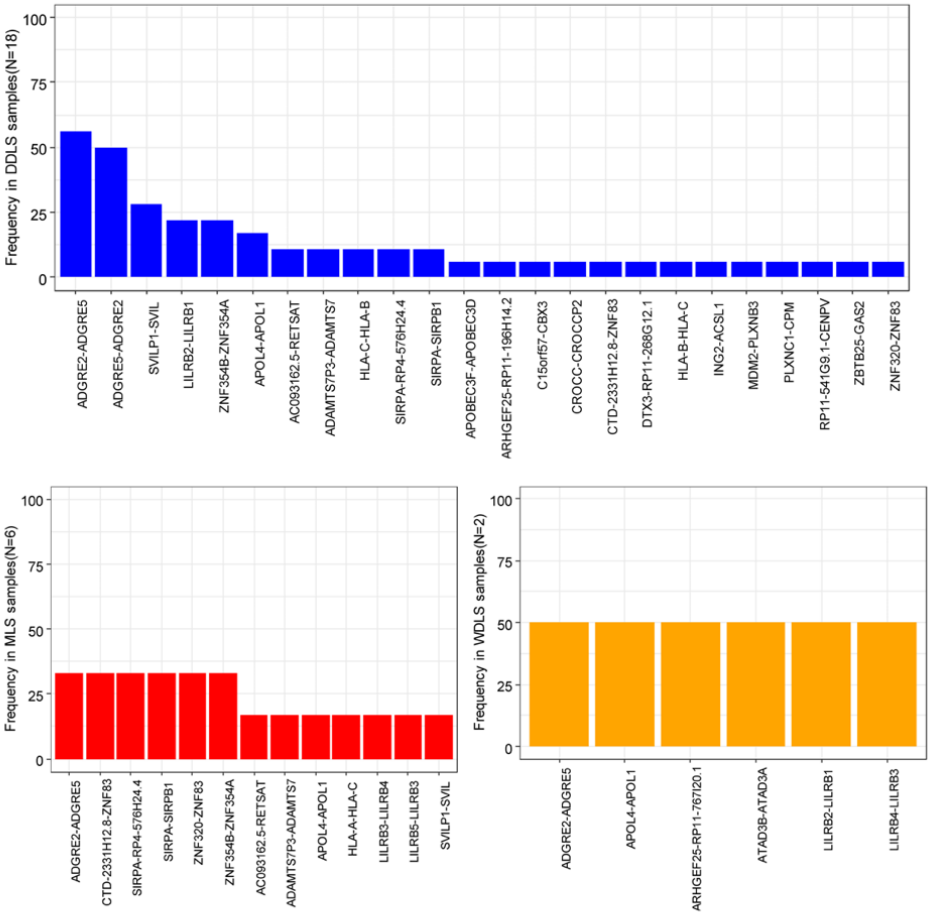
Supplementary figure 4-3: The underlying SVs for FUdGE predicted CMTs (direct, intron-retention, intergenic-retention) have support from WGS data in MCF7 and the primary breast tumor sample. The plot depicts number of junction and spanning reads from the requantification of SVs with WGS reads that generated CMTs that were expressed or confirmed with RNA-seq data. The number of expressed direct fusion transcripts were 1238 and 2822 in MCF7 and the primary breast tumor sample respectively. The number of expressed intron-retained fusion transcripts were 15 and 175 in MCF7 and the primary breast tumor sample respectively. The number of expressed intergenic region retained fusion transcripts were 24 and 101 in MCF7 and the primary breast tumor sample respectively.



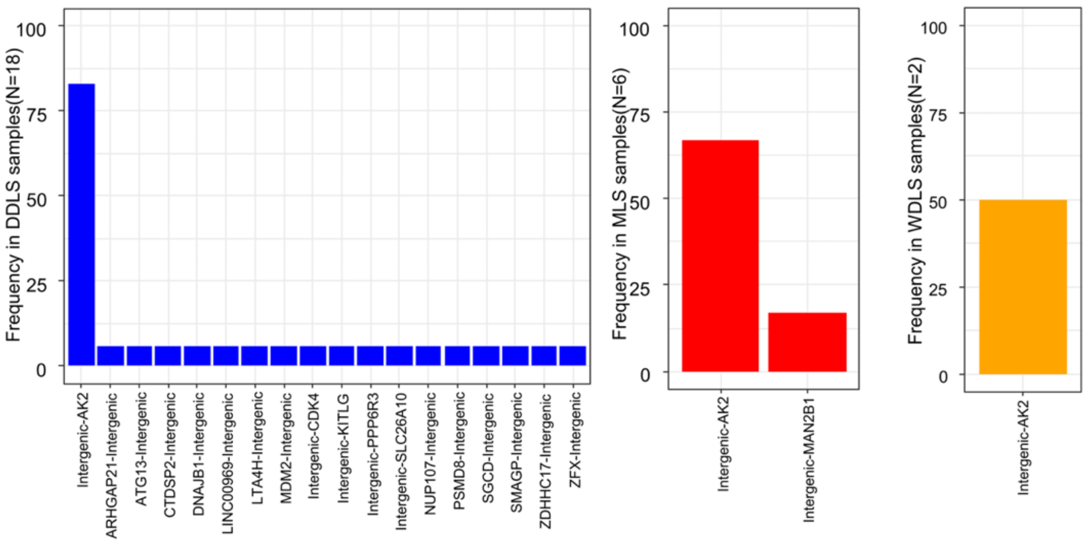
Supplementary figure 4-4: The junction and spanning reads support FUdGE predicted direct fusion transcripts that were validated by qRT-PCR/qPCR. The figure plots junction and spanning RNA-seq reads for the direct fusion transcripts that were positive or negative by qPCR validation. Figure A and B represents data from MCF7 and the primary breast tumor sample respectively.



