UNIVERSITAT DE BARCELONA

Identification and characterization of new complex patterns of structural DNA and RNA alterations in cancer

Luisa Fernanda Delgado Serrano

IDENTIFICATION AND CHARACTERIZATION OF NEW COMPLEX PATTERNS OF STRUCTURAL DNA AND RNA ALTERATIONS IN CANCER

2021

BY LUISA FERNANDA DELGADO SERRANO

"What could I say to you that would be of value, except that perhaps you seek too much, that as a result of your seeking you cannot find"

- Hermann Hesse.

Programa de doctorat Biomedicina (HDK05)

Facultat de Biologia, Universitat de Barcelona

Identification and characterization of new complex patterns of structural DNA and RNA alterations in cancer

Memòria presentada per Luisa Fernanda Delgado Serrano

per optar al grau de doctora per la Universitat de Barcelona

Tesi realitzada al

Barcelona Supercomputing Center (BSC)

Doctorand

Luisa Fernanda Delgado Serrano

Director

David Torrents Arenales

Tutor

Josep Lluís Gelpí Buchaca

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Universitat de Barcelona

Agradecimientos

Han sido tantas las personas que de una u otra manera me han enriquecido en esta etapa como estudiante de doctorado, que ninguna de estas palabras le harán honor ¡a tanto!

En primera instancia, quiero agradecer a David el haberme abierto las puertas no sólo a ser parte de su grupo, sino a emprender esta gran etapa de mi vida en un país 8,510 kilómetros de mi tierra natal. Su forma particular de ver a los estudiantes por sus cualidades hace que cada uno de nosotros experimente su potencial como investigador. También quiero agradecer al equipo coordinado por David Carreras en el BSC, con los que compartimos dos mundos distintos pero complementarios. En especial a Gonzalo, por aguantar trabajar conmigo desde tan distintas perspectivas profesionales y lograr concebir juntos una publicación.

Me gustaría agradecer también a todos los miembros de mi grupo con los que compartimos juntos estos años la experiencia doctoral, comenzando por Jordi, el viejo George, mi predecesor en tantas cosas, empezando por el piso desde el cual se escribieron todas estas líneas, gracias por las tantas charlas a cerca de lo que compartimos y de lo que nos hace tan distintos, un aprendizaje a la perseverancia y la verdadera amistad. Gracias por haberme escuchado a las tantas de la madrugada, por haberme levantado infinidad de veces el ánimo, por habernos conectado en nuestras penas y alegrías, y por dedicarme el tiempo, energía y trabajo que hiciesen falta para que yo pudiera avanzar en esta tesis doctoral. A Lorena, por compartir conmigo tantos cafecitos, en donde teníamos largas tertulias científicas, otras más personales, todas ellas me dejaron siempre algo de su mirada optimista y de su conocimiento en materia matemática. A Ana, no solo por ser la primera en tener paciencia para enseñarme sobre la genómica de cáncer, el programar en Python, sino también

por mostrarme su ciudad, sus costumbres y por haber compartido mis primeros años acá con risas y alegrías.

A Ignasi, por ser El editor excepcional de este trabajo, que siempre con la frase "Ningún problema Luisa", me enseñó que hay gente que es brillante en lo que hace porque le gusta lo que hace y le sale innato de ¡maravilla! A Cecilia, por adoptarme en los momentos necesarios para los consejos y guianzas y por sus revisiones y aportes en este documento. A Álvaro, por incorporarse al proyecto y aportar sus ideas y trabajo. A Romina, por su devota disposición para ayudar siempre, con sus respuestas tan completas y elegantes sobre datasets y pipelines, aparte de sus pastelitos deliciosos que endulzaban nuestras mañanas. A Iván, por sus enseñanzas en R. A Dani, Juan, Alex, Flo, por compartir charlas, comidas e ideas de todo tipo. A Mattia del grupo de Salva, por participar en este proyecto y ser tan generoso con su conocimiento y en su trabajo. Al grupo de Marta Melé, por brindarnos orientación en los análisis de RNA-seq.

Fuera del ámbito profesional, quiero agradecer a mi Juancho, un ser que con su gran luz, alegría y creatividad logró concebir la portada de este trabajo. A parte, le agradezco inmensamente el heredarme esa gran familia en Barna con la cual hoy en día me encuentro compartiendo: mis amigos Cata y Sergio, por brindarnos un cálido momento siempre de charlas de todo tipo; Julio, mi amigo y apoyo con el cual sentimos siempre a Latinoamérica una sola.

Agradecer también a mi amigo del viejo continente, Enrico, con el que aprendí el significado de "A saco" en estos años por Europa. A mis amigos Colombianos en Europa, en especial a Camilo, por ser un recordatorio de lo que soy y por dedicar su tiempo (hasta en pijamas) en las tantas discusiones que tuvimos sobre biología molecular. A Ivi "el helmanito", por sus conocimientos en el campo experimental y por sus atenciones en nuestros viajes.

A mi amiga Luz K, que aunque regadas por el mundo, sigue habiendo ese lazo irrompible que me mantiene a flote. A “mi ami", Juan Carlos, que, con sus consejos, sus conocimientos en inglés y su amistad, me ha llevado en mi camino. Todas las líneas de este trabajo tienen su sello impregnado y su toque sofisticado y científico.

A mi familia, ese núcleo familiar tan sólido con el que infinitamente me siento agradecida, por su apoyo a todo dar en mis decisiones y su amor inconmensurable con el que pude salir adelante en los momentos más difíciles. Porque aún en la distancia, han sido el eje central de mi vida. Y a mi abue, que aunque no nos dio el tiempo para un último beso y abrazo, si me dio su bendición, que ahora me acompaña para siempre.

Por último, dejo mis palabras mayores de agradecimiento a mi Javito, la pareja que no vine buscando pero que me encontré en este espacio y este tiempo. Agradezco infinitamente a él y a su familia, en especial a Pedrito, por hacerme sentir siempre como en casa, protegida y querida. Gracias por ser mi familia acá en España; Por darme las fuerzas y la entereza para seguir siempre adelante, por las charlas científicas y seudocientíficas tan enriquecedoras para mi mente y mi alma; Por haberme dado herramientas tan valiosas para mi claridad mental, mi luz interior y por haberme soportado y cuidado tanto en todos los altibajos que supuso para mi esta tesis. Sin él, nada de esto estaría hoy acá plasmado.

“Lo real de lo irreal”

Dedicada a Javi.

Thesis Trajectory

I would like to introduce the trajectory of my thesis exposing here the bases of our decisions and strategies we succeeded. Following the trajectory of our group in the study of structural variants (SVs) in tumor genomes, in this thesis, we wanted to go one step beyond and uncover complex structural events occurring in cancer at genomic and transcriptomic levels that could inform us about their mechanisms and their possible functional impact. Thus, motivated by the previous identification of a particular potential pattern of structural variation within the study of the Chronic Lymphocytic Leukemia genomes, and given our participation in the ICGC-TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium, my research was initially focused on the functional characterization of the potential pattern using this cohort. To this end, I had to develop a strategy to first isolate the chromosomal rearrangements involved in the pattern. During the first two years of my thesis, I identified the features determining this particular pattern of SVs. Because complex chromosomal rearrangements are often composed of multiple DNA breaks, their classification and interpretation functionally represent a big challenge, demanding more comprehensive approaches like AI. Therefore, we attempted to apply AI approaches in parallel, collaborating with two groups from the Computer Sciences Department at BSC. After two years of laying out the biological question, the application of AI in our research wasn't successful. However, from one of these collaborations, we developed a statistical strategy to classify complex rearrangements across different cancer types. All these results were part of the genomic characterization of DNA alterations in cancer in this thesis and are included in a manuscript which is currently under review.

From the third year of my thesis, we collaborated with Dr. De Mattos, covering the characterization of transcriptomes from patients with metastatic breast

cancer. With the aim of identifying the transcriptional alterations associated with metastasis, I performed two different approaches in parallel, analyzing the RNA-seq data from 82 metastatic samples from 10 patients. On one hand, I performed a gene expression analysis to find the transcriptional signatures in metastasis. On the other hand, we wanted to comprehensively characterize the transcriptomes of metastasis by identifying new transcripts in the form of fusion transcripts. We decided to develop the framework to fusion transcript prediction, allowing us to discover new insights into the transcriptomic landscape in breast cells.

Abstract

Human cancer arises as a result of genomic alterations that transform cells and make them to grow without control and to pathological levels. The characterization of such genomic changes has enabled understanding tumor development and identifying clinical biomarkers for prognosis and therapy. Many of the genomic and epigenomic alterations in cancer can also be observed through the analysis of the transcriptome, which gives a more functional approach. Next-generation sequencing technologies, such as whole-genome sequencing (WGS) and RNA-seq, have provided the opportunity to assess molecular characterization of distinct tumors, leading to the discovery of several molecular aberrations linked to the biology of tumors. These molecular alterations include cancer-driving mutations, complex chromosomal rearrangements, atypical transcriptional profiles, gene fusions, among others.

Despite the efforts to generate a comprehensive molecular atlas of tumor biology, there are still many unknown aspects, mainly due to technical and methodological limitations. In particular, regarding the genomic characterization of cancer genomes, structural variation is known to play a crucial role in carcinogenesis as a major component of cancer genome architecture. However, the full spectrum of complex rearrangements is still largely unknown due to the lack of straightforward frameworks that allow their identification, classification, and comprehensive characterization.

On the other hand, regarding the transcriptomic characterization in tumor progression, changes in gene expression have been found to be related to metastasis. Nevertheless, the entire landscape of mRNA alterations has not been explored, opening a gap in the discovery of new transcripts, such as fusion transcripts, that might be associated with metastasis and evolve into predictive markers.

In this thesis, we address these two limitations through two studies to better characterize structural genomic and transcriptomic alterations that contribute to tumor development.

With the particular aim to classify and isolate complex somatic rearrangements in the cancer genome, with a potential molecular mechanism behind, we first investigated the landscape of chromosomal rearrangement from 2,586 tumor genomes across 40 cancer types (from the ICGC-TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium). We developed a novel framework for clustering, classification, and characterization of SV patterns with distribution that match with potential specific molecular mechanisms behind, rather than with random distribution. As a result, we identified complex rearrangements likely triggered by single catastrophic events, such as Chromoplexy or Chromothripsis. Among these, we identified a new pattern that we named Chromotrikona, which involves reciprocal translocations between different chromosomes. These findings contribute to the understanding of the genomic structural processes leading complex genomic reorganizations with impact in cancer.

To search for RNA alterations associated with metastasis, we further conducted a study of RNA alterations in metastatic breast cancer, specifically in differential gene expression and fusion transcripts. We collaborated with Dr. De Mattos from IrsiCaixa, analyzing RNA-seq datasets of 82 metastatic breast samples obtained from 10 different patients, to search for evidence that could determine the metastasis in various tissues. Running in parallel, each one of these lines has provided interesting results. In particular, in the context of fusion transcripts, we have generated and applied a comprehensive strategy to identify and characterize transcript fusion events. The study has revealed a new pattern of massive multi-fusions of different transcripts in most of the metastases analyzed. This pattern has been experimentally validated and characterized. Surprisingly, it has been found beyond metastatic samples,

being also present in normal breast cells. Although it is not clear the origin and the real nature of this event, we hypothesize that it might occurs at RNA level and be tissue specific. Further analyses will be needed to provide a deeper insight of the impact of this event in human cellular phenotypes.

Contents

Abbreviations

AI	Artificial Intelligence
Alt-EJ	Alternative End-Joining
BAC	Bacterial Artificial Chromosomes
BAM	Binary Alignment Map
Bp	base pair
BSC	Barcelona Supercomputing Center
CGP	Cancer Genome Project
CLL	Chronic Lymphocytic Leukaemia
cDNA	complementary DNA
DNA	Deoxyribonucleic Acid
DSB	DNA double-strand break
ECM	Extracellular matrix
FE	Fold Enrichment
FoSTeS	Fork Stalling and Template Switching
HGP	Human Genome Project
HR	Homologous Recombination
ICGC	International Cancer Genome Consortium
KDE	Kernel Density Estimation
lncRNA	Long non-coding RNA
LOH	Loss Of Heterozygosity
MDS	Multidimensional Scaling Analysis
MMBIR	Microhomology-Mediated Break-Induced Replication

MMEJ	Microhomology-Mediated End Joining
NGS	Next-Generation Sequencing
NAHR	Nonallelic Homologous Recombination
NHEJ	Non-Homologous End Joining
PCAWG	Pan-Cancer Analysis of Whole Genomes
PE	Paired-End
RNA	Ribonucleic Acid
RNA-seq	RNA sequencing
RT	Reverse Transcriptase
RT-PCR	Reverse Transcription Polymerase Chain Reaction
SNVs	Single Nucleotide Variants
SSA	single-strand annealing
ssDNA	single-stranded DNA
SVs	Structural Variants
TCGA	The Cancer Genome Atlas
TPM	Transcripts Per Million
WES	Whole-exome Sequencing
WGS	Whole-genome Sequencing
SE	Single-End
SNV	Single Nucleotide Variant
WHO	World Health Organization

1 INTRODUCTION

1.1 The complexity of cancer and the growing field of precision medicine to tailor therapy

Global health is affected by numerous diseases which impact the quality of life and life expectancy worldwide. According to estimates from the World Health Organization (WHO), in 2020, cancer was responsible for nearly 10 million deaths per year, ranked as the second leading cause of death worldwide[1].

Cancer encompasses over 100 distinct diseases with heterogenous risk factors locates in different organs and subtissues and originates from different cell types[2]. Additionally, some cancer types (e.g., breast, colon, and Hodgkin's lymphoma) are under specific classifications according to their molecular subtypes[3–5]. Furthermore, there is variability within the same tumor depending on the particular location or stage of cancer.

This complexity and heterogeneity of cancer lend themselves to the field of precision medicine for its treatment. Precision cancer medicine is designed to develop therapies for a single subject or subject group based on unique characteristics from patients. These characteristics emerge from the rapidly expanding knowledge about the roles of genomics, other omics (transcriptomics, metabolomics, proteomics, etc.), and the immune system in cancer. The paradigm shift to tailor therapy in cancer has been based on the development of specialized treatments for each specific subtype of cancer, according to the unique molecular landscape[6].

Thus, the characterization of the patients from multiple omics levels allows the discovery of diagnostic and prognostic markers to detect and estimate cancerogenic risk. Taken together, all this information will ultimately serve to better predict how patients will respond to a particular treatment.