Supplementary figure 4-5: Most of the MLS samples in liposarcoma cohort are characterized with *FUS-DDIT3* fusion gene. The plots depict percentage of different direct fusion transcripts in three sub-types of liposarcoma i.e., DDLS, MLS and WDLS. The figure plots frequency of direct fusion transcripts that were confirmed with RNA-seq data with at least 3 junction or spanning reads within DDLS, MLS and WDLS samples.



Supplementary figure 4-6: The distribution of intron-retained transcripts expressed in the liposarcoma samples was confounded with fusions between paralogous genes. The plots depict percentage of different annotated sections of intron-retained CMTs in three sub-types of liposarcoma i.e., DDLS, MLS and WDLS. The figure plots frequency of intron-retained fusion transcripts that were confirmed with RNA-seq data with at least 3 junction or spanning reads within DDLS, MLS and WDLS samples.



Supplementary figure 4-7: The distribution of intergenic-retained fusion transcripts expressed in the liposarcoma samples was characterized with fusion of an intergenic region with the *AK2* gene. The plots depict percentage of different annotated sections of intergenic-retained CMTs in three sub-types of liposarcoma i.e., DDLS, MLS and WDLS. The figure plots frequency of intergenic-retained fusion transcripts that were confirmed with RNA-seq data with at least 3 junction or spanning reads within DDLS, MLS and WDLS samples.

4.6 Supplementary tables

Supplementary table 4-1: Distribution of the CMTs in MCF7 and the primary breast tumor sample that were confirmed with RNA-seq data. The table mentions number of various types of CMTs that were predicted by FUdGE and confirmed with RNA-seq data along with the location of the breakpoints of the underlying SVs in intergenic regions or within a gene.

Sample		Direct Fusion Transcripts			Intron-retention Fusion Transcripts	Intergenic-retention Fusion Transcripts
		Gene-Gene	Gene-Intergenic	Intergenic-Intergenic		
MCF7	Confirmed	845 (77.2%)	169(9.12%)	224 (13.7%)	15	24
	Possible CMTs	7680 (1.5%)	41470 (9.3%)	364384 (89.2%)	770	1001
	% confirmed CMTs	97% (N=1238)			1.2% (N=15)	1.8% (N=24)
Primary breast tumor	Confirmed	1363 (62.7%)	456 (11.9%)	1003 (25.4%)	175	101
	Possible CMTs	18145 (2.2%)	104450 (12.8%)	698140 (85%)	2267	3666
	% confirmed CMTs	91.1% (N = 2822)			5.6% (N=175)	3.3% (N=101)

Supplementary table 4-2: The distribution of recurrent direct fusion transcripts in DDLS, MLS and WDLS type of liposarcoma samples. The table represents different fusion partners of direct fusion transcripts in sub-types of liposarcoma (DDLS: N=18; MLS: N=6; WDLS: N=6) that were recurrently detected in respective sub-types.

Fusion partner 1 (FP1)	FP1-annotation	Fusion partner 2 (FP2)	FP2-annotation	Frequency	Comment	Type
RP4-592A1.2	Pseudogene	AK2	protein-coding gene	88.89%	Pseudogene-gene	DDLS
ADGRE5	protein-coding gene	ADGRE2	protein-coding gene	77.78%	Paralog-Paralog	DDLS
PYHIN5P	Pseudogene	PYHIN1	protein-coding gene	72.20%	Pseudogene-tumor suppressor gene	DDLS
TDG	protein-coding gene	TDGP1	Pseudogene	72.20%	gene-Pseudogene	DDLS
ABCA9	protein-coding gene	ABCA8	protein-coding gene	66.67%	Paralog-Paralog	DDLS
PTPN14	protein-coding gene	AP3S1	Pseudogene	66.67%	Tumor suppressor gene-Pseudogene	DDLS
RP11-365D23.4	Pseudogene	AP3S1	Pseudogene	66.67%	Pseudogene-Pseudogene	DDLS
ADGRE2	protein-coding gene	ADGRE5	protein-coding gene	61.11%	Paralog-Paralog	DDLS
CTSC	protein-coding gene	RAB38	protein-coding gene	61.11%	gene-oncogene	DDLS
HLA-C	protein-coding gene	HLA-B	protein-coding gene	61.11%	Paralog-Paralog	DDLS
ZNF100	protein-coding gene	RP11-420K14.1	Pseudogene	61.11%	gene-Pseudogene	DDLS
CTC-513N18.7	protein-coding gene	ZNF66	protein-coding gene	55.55%	Paralog-Paralog	DDLS
LINC00969	Long non-coding RNA	SDHAP1	Pseudogene	55.55%	Long non-coding RNA-Pseudogene	DDLS
PARP4P2	Pseudogene	PARP4	protein-coding gene	55.55%	Pseudogene-gene	DDLS
RP11-776A13.4	Pseudogene	TMC1	protein-coding gene	55.55%	Pseudogene-gene	DDLS
LILRB2	protein-coding gene	LILRB1	protein-coding gene	50%	Paralog-Paralog	DDLS
RNF216	protein-coding gene	RNF216P1	Pseudogene	50%	gene-Pseudogene	DDLS
FUS	protein-coding gene	DDIT3	protein-coding gene	83.33%	gene-gene	MLS
CTC-513N18.7	protein-coding gene (ZNF626)	ZNF66	protein-coding gene	66.66%	Paralog-Paralog	MLS

FAM53A	protein-coding gene	RP11-1398P2.1	Long non-coding RNA	66.66%	gene-Long non-coding RNA	MLS
PARP4P2	Pseudogene	PARP4	protein-coding gene	66.66%	Pseudogene-gene	MLS
PYHIN5P	Pseudogene	PYHIN1	protein-coding gene	66.66%	Pseudogene-tumor suppressor gene	MLS
RP4-592A1.2	Pseudogene	AK2	protein-coding gene	66.66%	Pseudogene-gene	MLS
ZNF813	protein-coding gene	ZNF765	protein-coding gene	66.66%	Paralog-Paralog	MLS
BPTF	protein-coding gene	AMZ2	protein-coding gene	50.00%	gene-gene	MLS
CES1	protein-coding gene	CES1P1	Pseudogene	50.00%	gene-Pseudogene	MLS
FAM127B	protein-coding gene	FAM127C	protein-coding gene	50.00%	Paralog-Paralog	MLS
HLA-C	protein-coding gene	HLA-B	protein-coding gene	50.00%	Paralog-Paralog	MLS
LILRA2	protein-coding gene	AC010518.2	Pseudogene	50.00%	gene-Pseudogene	MLS
LINC00969	Long non-coding RNA	SDHAP1	Pseudogene	50.00%	Long non-coding RNA-Pseudogene	MLS
PTPN14	protein-coding gene	AP3S1	Pseudogene	50.00%	Tumor suppressor gene-Pseudogene	MLS
RP11-365D23.4	Pseudogene	AP3S1	Pseudogene	50.00%	Pseudogene-Pseudogene	MLS
SVILP1	Pseudogene	SVIL	protein-coding gene	50.00%	Pseudogene-gene	MLS
TDG	protein-coding gene	TDGP1	Pseudogene	50.00%	gene-Pseudogene	MLS
ZNF100	protein-coding gene	RP11-420K14.1	Pseudogene	50.00%	gene-Pseudogene	MLS
ZNF702P	Pseudogene	ZNF83	protein-coding gene	50.00%	Pseudogene-gene	MLS
ABCA9	protein-coding gene	ABCA8	protein-coding gene	100.00%	Paralog-Paralog	WDLS
ARHGAP11B	protein-coding gene	ARHGAP11A	protein-coding gene	100.00%	Paralog-Paralog	WDLS
CES1	protein-coding gene	CES1P1	Pseudogene	100.00%	gene-Pseudogene	WDLS
CTC-513N18.7	protein-coding gene (ZNF626)	ZNF66	protein-coding gene	100.00%	Paralog-Paralog	WDLS

CTSC	protein-coding gene	RAB38	protein-coding gene	100.00%	gene-oncogene	WDLS
HLA-DQA1	protein-coding gene	HLA-DQA2	protein-coding gene	100.00%	Paralog-Paralog	WDLS
PARP4P2	Pseudogene	PARP4	protein-coding gene	100.00%	Pseudogene-gene	WDLS
PRKRIP1	protein-coding gene	PMS2P4	Pseudogene	100.00%	gene-Pseudogene	WDLS
PTPN14	protein-coding gene	AP3S1	Pseudogene	100.00%	Tumor suppressor gene-Pseudogene	WDLS
PYHIN5P	Pseudogene	PYHIN1	protein-coding gene	100.00%	Pseudogene-tumor suppressor gene	WDLS
RP11-365D23.4	Pseudogene	AP3S1	Pseudogene	100.00%	Pseudogene-Pseudogene	WDLS
TDG	protein-coding gene	TDGP1	Pseudogene	100.00%	gene-Pseudogene	WDLS
TMEM218	protein-coding gene	ROBO4	protein-coding gene	100.00%	gene-gene	WDLS
ZNF100	protein-coding gene	RP11-420K14.1	Pseudogene	100.00%	gene-Pseudogene	WDLS

5 Future outlook

The technologies used for the identification of SVs have evolved over time. With the latest technologies it is now possible to define types of SVs with nucleotide base resolution of breakpoints. The most popular amongst them has been Illumina's short-read sequencing technology. However, this technology suffers from a high false discovery rate which can be mitigated with the long read sequencing technology by PacBio and Oxford Nanopore. It enables discovery of novel mutations especially in the high complexity regions of the human genome that are difficult to resolve by short-reads. Nevertheless, the long read technology offers both advantages and disadvantages in detection of SVs. Its lower accuracy rate and high cost are some disadvantages. Consequently, this technology is primarily used for research or validation of mutations, while short-read sequencing has been used in several large consortiums/studies like 1000 Genomes and PCAWG. However, due to the longer read length it does offer an advantage in resolving SVs, especially the ones originating from high complexity regions of the genome and novel type of SVs like insertions. It is undeniable that the usage of the long read sequencing technology expands the detectable mutational landscape of a genome. As a future outlook for the detection of SVs, the combination of methods like short-reads paired with low coverage long read sequencing would in my opinion allow the best utilisation of both technologies with lower cost in comparison to high coverage sequencing of a genome from individual technologies.

In complex diseases like cancer, SVs are often complicated in nature with back-to-back variations and highly rearranged genomes. The short-read sequencing technology and related SVs tools often struggle to resolve such variations that stitch several pieces of information together for interpretation. Some recent specialised algorithmic studies attempted to resolve them using short-reads, but long read sequencing technology has proven to be more efficient in resolving such SVs. Nevertheless, the interpretation of such complex SVs remains challenging. For example, interspersed duplication events can be interpreted as deletion and duplication or deletion and inversion (if the segments are located on the same chromosome) or deletion and translocation (if the segment is deleted and inserted on another chromosome). These types of interspersed duplications will have different read mapping signals that are difficult to resolve irrespective of the sequencing technology they are derived from. Advanced machine learning algorithms like deep learning might be able to offer some rebate in this aspect. However, the biggest hurdle in using deep learning algorithms is the lack of sufficient number of experimentally validated complex SVs. It would be possible to use deep learning algorithms with input mappings of short-reads/long-reads sequencing for resolution of complex SVs in the future.