1.1.1 Next-generation sequencing and the omics: finding biomarkers and molecular mechanisms behind tumor biology

Since the milestone discovery of the DNA structure[7,8], significant advances have been achieved in understanding the complexity and diversity of the genome in health and disease. In the genomic field, the major accomplishment has been sequencing the human genome through international endeavors such as the Human Genome Project (HGP)[9–11]. This was a major undertaking by the International Human Genome Sequencing Consortium (IHGSC), in which over 200 collaborating labs in 19 countries discovered new information about the structure and organization of the genome (started in 1990 and ended in 2003). This data was obtained using bacterial artificial chromosomes (BAC) and Sanger sequencing. BAC vectors enable to determine the chromosomal location of DNA fragments, whereas Sanger sequencing facilitates their base-by-base identification.

Although, these methods were essential in early sequencing efforts, the completion of the HGP evidenced the limited throughput and the high cost of sequencing as major barriers to address the plethora of biological questions that arose. Next-Generation Sequencing (NGS) technologies have emerged to overcome these problems, providing cost-effective tools capable of high-throughput, in-parallel DNA sequencing. Millions to billions of DNA nucleotides can be sequenced in parallel, yielding substantially more throughput and minimizing the need for the fragment-cloning methods as used with Sanger sequencing[12].

1.1.1.1 High-Throughput Sequencing Technologies: The Illumina system

Different NGS technologies have evolved over the past 15 years, including sequencing by hybridization, mass spectrometry sequencing, sequencing by nanopores, and sequencing by ligation[13]. More recently, DNA sequencing by synthesis approaches have been widely explored, which detect single nucleotides as they are incorporated into growing DNA strands during the polymerase reaction.

The Illumina system accounts for the most extensively used method for sequencing compared to other platforms[14,15]. It is a sequencing by synthesis approach, in which DNA polymerase catalyzes the incorporation of fluorescently labeled deoxyribonucleotide triphosphates (dNTPs) into a DNA template strand, over sequential cycles of DNA synthesis. During each cycle, the incorporation of the nucleotides is identified by fluorophore excitation across millions of DNA fragments in a massively parallel fashion. The use of dNTPs ensures that for each cycle, a single nucleotide is incorporated in the elongating complementary strand; the ribose 3′-OH group is blocked, thus preventing elongation[16]. It provides high accuracy and high yield of error-free reads into the base calls.

Briefly, there are three essential steps in the Illumina NGS technology[17]. First, a (i) *Library Preparation* is carried out by a random fragmentation of the DNA sample, followed by 5′and 3′adapter ligation. These adaptors are further useful in the second step to fix the DNA fragments into a flow cell containing surface-bound oligos complementary to the library adapters. Generally, before loading the library into the flow cell, the adapter-ligated fragments are PCR amplified in order to concentrate the sample. In the second step named (ii) *Cluster Generation*, each DNA fragment is amplified into distinct clonal clusters by bridge amplification; then, the templates are available for (iii) *Sequencing*. This

last step is composed of sequential cycles, where during each cycle, a mixture of all the four labeled and 3′-blocked dNTPs are added. After incorporating of a single dNTP to each growing strand, the flow cell is imaged to identify which dNTP was incorporated at each cluster. The fluorophore and blocking group are then removed, and a new cycle begins. The “n” times of repeating the cycle, the sequencing read length of “n” bases will be obtained.

Additionally, the DNA fragments can be sequenced from one end (Single-read sequencing) or both ends matched to each other (Paired-end sequencing). These two sequencing strategies enable to perform different sequence data analyses. For instance, Paired-end sequencing generates more accurate alignable sequence data, facilitating the alignment algorithms mapping the reads over the reference sequence (Figure 1).

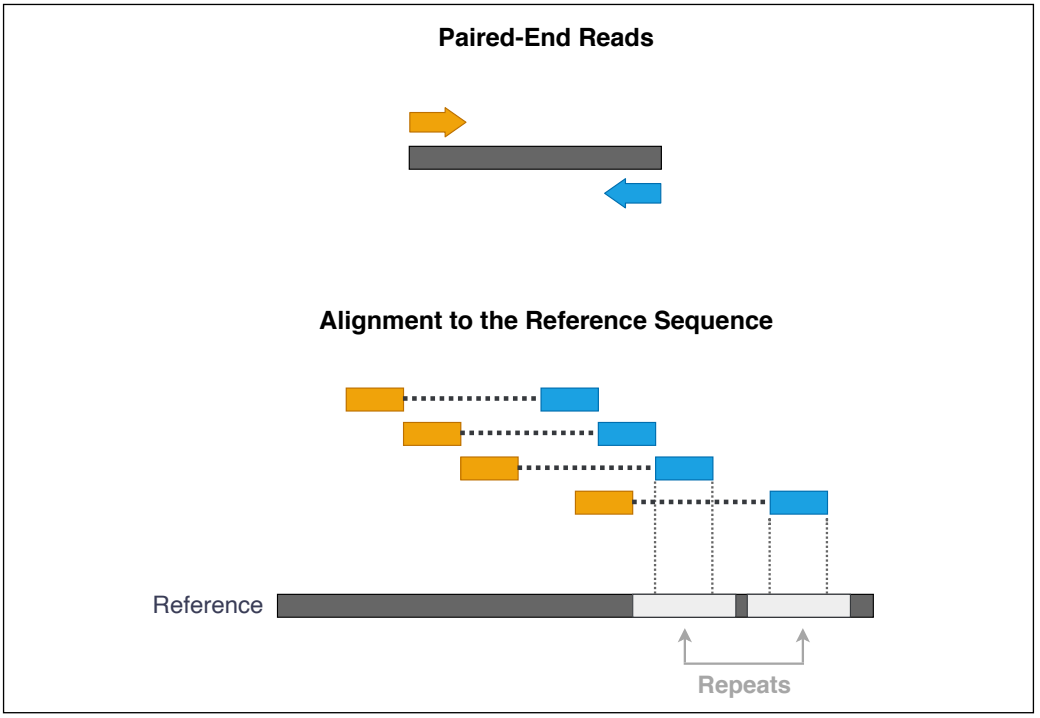


Figure 1. Paired-end sequencing leads to sequence both ends of the DNA fragment. Since the distance between the two paired reads is known, alignment algorithms can better map the reads over low complex reference sequences, such as repetitive regions. Image adapted from reference[18].

Thus, paired-end sequencing allows discovery applications such as detection of genomic alterations, repetitive sequence elements, gene fusions, and novel transcripts.

1.1.1.2 Omics technologies to study cancer development

With this advance in sequencing technologies, genomic approaches have been widely adopted in biomedical research, successfully identifying several gene and loci alterations involved in tumor development[19–22]. In addition, different molecular mechanisms have been identified associated with the myriad of acquired genetic alterations in neoplasia. These findings have revealed insights for new approaches to cancer diagnosis, treatment, and prognosis[23–27].

The rapid progress in NGS technology has provided the development of different NGS-based strategies regarding the size of the interrogated genome. These strategies include capturing the few protein-coding regions of a selected panel of genes (tens to hundreds), targeting the entire genetic code of an individual, which is called whole-genome sequencing (WGS), and sequencing the exonic regions, called whole-exome sequencing (WES)[28–30].