It is crucial to understand the impact of genomic mutations on the transcriptome or proteome in the context of a disease. One of the functional consequences of SVs is the formation of chimeric transcripts. In this case the presented software tool, FUdGE, can be beneficial to scientists studying the impact of somatic SVs in terms of novel chimeric transcripts created and dominant in a subtype of disease. The types of disease dominated by a characteristic SV or a chimeric transcript can benefit from non-targetable therapies like immunotherapy. In the past, immunotherapy drugs have proven to be effective for treatment of cancer patients with high mutational burden in terms of SNVs (like melanoma). Thus, it is plausible that diseases dominated by SVs or chimeric transcripts would be curable with an immunotherapy drug targeting it. However, the disease cases with lower mutational burden like sarcoma and rare diseases can also benefit from novel targeted therapies for somatic SVs/chimeric transcripts. One such approach would include studying the translation of somatic chimeric transcripts into neo-antigens and their presentation to immune cells in the body. Currently, such personalised treatments include only the neo-antigens derived from SNVs. I believe that the drastic rearrangement of the genome with SVs and chimeric transcripts derived from them, could lead to neo-antigens with much stronger immune response in comparison to SNVs derived neo-antigens. There are limited studies displaying strong immune responses to fusion gene derived neo-antigens and henceforth, research in this direction can lead to better therapies for all the patients. A logical step forward would be to study the impact of SVs derived chimeric transcripts in cancer patients in terms of immune response.

Overall, the prediction of drug targets for many genetic diseases is multi-facet task that also requires scientists to gather the complete mutational profile of the genome and study its impact. In the past, the mutational landscape arising from SNVs was studied widely. The addition of SVs to this mutational landscape and its impact would allow novel targets for treatment of cancer. Apart from the generation of chimeric transcripts, SVs can also affect the 3D structure of the genome that can further impact DNA-DNA interactions and the expression of genes. This aspect can further widen our knowledge in understanding the effect of SVs in the genome. The more we discover and understand the impact of different mutations in a disease genome, the closer we would get to finding right targets for curing those diseases.

6 References

1. Pott P. The Chirurgical Works. Vol. 1. T. Lowndes, J. Johnson, G. Robinson, T. Cadell, T. Evans, W. Fox, J. Bew and …; 1779.
2. Bignold LP. Variation,“evolution”, immortality and genetic instabilities in tumour cells. Cancer Lett. 2007;253(2):155–69.
3. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526(7571):75–81.
4. Yi K, Ju YS. Patterns and mechanisms of structural variations in human cancer. Exp Mol Med. 2018;50(8):1–11.
5. Pan-cancer analysis of whole genomes. Nature. 2020;578(7793):82–93.
6. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. cell. 2011;144(1):27–40.
7. Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, et al. Punctuated evolution of prostate cancer genomes. Cell. 2013;153(3):666–77.
8. Zakov S, Kinsella M, Bafna V. An algorithmic approach for breakage-fusion-bridge detection in tumor genomes. Proceedings of the National Academy of Sciences. 2013;110(14):5546–51.
9. Hadi K, Yao X, Behr JM, Deshpande A, Xanthopoulakis C, Tian H, et al. Distinct Classes of Complex Structural Variation Uncovered across Thousands of Cancer Genome Graphs. Cell. 2020;183(1):197-210.e32.
10. Nowell PC. A minute chromosome in human chronic granulogytic leukemia. Science. 1960;132:1497.
11. Langer-Safer PR, Levine M, Ward DC. Immunological method for mapping genes on Drosophila polytene chromosomes. Proceedings of the National Academy of Sciences. 1982;79(14):4381–5.
12. Gisselsson D, Pettersson L, Höglund M, Heidenblad M, Gorunova L, Wiegant J, et al. Chromosomal breakage-fusion-bridge events cause genetic intratumor heterogeneity. Proceedings of the National Academy of Sciences. 2000;97(10):5357–62.
13. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, et al. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science (1979). 1992;258(5083):818–21.
14. Bignell GR, Huang J, Greshock J, Watt S, Butler A, West S, et al. High-resolution analysis of DNA copy number using oligonucleotide microarrays. Genome Res. 2004;14(2):287–95.
15. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics. 2012;13(1):36–46.
16. Zheng GXY, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat Biotechnol. 2016;34(3):303–11.
17. Ottaviani D, LeCain M, Sheer D. The role of microhomology in genomic structural variation. Trends Genet. 2014;30(3):85–94.
18. Hanscom T, McVey M. Regulation of Error-Prone DNA Double-Strand Break Repair and Its Impact on Genome Evolution. Cells. 2020;9(7):1657.
19. Wright WD, Shah SS, Heyer WD. Homologous recombination and the repair of DNA double-strand breaks. Journal of Biological Chemistry. 2018;293(27):10524–35.

20. White TB, Morales ME, Deininger PL. Alu elements and DNA double-strand break repair. Mobile Genetic Elements. 2015;5(6):81–5.

21. Costantino L, Sotiriou SK, Rantala JK, Magin S, Mladenov E, Helleday T, et al. Break-induced replication repair of damaged forks induces genomic duplications in human cells. Science (1979). 2014;343(6166):88–91.

22. Mehta A, Beach A, Haber JE. Homology requirements and competition between gene conversion and break-induced replication during double-strand break repair. Mol Cell. 2017;65(3):515–26.

23. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nature Reviews Genetics. 2009;10(8):551–64.

24. Wyatt DW, Feng W, Conlin MP, Yousefzadeh MJ, Roberts SA, Mieczkowski P, et al. Essential roles for polymerase θ-mediated end joining in the repair of chromosome breaks. Mol Cell. 2016;63(4):662–73.

25. Chang HHY, Pannunzio NR, Adachi N, Lieber MR. Non-homologous DNA end joining and alternative pathways to double-strand break repair. Nature reviews Molecular cell biology. 2017;18(8):495–506.

26. Yu AM, McVey M. Synthesis-dependent microhomology-mediated end joining accounts for multiple types of repair junctions. Nucleic Acids Res. 2010;38(17):5706–17.

27. Bhargava R, Onyango DO, Stark JM. Regulation of single-strand annealing and its role in genome maintenance. Trends in Genetics. 2016;32(9):566–75.

28. Zelensky AN, Schimmel J, Kool H, Kanaar R, Tijsterman M. Inactivation of Pol θ and C-NHEJ eliminates off-target integration of exogenous DNA. Nat Commun. 2017;8(1):1–7.

29. Zámborszky J, Szikriszt B, Gervai JZ, Pipek O, Póti Á, Krzystanek M, et al. Loss of BRCA1 or BRCA2 markedly increases the rate of base substitution mutagenesis and has distinct effects on genomic deletions. Oncogene. 2017;36(6):746–55.

30. Sasaki T, Rodig SJ, Chirieac LR, Jänne PA. The biology and treatment of EML4-ALK non-small cell lung cancer. Eur J Cancer. 2010;46(10):1773–80.

31. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013;45(10):1134–40.

32. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010;463(7283):899–905.

33. Jost D, Vaillant C, Meister P. Coupling 1D modifications and 3D nuclear organization: data, models and function. Curr Opin Cell Biol. 2017;44:20–7.

34. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013;45(10):1134–40.

35. Lupiáñez DG, Spielmann M, Mundlos S. Breaking TADs: how alterations of chromatin domains result in disease. Trends in Genetics. 2016;32(4):225–37.

36. Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, Schöpflin R, et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature. 2016;538(7624):265–9.

37. Northcott PA, Lee C, Zichner T, Stütz AM, Erkek S, Kawauchi D, et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature. 2014;511(7510):428–34.

38. Gröschel S, Sanders MA, Hoogenboezem R, de Wit E, Bouwman BAM, Erpelinck C, et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. Cell. 2014;157(2):369–81.

39. Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, et al. Widespread intron retention in mammals functionally tunes transcriptomes. Genome Res. 2014;24(11):1774–86.

40. Boutz PL, Bhutkar A, Sharp PA. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. Genes Dev. 2015;29(1):63–80.

41. Dvinge H, Bradley RK. Widespread intron retention diversifies most cancer transcriptomes. Genome Med. 2015;7(1):1–13.

42. Tan DJ, Mitra M, Chiu AM, Coller HA. Intron retention is a robust marker of intertumoral heterogeneity in pancreatic ductal adenocarcinoma. NPJ Genom Med. 2020;5(1):1–17.

43. Ishida Y, Agata Y, Shibahara K, Honjo T. Induced expression of PD-1, a novel member of the immunoglobulin gene superfamily, upon programmed cell death. EMBO J. 1992;11(11):3887–95.

44. Leach DR, Krummel MF, Allison JP. Enhancement of antitumor immunity by CTLA-4 blockade. Science (1979). 1996;271(5256):1734–6.

45. Türeci Ö, Vormehr M, Diken M, Kreiter S, Huber C, Sahin U. Targeting the heterogeneity of cancer with individualized neoepitope vaccines. Clinical Cancer Research. 2016;22(8):1885–96.

46. Mansfield AS, Peikert T, Vasmatzis G. Chromosomal rearrangements and their neoantigenic potential in mesothelioma. Translational Lung Cancer Research. 2020;9(Suppl 1):S92.

47. Yang W, Lee KW, Srivastava RM, Kuo F, Krishna C, Chowell D, et al. Immunogenic neoantigens derived from gene fusions stimulate T cell responses. Nat Med. 2019;25(5):767–75.

48. Weber D, Ibn-Salem J, Sorn P, Suchan M, Holtsträter C, Lahrmann U, et al. Accurate detection of tumor-specific gene fusions reveals strongly immunogenic personal neo-antigens. Nature Biotechnology. 2022;1–9.

49. Sbaraglia M, Bellan E, Dei Tos AP. The 2020 WHO classification of soft tissue tumours: news and perspectives. Pathologica. 2021;113(2):70.

50. Mandahl N, Magnusson L, Nilsson J, Viklund B, Arbajian E, von Steyern FV, et al. Scattered genomic amplification in dedifferentiated liposarcoma. Mol Cytogenet. 2017;10(1):1–10.

51. Jour G, Gullet A, Liu M, Hoch BL. Prognostic relevance of Fédération Nationale des Centres de Lutte Contre le Cancer grade and MDM2 amplification levels in dedifferentiated liposarcoma: a study of 50 cases. Modern Pathology. 2015;28(1):37–47.

52. Pedeutour F, Forus A, Coindre J, Berner J, Nicolo G, Michiels J, et al. Structure of the supernumerary ring and giant rod chromosomes in adipose tissue tumors. Genes, Chromosomes and Cancer. 1999;24(1):30–41.

53. Ricciotti RW, Baraff AJ, Jour G, Kyriss M, Wu Y, Liu Y, et al. High amplification levels of MDM2 and CDK4 correlate with poor outcome in patients with dedifferentiated liposarcoma: a cytogenomic microarray analysis of 47 cases. Cancer Genet. 2017;218:69–80.

54. Binh MBN, Sastre-Garau X, Guillou L, de Pinieux G, Terrier P, Lagacé R, et al. MDM2 and CDK4 immunostainings are useful adjuncts in diagnosing well-differentiated and dedifferentiated liposarcoma subtypes: a comparative analysis of 559 soft tissue neoplasms with genetic data. Am J Surg Pathol. 2005;29(10):1340–7.