To better characterize tumor cells and their functional abnormalities, other high-throughput omics technologies evolved to study other biomolecules, such as epigenomics, focus on chromatin structure for epigenetic markers, proteomics for proteins and peptides, metabolomics for low-molecular-weight metabolites and transcriptomics for expressed RNA molecules (transcriptome). Mainly, NGS-based transcriptome analysis (RNA-seq) has been extensively used to profile and quantitatively analyze the transcriptome of cells and tissues, determining how gene expression can be used to diagnose cancer and identify underlined molecular mechanisms in disease development and progression. Several studies have revealed particular gene expression profiles that can help prognosis or recurrence risk and predict treatment response for different cancer types, such as breast cancer[31], colorectal cancer[32], glioblastoma multiforme[33],

and non-small lung cancer[34]. More recently, other studies have extended gene expression analysis to single cells[35], providing insights into cell heterogeneity in cancer[36] that may further lead the clinical decision-making.

Overall, this tremendous progress in omics methodologies has opened a new gate towards better understanding the underlying molecular mechanisms contributing to tumor development. Recent advances in sequencing technologies now allow the effective analysis of individual tumors, even at the single-cell level. The growing set of data derived from these efforts serves as a main pillar to foster the new era of precision medicine.

1.2 Cancer and genome

At the dawn of the twentieth century, the German cytologist Theodor Boveri (1862-1915) proposed that malignant tumors are caused by certain chromosomal aberrations that provoke cells to divide uncontrollably[37]. During the following decades, this concept competed with the notion that cancer was primarily caused by viruses such as Rous Sarcoma Virus (RSV), discovered by Peyton Rous (1879-1970). Rous found that this virus could induce cancer in healthy chickens by injecting of cell-free filtrates of tumor extracts[38]. Further studies showed that the viral tyrosine kinase encoded by the *SRC* gene (*v-SRC*, from viral *SRC*) transferred malignant properties to normal avian cells, promoting the neoplastic transformation[39]. In 1976, Bishop and Varmus made the pivotal and unexpected discovery that the normal chickens harbored a structurally closely related *SRC* gene (*c-SRC,* cellular *SRC*)[40]. This finding changed the idea about the origin of cancer, from considering foreign substances as its leading cause (viral origin), to the notion that cancer might arise from the altered function of genes present in normal cells.

The discovery that mutated cancer-causing genes, termed *oncogenes*, were in fact variants of normal cellular genes (*proto-oncogenes*), suggested that mutations in the genome might yield a mechanism in which the tumor-promoting function of proto-oncogenes could be activated. This notion was also reinforced by the demonstration that agents that cause damage in the DNA, generating mutations, also act as carcinogens[41].

In human cancer, the identification of the first naturally occurring oncogene emerged from studies of the *HRAS* gene in bladder cancer. DNA sequencing experiments led to confirm that a single nucleotide change at codon 12 had the ability to transform normal into tumoral cells[42,43]. From here, cancer genomics represented the basis for studying tumor development in human cells.

Additionally, another class of genes with a key role in cancer induction was revealed, named *tumor-suppressor genes.* A classic case is retinoblastoma, caused by loss of function of *RB*, the first tumor-suppressor gene to be discovered. In 1971, Alfred Knudson demonstrated that not just the activation of oncogenes could cause cancer, but also the loss of gene function trigger cancer. Knudson studying hereditary retinoblastoma in childhood, found that the protein encoded by the *RB* gene was inactivated when both alleles were mutated, resulting in tumor development[44]. Since then, tumor-suppressor genes are known as the genes that encode proteins involved in inhibiting cell proliferation.

Although all these genomic studies also suggested a complex mutational burden in cancer, with a variety of genes contributing to carcinogenesis across and within tumor types, all cancer types share some characteristics that represent the hallmark of this disease. Most tumor diseases develop as a result of genomic abnormalities of the cell genomes, followed by a natural selection that allows them to grow, proliferate, and spread throughout tissues[2]. The mutations acquired and accumulated over the lifetime of the cells are called

somatic variants[45], thus discriminating them from *germline mutations,* which are those inherited from parents and transmitted to offspring. Most somatic mutations are not directly associated with a growth advantage (known as *passengers*), but only a small fraction confers a selective advantage on the cell, rising survival, and proliferation (named *driver mutations*)[46].

Identifying and characterizing somatic events associated with cancer has become fundamental to understanding the biology behind tumor development and identifying the underlying molecular mechanisms and potential biomarkers.

1.2.1 Somatic variation in cancer: understanding tumor development

The longstanding recognition that genomic alterations can give rise to cancer prompted the studies to focus on discovering somatic mutations. However, somatic mutations are not exclusive of cancer genomes, as they can occur in the genome of all dividing cells, both normal and neoplastic[45]. Additionally, somatic mutations can arise through exposure to exogenous or endogenous mutagens.

The repertoire of somatic variants that a tumoral cell harbors, can be identified by comparing its genome with the genome of 'normal' (non-mutated) cells from the same patient, where somatic mutations are absent[47]. Using this approach, a variety of somatic mutations has been discovered in cancer genomes, evidencing that its distribution exhibits substantial heterogeneity among cancer genomes.

1.2.1.1 International initiatives and cancer projects

As mentioned above, the sequencing of the human genome and the emergence of NGS technologies have enabled the detection of thousands of mutations in single samples and large cohorts, helping to define the genomic landscape of cancer.

Cancer genomics engaged systematic studies of (some or all of) the genome to discover loci of recurrent derangement in specific cancer types. Founder genomic studies at the Sanger Institute and Johns Hopkins uncovered genes altered frequently in melanoma[48] and colon cancer[49], respectively. Next, several studies in lung cancer found frequent activating mutations related to patient treatment outcomes[50–52].

In the early 2000s, the Wellcome Trust Sanger Institute launched the Cancer Genome Project (CGP, https://www.sanger.ac.uk/group/cancer-genome-project/). This project identifies the somatic mutations through sequencing the human genome with NGS, aiming at characterizing genetic changes and mutational processes in tumors. As its outcome, a public database of genomic changes across cancer types is now available. Somatic mutations discovered in this project and additional related projects are currently included in COSMIC[53], the most comprehensive database of somatic mutations involved in human cancer.

In the United States, The Cancer Genome Atlas (TCGA) program[54] started in 2006 through the joint effort between the National Cancer Institute and the National Human Genome Research Institute. The TCGA program has collected genomic, epigenomic, transcriptomic, and proteomic data, characterizing over 20,000 primary cancer and matched normal samples across 33 cancer types. With integrative analyses of this data, insights into new biomarkers for

prognosis and therapy, and genes related to hereditary malignancies, have been emerged.

Following the launch of the CGP and the TCGA projects, the International Cancer Genome Consortium (ICGC)[55] was launched to coordinate large-scale cancer genome studies in 50 different cancer types and/or subtypes across the world. Specifically, the ICGC coordinates 90 projects from 16 countries and two European consortia (Figure 2), in which more than 25,000 cancer genomes at the genomic, epigenomic, and transcriptomic levels have been analyzed.