55. Bill KLJ, Seligson ND, Hays JL, Awasthi A, Demoret B, Stets CW, et al. Degree of MDM2 amplification affects clinical outcomes in dedifferentiated liposarcoma. The Oncologist. 2019;24(7):989–96.

56. Lee SE, Kim YJ, Kwon MJ, Choi DI, Lee J, Cho J, et al. High level of CDK4 amplification is a poor prognostic factor in well-differentiated and dedifferentiated liposarcoma. 2014;

57. Amin-Mansour A, George S, Sioletic S, Carter SL, Rosenberg M, Taylor-Weiner A, et al. Genomic evolutionary patterns of leiomyosarcoma and liposarcoma. Clinical Cancer Research. 2019;25(16):5135–42.

58. Abeshouse A, Adebamowo C, Adebamowo SN, Akbani R, Akeredolu T, Ally A, et al. Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. Cell. 2017;171(4):950–65.

59. Hirata M, Asano N, Katayama K, Yoshida A, Tsuda Y, Sekimizu M, et al. Integrated exome and RNA sequencing of dedifferentiated liposarcoma. Nat Commun. 2019;10(1):1–12.

60. Jo VY, Fletcher CDM. WHO classification of soft tissue tumours: an update based on the 2013 (4th) edition. Pathology. 2014;46(2):95–104.

61. Panagopoulos I, Mandahl N, Mitelman F, Aman P. Two distinct FUS breakpoint clusters in myxoid liposarcoma and acute myeloid leukemia with the translocations t (12; 16) and t (16; 21). Oncogene. 1995;11(6):1133–7.

62. Antonescu CR, Tschernyavsky SJ, Decuseara R, Leung DH, Woodruff JM, Brennan MF, et al. Prognostic impact of P53 status, TLS-CHOP fusion transcript structure, and histological grade in myxoid liposarcoma: a molecular and clinicopathologic study of 82 cases. Clinical Cancer Research. 2001;7(12):3977–87.

63. Perez-Losada J, Sanchez-Martin M, Rodriguez-Garcia MA, Perez-Mancera PA, Pintado B, Flores T, et al. Liposarcoma initiated by FUS/TLS-CHOP: the FUS/TLS domain plays a critical role in the pathogenesis of liposarcoma. Oncogene. 2000;19(52):6015–22.

64. Lee ATJ, Thway K, Huang PH, Jones RL. Clinical and molecular spectrum of liposarcoma. Journal of Clinical Oncology. 2018;36(2):151.

65. Creytens D, Folpe AL, Koelsche C, Mentzel T, Ferdinande L, van Gorp JM, et al. Myxoid pleomorphic liposarcoma—a clinicopathologic, immunohistochemical, molecular genetic and epigenetic study of 12 cases, suggesting a possible relationship with conventional pleomorphic liposarcoma. Modern Pathology. 2021;34(11):2043–9.

66. Jones RL, Fisher C, Al-Muderis O, Judson IR. Differential sensitivity of liposarcoma subtypes to chemotherapy. Eur J Cancer. 2005;41(18):2853–60.

67. Bui NQ, Przybyl J, Trabucco SE, Frampton G, Hastie T, van de Rijn M, et al. A clinico-genomic analysis of soft tissue sarcoma patients reveals CDKN2A deletion as a biomarker for poor prognosis. Clin Sarcoma Res. 2019;9(1):1–11.

68. Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, et al. Patterns of somatic structural variation in human cancer genomes. Nature. 2020;578(7793):112–21.

69. Metzker ML. Sequencing technologies — the next generation. Nature Reviews Genetics. 2010;11(1):31–46.

70. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, et al. The potential and challenges of nanopore sequencing. Nat Biotechnol. 2008;26(10):1146–53.

71. Selvaraj S, R Dixon J, Bansal V, Ren B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nat Biotechnol. 2013;31(12):1111–8.

72. Sethi R, Becker J, Graaf J de, Löwer M, Suchan M, Sahin U, et al. Integrative analysis of structural variations using short-reads and linked-reads yields highly specific and sensitive predictions. PLOS Computational Biology. 2020;16(11):e1008397.

73. van Belzen IAEM, Schönhuth A, Kemmeren P, Hehir-Kwa JY. Structural variant detection in cancer genomes. NPJ Precis Oncol. 2021;5(1):15.

74. Cameron DL, di Stefano L, Papenfuss AT. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. Nature Communications. 2019;10(1):3240.

75. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biology. 2019;20(1):117.

76. Becker T, Lee WP, Leone J, Zhu Q, Zhang C, Liu S, et al. FusorSV. Genome Biology. 2018;19(1):38.

77. Lopez G, Egolf LE, Giorgi FM, Diskin SJ, Margolin AA. svpluscnv. Bioinformatics. 2021;37(13):1912–4.

78. Shao H, Ganesamoorthy D, Duarte T, Cao MD, Hoggart CJ, Coin LJM. npInv. BMC Bioinformatics. 2018;19(1):261.

79. Spies N, Weng Z, Bishara A, McDaniel J, Catoe D, Zook JM, et al. Genome-wide reconstruction of complex structural variants using read clouds. Nat Methods. 2017;14(9):915–20.

80. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO. DELLY. Bioinformatics. 2012;28(18):i333–9.

81. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY. Genome Biology. 2014;15(6):R84.

82. Wala JA, Bandopadhayay P, Greenwald NF, O'Rourke R, Sharpe T, Stewart C, et al. SvABA. Genome Research. 2018;28(4):581–91

83. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta. Bioinformatics. 2016;32(8):1220–2.