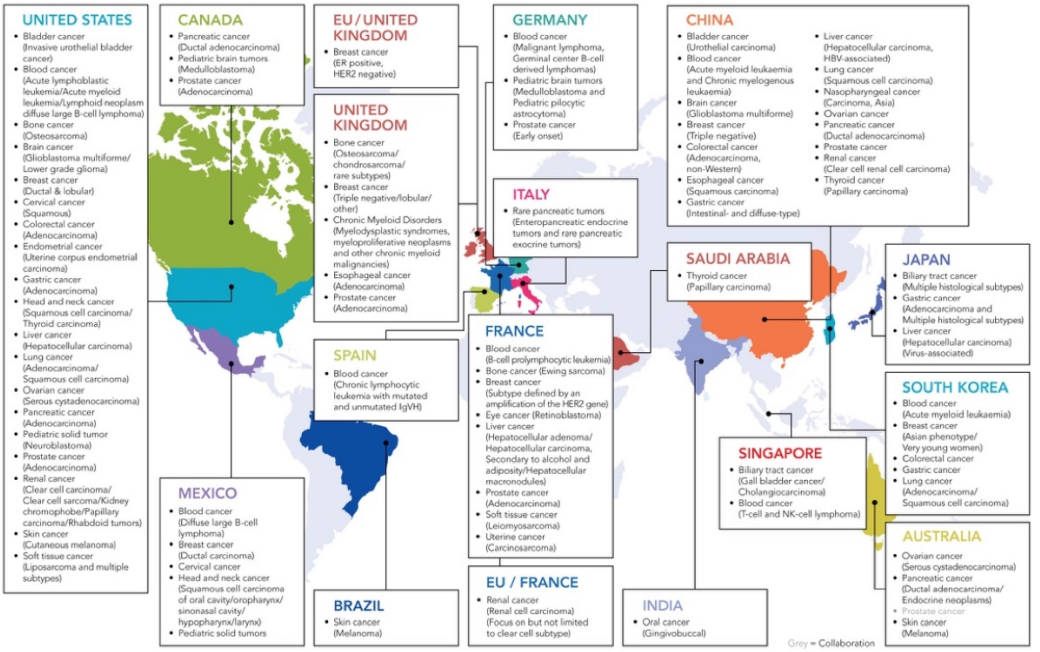


Figure 2. Countries and projects contributing to the ICGC. Map displaying the cancer projects from 16 countries included in the ICGC. Image taken from https://icgc.org.

The main goal of this international initiative was to cover the following aspects: (i) to avoid duplication of efforts or incomplete studies from independent cancer genome initiatives; (ii) to fill the gap of method standardization between studies, providing datasets to be compared and merged; and (iii) to expand the dissemination of data sets and methods across research groups.

With the increase in WGS studies from individual ICGC and TCGA working groups, the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium was established in 2013 to undertake integrative analyses of genomic features across tumor types[56]. Here, 2,658 whole-cancer genomes and their matching normal tissues across 38 tumor types were analyzed through different fronts, focused on the impact of somatic and germline variations in tumor development. Several outcomes arose from the PCAWG consortium, including mutational catalogs across cancer types gathered into the ICGC Data Portal and, standardized frameworks and different bioinformatic tools to identify genomic variants and other types of cancer-related alterations.

Overall, all projects and international collaborations have revealed a comprehensive repertoire of oncogenic mutations, uncovering mutational patterns and their underlying mechanisms. Together, these insights can define clinically relevant prognosis and therapeutic management features and provide the key to developing new cancer therapies.

1.2.1.2 Categories of genetic variants

Genomic variation could be classified according to its size, where two major categories emerge[57]. The first class of variants encompasses small variants (Figure 3A): (i) a substitution of one base by another, named Single-Nucleotide Variant (SNV) and (ii) Indels, which refers to short insertions and deletions smaller than 50bp.

The second category corresponds to large-scale structural alterations (Figure 3B), known as Structural Variants (SVs), referring to chromosomal rearrangements[58]. In terms of copy number change, these rearrangements can be balanced or unbalanced.

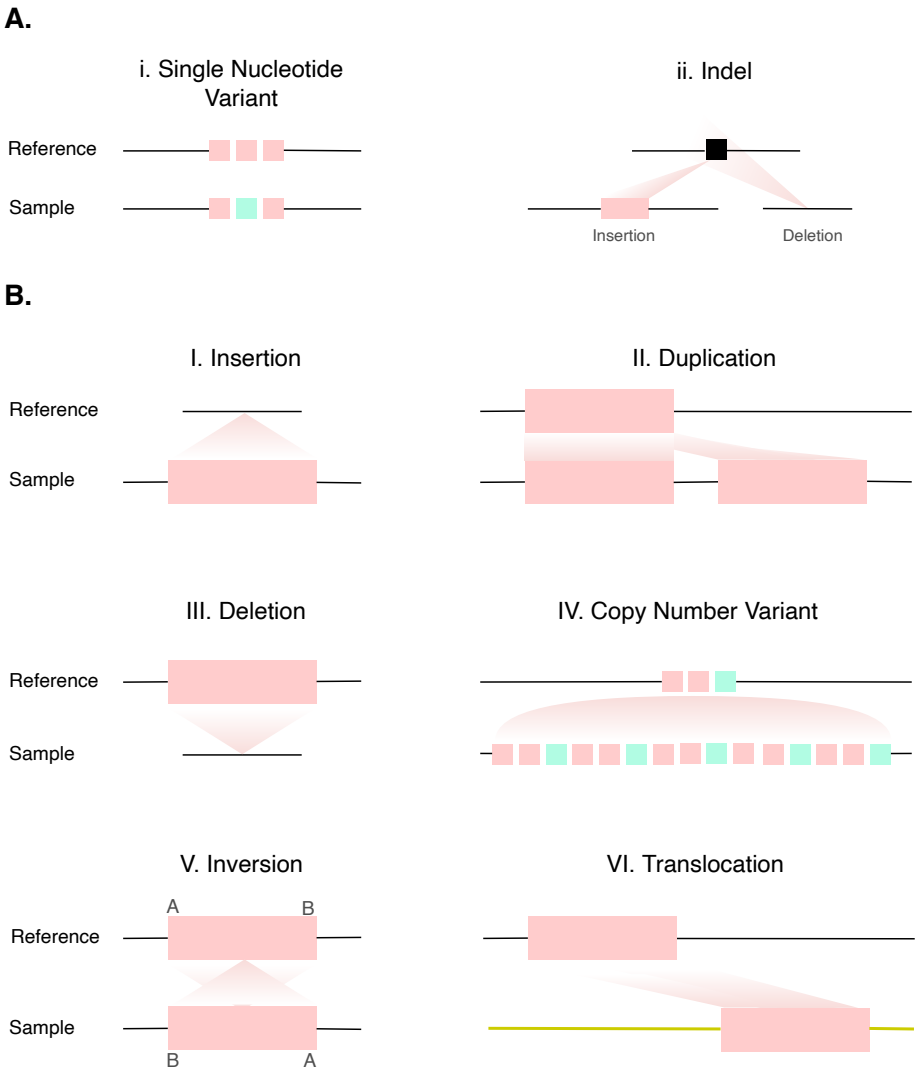


Figure 3. Types of somatic variants. A. Small Variants. (i) Single Nucleotide Variant (SNV) and, **(ii)** Indel, corresponding to small deletions or insertions. **B. Structural Variants (SVs)**, corresponding to large chromosomal rearrangements including **(I)** Insertion, **(II)** Duplication, **(III)** Deletion, **(IV)** Copy Number Variant (CNV), **(V)** Inversion and **(VI)** Translocation.

The SVs occurring intrachromosomally associated with a copy number gain are: (I) Insertions, which refers to an insertion of a segment of DNA into the chromosome; and (II) Duplications, consisting in extra copies of some genomic region. Opposite, the rearrangements that involve copy number loss refer to (III)

Deletions, where a loss of a genomic region occurs. Another SV involving copy number changes is (IV) Copy Number Variation (CNV), in which the number of copies of a particular chromosomal region increases from the two normal copies of a diploid genome to several copies, or contrary, decreases, resulting sometimes into a complete loss of DNA fragments (loss of heterozygosity).