84. Xi R, Lee S, Xia Y, Kim TM, Park PJ. Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. Nucleic Acids Research. 2016;44(13):6274–86.

85. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215(3):403–10.

86. Zhao X, Emery SB, Myers B, Kidd JM, Mills RE. Resolving complex structural genomic rearrangements using a randomized approach. Genome Biology. 2016;17(1):126.

87. Nattestad M, Goodwin S, Ng K, Baslan T, Sedlazeck FJ, Rescheneder P, et al. Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. Genome Research. 2018;28(8):1126–35.

88. Cortés-Ciriano I, Lee JJK, Xi R, Jain D, Jung YL, Yang L, et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. Nat Genet. 2020;52(3):331–41.

89. Davoli T, Uno H, Wooten EC, Elledge SJ. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. Science. 2017;355(6322).

90. Zhang CZ, Spektor A, Cornils H, Francis JM, Jackson EK, Liu S, et al. Chromothripsis from DNA damage in micronuclei. Nature. 2015;522(7555):179–84.

91. Wickham H. ggplot2. Vol. 3. Wiley Interdisciplinary Reviews: Computational Statistics; 2011.

92. Sahin U, Derhovanessian E, Miller M, Kloke BP, Simon P, LÃ¶wer M, et al. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. Nature. 2017;547(7662):222–6.

93. Dumbrava EI, Meric-Bernstam F. Personalized cancer therapy—leveraging a knowledge base for clinical decision-making. Molecular Case Studies. 2018;4(2):a001578.

94. Soverini S, Mancini M, Bavaro L, Cavo M, Martinelli G. Chronic myeloid leukemia: the paradigm of targeting oncogenic tyrosine kinase signaling and counteracting resistance for successful cancer therapy. Mol Cancer. 2018;17(1):1–15.

95. Yun JW, Yang L, Park HY, Lee CW, Cha H, Shin HT, et al. Dysregulation of cancer genes by recurrent intergenic fusions. Genome Biol. 2020;21(1):1–20.

96. Zhang J, White NM, Schmidt HK, Fulton RS, Tomlinson C, Warren WC, et al. INTEGRATE: gene fusion discovery using whole genome and transcriptome data. Genome Res. 2016;26(1):108–18.

97. McPherson A, Wu C, Wyatt AW, Shah S, Collins C, Sahinalp SC. nFuse: discovery of complex genomic rearrangements in cancer using high-throughput sequencing. Genome Res. 2012;22(11):2250–61.

98. Rabbitts TH, Forster A, Larson R, Nathan P. Fusion of the dominant negative transcription regulator CHOP with a novel gene FUS by translocation t (12; 16) in malignant liposarcoma. Nat Genet. 1993;4(2):175–80.

99. Åman P, Ron D, Mandahl N, Fioretos T, Heim S, Arheden K, et al. Rearrangement of the transcription factor gene CHOP in myxoid liposarcomas with t (12; 16)(q13; p11). Genes, chromosomes and cancer. 1992;5(4):278–85.

100. Dormann D, Haass C. Fused in sarcoma (FUS): an oncogene goes awry in neurodegeneration. Molecular and Cellular Neuroscience. 2013;56:475–86.

101. Yang Y, Liu L, Naik I, Braunstein Z, Zhong J, Ren B. Transcription factor C/EBP homologous protein in health and diseases. Front Immunol. 2017;8:1612.

102. Grosso AR, Leite AP, Carvalho S, Matos MR, Martins FB, Vitor AC, et al. Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma. Elife. 2015;4:e09214.

103. Yan Z, Fan G, Li H, Jiao Y, Fu W, Weng J, et al. The CTSC-RAB38 Fusion Transcript Is Associated With the Risk of Hemorrhage in Brain Arteriovenous Malformations. Journal of Neuropathology & Experimental Neurology. 2021;80(1):71–8.

104. Liu H, Pu Y, Amina Q, Wang Q, Zhang M, Song J, et al. Prognostic and therapeutic potential of Adenylate kinase 2 in lung adenocarcinoma. Sci Rep. 2019;9(1):1–10.

105. Kumar S, Razzaq SK, Vo AD, Gautam M, Li H. Identifying fusion transcripts using next generation sequencing. Wiley Interdisciplinary Reviews: RNA. 2016;7(6):811–23.

106. Agostini F, Zagalak J, Attig J, Ule J, Luscombe NM. Intergenic RNA mainly derives from nascent transcripts of known genes. Genome Biol. 2021;22(1):1–19.

107. Wei Z, Zhou C, Zhang Z, Guan M, Zhang C, Liu Z, et al. The landscape of tumor fusion neoantigens: a pan-cancer analysis. Iscience. 2019;21:249–60.

108. Smart AC, Margolis CA, Pimentel H, He MX, Miao D, Adeegbe D, et al. Intron retention is a source of neoepitopes in cancer. Nat Biotechnol. 2018;36(11):1056–8.

109. Dong C, Cesarano A, Bombaci G, Reiter JL, Yu CY, Wang Y, et al. Intron retention-induced neoantigen load correlates with unfavorable prognosis in multiple myeloma. Oncogene. 2021;40(42):6130–8.

110. Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. Genome Biol. 2019;20(1):1–16.

111. Anastasiadou E, Jacob LS, Slack FJ. Non-coding RNA networks in cancer. Nature Reviews Cancer. 2018;18(1):5–18.

112. Szalmas A, Tomaić V, Basukala O, Massimi P, Mittal S, Konya J, et al. The PTPN14 tumor suppressor is a degradation target of human papillomavirus E7. J Virol. 2017;91(7):e00057-17.

113. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.