In the case of balanced chromosome rearrangements, (V) Inversions are SVs in which a chromosome segment breaks off and reinserted in the same locus but in the reverse direction.

The SVs involving two different chromosomes (interchromosomal rearrangements) correspond to (VI) Translocations, where a chromosome breaks and a portion of it attaches to another chromosome. Translocations can be unbalanced, but sometimes pieces from two chromosomes swap places with each other (balanced translocations).

Besides these well-known types of SVs, the SV landscape has been considerably expanded during the last decade due to NGS, and more precisely, the development of algorithms to detect SVs through pair-end reads[59,60]. This approach has led to detect large-scale and more complex structural variation at high resolution, not previously appreciated.

1.2.1.3 The landscape of the SVs

As mentioned above, genomic alterations associated with the offset or progression of tumors consist of a broad spectrum of types and sizes. These vary from single nucleotide variants (SNVs) to larger structural variants (SVs) that affect genome organization. SVs are major contributors to the overall landscape of genomic variation, impacting more base pairs in the genome than all SNVs[61]. Several studies have demonstrated the implication of SVs in severe human disease, including both solid tumors and hematopoietic malignancies[62–67]. Indeed, these somatic alterations have different avenues by which they can

facilitate tumor development. SVs can potentially hit cancer-related genes, by directly disrupting tumor suppressors or by activating proto-oncogenes, for example through the generation of multiple gene copies that increase their expression. It has been reported that, in cancer genomes, those genomic regions with copy number gains (or amplified regions) are enriched for proto-oncogenes, whereas the copy number lost regions are enriched in tumor suppressor genes[68]. Beyond the impact of SVs in the direct amplification or deletion of genomic regions with genes, these variants can also modify the structure of the genome, for example by disrupting the boundaries of topologically associated domains (TADs), leading to enhancer hijacking[69] and gene fusions[70].

These implications highlighted the role of a wide spectrum of somatic SVs in carcinogenesis. Gathering clinical features and structural variations yields the opportunity for cancer diagnosis, tumor stratification, prognosis, and precision treatment[71]. Although there are different categories of chromosomal rearrangements, the formation of SVs is commonly encompassed in the occurrence of DNA double-strand breaks (DSB) and improper repair or rejoining of broken chromosomes[72]. The number of breakpoints involved, and the rearrangement patterns are two significant features for classifying structural variants. In terms of the number of breakpoints, structural variants can be classified into simple (as mentioned in the above section, i.e., deletions, duplications, inversions, and translocations) or complex SVs (as further described in section 1.2.1.5, i.e., chromothripsis and chromoplexy among others). In turn, these rearrangement patterns can also include “cut-and-paste” events, such as reciprocal inversions and balanced translocations, or “copy-and-paste” patterns as duplications.

However, cancer genomes often harbor hundreds of somatic chromosomal rearrangements, a complex scenario in which many of them cannot be easily classified into simple or complex structural variant classes. Our understanding

of the classification of chromosomal rearrangements and their accurate inference, thus, remains incomplete. Therefore, the characterization of SVs is pivotal to provide new insights into the knowledge of the functional impact, the underlying mechanisms, and the patterns that can explain the recurrence of these aberrations in cancer genomes.

1.2.1.4 Mechanisms of chromosome rearrangements

SVs emerge from different mutational mechanisms that include DNA recombination-, replication- and repair-related processes. Chromosomal breaks are relevant intermediates in such processes: DNA double-strand breaks can arise from both cell-intrinsic and extrinsic causes. The exogenous sources include oxidative stress, ionizing irradiation and hyperosmolality. Endogenous origin mainly occurs in cell processes such as replication and transcription. In a replicating cell, a major cause of DSBs occurs when a DNA replication fork collapses, deriving in the intervention of nucleases on the stalled fork[73,74]. Beyond that, in normal physiology, nucleases can also act in several additional ways producing DSBs. For instance, site-specific DSBs intervene in chromosome rearrangements during meiosis[75] and also promote V(D)J recombination in developing lymphocytes during immunological cell maturation[76]. Other intrinsic origins include the decline of telomere protection during aging, or malfunctional complexes that lead to free DNA ends[77].

Thus, in normal physiology, the generation of DSBs is a quite frequent event. To compensate, the cell monitors the genome integrity and actives its DSB repair machinery, which could be dysfunctional and repair improperly. There are three main groups of DSB repair pathways through which SVs can arise: (1) homologous recombination (HR), (2) non-homologous end joining (NHEJ) and (3) replication-based mechanisms (Figure 4).

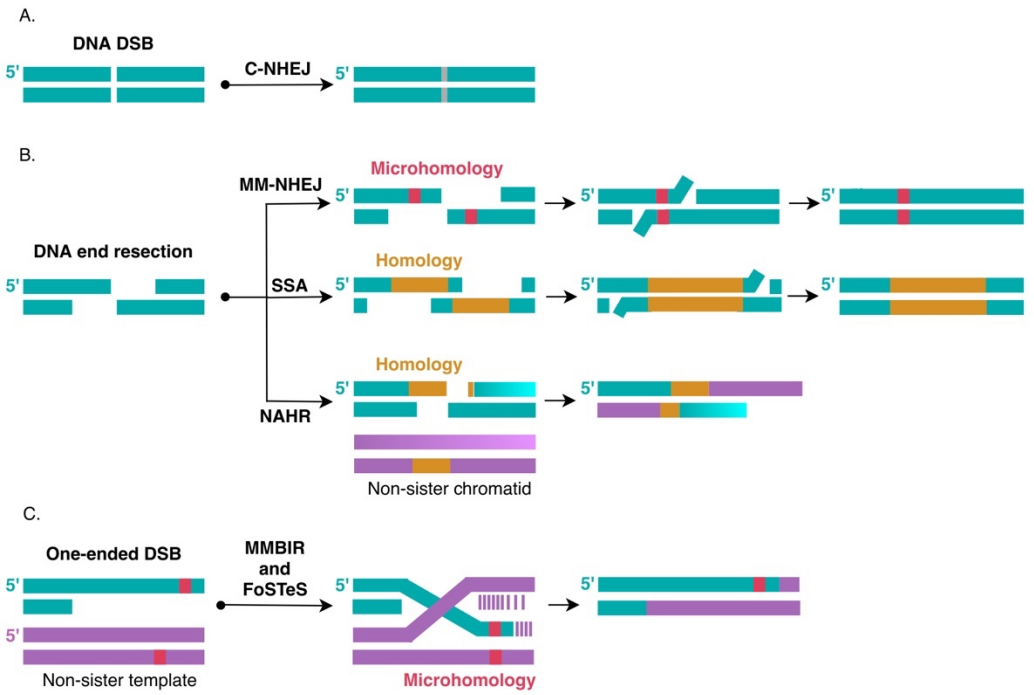


Figure 4. DSB repair pathways involved in the formation of structural variation. When a DSB arises, the cell actives mechanisms to repair it. **A.** The canonical NHEJ (C-NHEJ) repairs DSBs by ligating the two DNA ends, sometimes introducing insertions or deletions at the ligation site. **B.** DNA end resection could be repaired by three mechanisms: when the 3' ssDNA overhangs have microhomology, they anneal to guide the repair by microhomology-mediated end joining (MMEJ). Moreover, when the homology is extended, single-strand annealing (SSA) acts to repair the broken chromosome. In contrast, the third mechanism, nonallelic homologous recombination (NAHR), uses homology recombination (HR) to repair the end resection, sometimes taking a nonallelic locus as a template and therefore, introducing structural alterations. **C.** Besides two-ended DSBs, the DSB repair system deal with one-ended breaks, resulting from broken replication forks, by invading the 3′ ssDNA end to the donor chromosome using microhomology and subsequent replication. Error-prone mechanisms could derive, named Microhomology-Mediated Break-Induced Replication (MMBIR) and the Fork Stalling and Template Switching (FoSTeS).

1. *Homologous recombination*

HR is the most common DNA repair mechanism, which uses the sister chromatid as a template to synthesize the missing DNA strand in the DNA break. This mechanism is generally accurate when the correct

template is used. However, the pairing could also happen with other highly similar sequences (e.g., repeats) by nonallelic homologous recombination (NAHR). Thus, once a DSB occurs, homology search may find a nearby paralog of the DNA strand, leading to NAHR and further rearrangements: deletions, duplications, inversions, or translocations[78,79]. NAHR events are enriched in open chromatin, suggesting that these DSBs arise in the cell transcription process[80]. The type of rearrangement emerging depends on the spatial proximity between the two homologous sequences and the orientation of the paralog with respect to the DSB[79].

Furthermore, deficiency in homologous recombination appears to be a major trigger of cancer genome instability[81]. This deficiency lies in affectations or faulty mechanisms. For instance, it has been shown that loss-of-function mutations of *BRCA1*, *BRCA2*, Fanconi anemia genes, and several other genes required in HR, are implicated in cancer predisposition in the human population[82,83].

2. *Non-homologous end joining and alternative end-joining repairs*

The canonical NHEJ (C-NHEJ) is a nonreplicative rapid repair mechanism by which the two DNA ends are joined. During the DNA ligation process, small insertions or deletions at the junction can occur, originating these kinds of structural variations.

The C-NHEJ is a process that is generally effective except when the DNA ends encompass single-stranded DNA (ssDNA) overhangs. Here, an alternative end-joining (Alt-EJ) mediates the repair[84] by using a microhomology-mediated end joining (MMEJ). One or more complementary base pairs shared between the two 3' ssDNA

overhangs, anneal to guide the repair [85]. If the ssDNA overhangs share extensive homology, this can bridge DSB ends by single-strand annealing (SSA)[86]. In both MMEJ and SSA processes, the ssDNA tails that do not anneal are digested away, leading to deletions as the most frequent mutations in these error-prone mechanisms.

3. *Replication-based mechanisms*

The third DSB repair mechanism is also related to microhomology but during replication. In this regard, a stalled or collapsed replication fork breaks, generating a one-ended DSB (only one DNA end instead of two DNA ends), which can trigger break-induced replication (BIR), a highly error-prone HR-mediated replicative response. Here, using microhomology, the 3′ ssDNA end invades the donor chromosome, normally the sister chromatid, which serves as a template to replicate the missing DNA.

However, the microhomology can be used to invade another DNA template besides the sister chromatid, through Microhomology-Mediated Break-Induced Replication (MMBIR)[87]. Moreover, in fork stalling and template switching (FoSTeS), a few rounds of microhomology-mediated template switching can drive the strand invasion of non-sister templates using microhomology-containing regions, accounting for chromosomal rearrangements[88].

Then, the failure to repair by BIR and microhomology-mediated (MM) template switching can account for MMBIR and FoSTeS, giving rise to single duplications, translocations and more complex chromosomal rearrangements[89].

1.2.1.5 Beyond single SVs: Complex genome rearrangements

The typical transformation of a normal cell into a tumor cell generally involves a sequence of independent events that converge into an uncontrolled cell growth. However, there are reasons to consider other scenarios, where many genomic alterations are acquired all at once. This scenario could be advantageous for the tumoral cell, shortening the time needed to reach a large phenotypic effect resulting from the combination of smaller-effect mutations. Another possible advantage is that, by acquiring mutations through a "single-hit" event, the cell can circumvent the deleterious fitness effects of the intermediate steps in tumor development. In this regard, it has been shown that several SVs are not independent of each other, and that they are not acquired randomly and gradually. Instead, they are acquired through a single-hit event, involving several DNA breaks, usually deriving in complex genome rearrangements[90–92].

In 2011, Stephens *et al*., studying a genome from a patient with chronic lymphocytic leukemia (CLL)[91], observed a striking pattern of chromosomal rearrangements that encompassed 42 highly localized translocations on the long arm of chromosome 4. The altered region of chromosome 4 was present in only two copy number states, with many transitions between them. The alternation within the two states was accompanied by loss and preservation of heterozygosity. Therefore, the observed pattern, characterized by clustered inversions, deletions and tandem duplications, was hard to explain by independent rearrangements accumulated gradually. In contrast, a more parsimonious explanation arose under the term chromothripsis: one chromosome is shattered into chunks of different lengths in a single catastrophic event; this results in multiple DSBs, which are repaired likely by highly error-prone NHEJ mechanisms. The derivative chromosome harbors a subset of the shattered segments in random orientations, and the segments

that are not retained, are either lost or included in extrachromosomal circular DNA elements (double-minute chromosomes).

After being discovered, chromothripsis was further studied in other primary tumors, confirming that it was a widespread and previously disregarded event[93–96]. In fact, some cancer genomes previously analyzed contained features that resembled chromothripsis[97,98]. This raised the questions of how and when did chromothripsis originate. The criteria to infer chromothriptic events were three and were subjected in order to estimate its frequency across cancer types[99]. As mentioned above, 1) one hallmark of chromothripsis was the clustering of rearrangements. However, some clustered SVs could originate from other processes over several cell division cycles, such as breakage-fusion-bridge (BFB) cycles or chromosome fragile sites[100,101]. Then, 2) an additional criterion was the copy number states, with interspersed retention and loss of heterozygosity (LOH). Finally, 3) given that the altered chromosomal region was limited to a single chromatid, only one derivative haplotype should contain the rearrangements.

Recently, as part of the PCAWG Consortium of ICGC and TCGA, Cortés-Ciriano and collaborators[102] carried out an extensive survey of chromothripsis, revealing that the prevalence and heterogeneity of this pattern were beyond previously appreciated. Here, they described non-canonical chromothriptic events which involved multiple chromosomes and signatures of replication-associated processes.

The identification of chromothripsis opened the gate to explore different complex landscapes of chromosomal rearrangements, in which many challenges remain. Indeed, a different pattern of complex SVs was described in 2011 by Liu *et al*.[90] investigating subjects with congenital developmental defects. In contrast to chromothripsis, this study reported frequent copy number gains with interspersed genomic regions, with duplication or triplication

of one parental allele and without LOH. This complex rearrangement pattern was named chromoanasynthesis. The copy number gains due to the extra synthesis of DNA segments suggested MMBIR as the principal mechanism by which this pattern arose. Here, the additional allele copies were integrated through a template-switching event in DNA replication rather than by ligation of shattered fragments (as via NHEJ in chromothripsis).