Riccha Sethi

Bioinformatician

Address: 35 Somner Close
Canterbury
UK
Date of birth: 24 April 1986
☎ *Phone: +44 (7392013648)*
✉ *Mail: ricchasethi@gmail.com*
Nationality: Indian

Research Interests

Computational and machine learning approaches for solving interesting problems in bioinformatics.

Education

Since 2016 **PhD in Bioinformatics**.

TRON gGmbH (Translational Oncology), Johannes Gutenberg University, Mainz, Germany

Thesis: Identification of structural variations and fusion genes from whole genome sequencing data of cancer patients

Advisors: Ugur Sahin, Martin Löwer, David Weber

2014-2016 **Master in Bioinformatics**.

Center for Bioinformatics (ZBI), Saarland University, Saarbrücken, Germany

GPA: 1.8/5 (inverted scale, 1.0 being the highest grade)

2009-2011 **Master of Engineering (Biotechnology)**.

Birla Institute of Technology and Science (BITS), Pilani, India

CGPA: 8.73/10

2004-2008 **Bachelor of Technology (Biotechnology)**.

Amity Institute of Biotechnology, Amity University, Noida, India

CGPA: 8.58/10

Work Experience

02/2022-Present **Bioinformatics Data Scientist**, *BenevolentAI*, United Kingdom.

Role: Process bulk transcriptomics, single cell RNA sequencing and proteomics data for drug discovery using AI based models

11/2016-09/2021 **Doctoral Research Assistant**, *TRON, Johannes Gutenberg University*, Germany.

Topics: Structural variations (SV) from whole genome sequencing data (WGS), benchmarking of Illumina short-reads with 10X Genomics linked-reads sequencing, machine learning approach for reliable prediction of SV, direct prediction of fusion genes, intron-retention and non-coding fusion transcripts from WGS

Advisors: Ugur Sahin, Martin Löwer, David Weber

10/2020-Present **Mother**, *Maternity Leave (09/2020-07/2021)*.

Role played: Acquired skills like maternal instincts, efficient organization, effective delivery of professional goals, time management, teaching and many more

04/2016- **Research Assistant**, *ZBI, Saarland University*, Germany.

07/2016 Topic: Linking hematopoietic differentiation to co-expressed sets of pluripotency-associated and imprinted genes and to regulatory microRNA-transcription factor motifs

Advisor: Volkhard Helms

08/2014- **Student Researcher**, *ZBI, Saarland University*, Germany.

03/2016 Topic: Synthetic data generation for evaluation of state-of-the-art haplotype phasing tools

Advisor: Tobias Marschall

07/2011- **Senior Executive, Quality Assurance**, *Biocon Limited*, India.

11/2013 Role played: Investigation of deviations and process changes in manufacturing of pharmaceutical drugs, incorporation of corrective and preventive action using statistical tools, face health authority audits like EU-GMP, handle change control request for improvement in drug manufacturing

01/2011- **Research Intern**, *Abexome Biosciences*, India.

06/2011 Topic: Production of monoclonal antibody for biological use

Advisor: Brijesh N Bhatt

08/2009- **Student Researcher**, *BITS, Pilani*, Rajasthan, India.

12/2010 Topic: Effects of morphine analogs on immune system cell line

Advisor: Uma Dubey

07/2008- **Trainee Scientist**, *NAM S&T Centre*, Delhi, India.

11/2008 Role played: Planning, implementation, evaluation and assessment of scientific programs of the centre

Publications

1. **Integrative analysis of structural variations using short-reads and linked-reads yields highly specific and sensitive predictions**

   with M. Löwer, U. Sahin and D. Weber

   PLOS Computational Biology, 2020

2. **STIM and ORAI genes, interactions with transcription factors, differential gene expression and co-expression analysis on breast invasive carcinoma dataset**

   with R. Mohamed, M. Hamed and V. Helms

   Front. Pharmacol. Conference Abstract: International Conference on Drug Discovery and Translational Medicine 2018 (ICDDTM '18) "Seizing Opportunities and Addressing Challenges of Precision Medicine", 2018

3. **Linking Hematopoietic Differentiation to Co-Expressed Sets of Pluripotency-Associated and Imprinted Genes and to Regulatory microRNA-Transcription Factor Motifs**

   with M. Hamed, V. Helms

   PLoS One, 2017

4. **Designer promoter: An artwork of cis-engineering**

   with R. Mehrotra, G. Gupta, N. Kumar and S. Mehrotra

   Plant Molecular Biology, 2011

Selected talks

1. "Direct detection of fusion gene neoantigens from whole genome sequencing" at Cancer Immunotherapy, CIMT (2019), Germany
2. "Direct identification of fusion genes from whole genome sequencing data" at Genome Informatics (2018), Wellcome Genome Campus, UK
3. "FuseSV: a pipeline to integrate structural variations from different callers" at Cancer Genomics conference (2017), EMBL, Germany
4. Presented and taught several topics (like sequencing technologies, neo-antigens generation and analysis, biological mechanisms for structural variations) to colleagues

Skills

1. Experience in application and usage of machine learning algorithms and related libraries (NumPy, Pandas, SciPy, scikit-learn, CARET etc.)
2. Experience in Unix/Linux systems
3. Experience in AWS, Kubeflow, DNAnexus and Amazon Redshift
4. Experience in standard NGS/DNA sequencing/RNA sequencing bioinformatics toolsets
5. Experience in data wrangling and robustness of code using unit tests
6. Knowledge of Git version control system and pipeline development in Nextflow
7. Knowledge in cancer genomics, immunology, machine learning, NGS and molecular biology
8. Experience in laboratory experimentation like PCR, SDS-PAGE, Western blot, ELISA etc.
9. Ability to colloborate in a team and work independently as well
10. Strong communication and organization skills

Programming languages

Python, R and SQL