Even though both chromothripsis and chromoanasynthesis affect mostly limited portions of the genome, they are different patterns of complex rearrangements that appear to have an all-at-once origin. In this regard, a third single catastrophic event was described in prostate cancer genomes, termed chromoplexy[92]. This event was characterized by a closed chain of translocations that could involve several chromosomes. The sequence hallmarks of chromoplexy suggested that all the DNA segments involved in the translocations arose from DNA DSBs, likely induced by a transcription-related mechanism, occurring in a spatially and temporally constrained fashion. In this regard, Baca *et al*. developed an algorithm to find such rearrangements called ChainFinder, which calculates the probability that two contiguous translocations have originated from independent breaks.

Overall, the discovery of all the above-delineated patterns has expanded our view about tumor evolution. In contrast to the classical understanding of a gradualism model of sequential accumulation of mutations, the acquisition of large-scale genomic alterations is likely to facilitate a rapid adaptation in the form of “punctuated” tumor evolution (Figure 5).

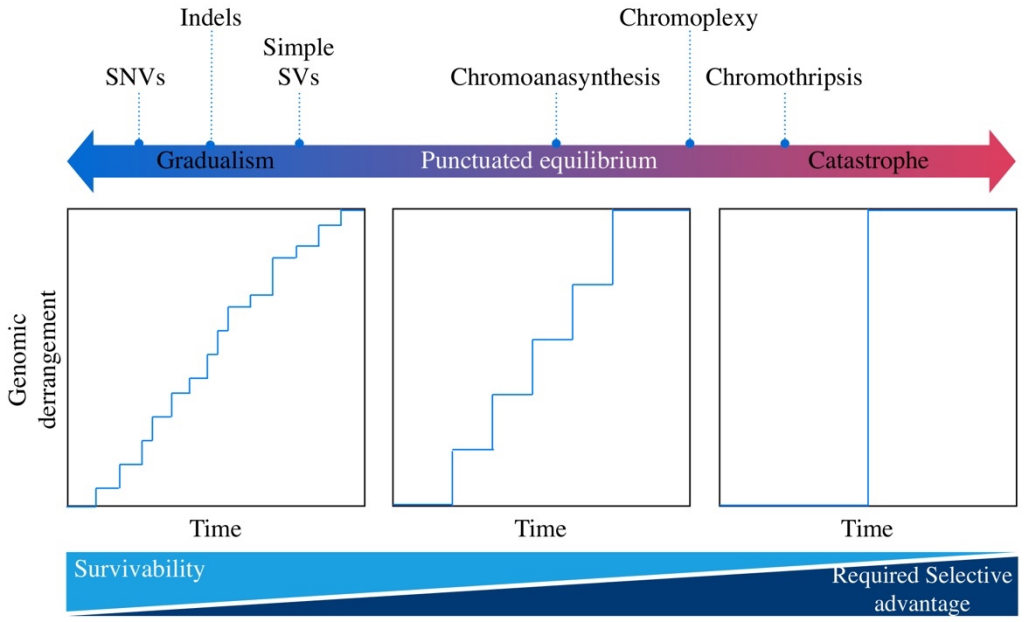


Figure 5. Model of Tumor genome evolution. Genomic alterations can be acquired gradually over several cell division cycles, but also through single-hit events, where several SVs can co-occur (as indicated at the top). In order to compensate for larger-scale rearrangements, the tumoral cell likely needs to involve oncogenic alterations (as shown at the bottom). Image adapted from Baca *et al.*[92].

However, the comprehension of the origin and the impact of catastrophic chromosomal rearrangements is incomplete, with still many unanswered questions. How common are complex rearrangements? Are there more undiscovered catastrophic events associated with cancer genomes? To address these questions, we must provide more insights regarding the mechanisms underlying these patterns, the hallmarks of such complex rearrangements, and their characterization across different cancer types.

1.3 Cancer and transcriptome

While genomic data is crucial to develop a comprehensive understanding of cancer development and treatment outcomes, bridging the gap between genotypic effect and phenotypic event is accomplished through the analysis of the data generated by other omics such as transcriptomics.

The gene expression profile in a cell or tissue reflects its functional state. The transcriptome is the set of RNA molecules, including ribosomal RNA (rRNA), messenger RNA (mRNA), transfer RNA (tRNA), micro RNA (miRNA), and others non-coding RNAs (ncRNA). Contemporary approaches for collecting transcriptomic information include microarray and RNA-seq methods, which mainly allow for measuring the RNA expression level from cells. Microarray analysis is limited because it requires prior knowledge of the gene's sequence, whereas RNA-seq does not, being useful for quantitative analysis of total RNA, as well as discovering new transcripts[103]. Within RNA-seq technologies, bulk RNA-seq has been widely used to study gene expression patterns at high level, whereas single-cell RNA sequencing (scRNA-seq) enables the exploration of gene expression profiles at the single-cell level[104]. The use of these RNA-seq technologies has largely facilitated the understanding of gene expression in diverse human tissues (including normal and cancer) at high resolutions.

1.3.1 Projects and resources to study the human transcriptome

The expressed information of the transcriptome is vastly varied through the different cells of a multicellular organism such as the human, and depends on the cell type regarding its functional and temporal state. Different projects have focused on characterizing the functional elements encoded in the human genome, and the gene expression profiles across different tissues.

Two leading projects, the Encyclopedia of DNA Elements (ENCODE)[105] and Genotype-Tissue Expression (GTEx)[106], have mapped functional elements at high resolution and the regulation of gene expression in diverse human tissue types.

The ENCODE project aims at depicting the genomic regions encoding a defined product (e.g., protein, ncRNA) or displaying biochemical signature (e.g., protein binding), denominated functional elements. Thus, this project has provided a comprehensive knowledge about the organization and regulation of our genes and genome, representing a resource of functional annotations for biomedical research.

The Genotype-Tissue Expression (GTEx) project has collected RNA sequence data from about 1000 individual across 54 non-diseased tissues, generating a comprehensive database to investigate tissue-specific gene expression and regulation. This has provided a better understanding of the heterogeneity among human tissues and, the complexity and variation in the regulation of genome expression.

1.3.2 The role of transcriptomics in cancer

Besides the above-mentioned efforts to characterize functional elements, the human transcriptome has also been subjected to understanding human health, due to alterations in gene expression are believed to contribute to diseases, including cancer. For instance, altered expression levels of specific isoforms or alleles have been identified in solid tumors, including colorectal cancer[107], kidney renal clear cell carcinoma[108], as well as hematological malignancies, such as chronic lymphocytic leukemia (CLL)[109].

The abnormal splicing and other editing events in specific cell types have been associated also with tumorigenesis[110]. In this regard, defects in alternative splicing in human tumors are mainly due to either somatic mutations in splicing-

regulatory elements (or motifs), or alterations (direct or indirect, like mutations in the oncogene *MYC*, which controls transcriptionally multiple splicing factors) in components of the splicing machinery[111].

Similarly, different classes of non-coding RNAs (ncRNAs) have been found to be associated with cancer. Given that they have key roles in gene regulation, ncRNAs are emerging as new actors in cancer development, demonstrating their potential roles in both oncogenic and tumor-suppressive pathways[112,113]. Different genetic alterations associated with ncRNAs have been found in different tumors. For instance, deletions of the miR-15/16 tumor suppressors in CLL[114]; conversely, amplifications of loci encoding oncogenic ncRNAs are also found in cancer, including amplification of long non-coding RNAs (lncRNAs) *FAL1* [115] and *PVT1*[116]. Besides somatic mutations, up- or downregulation of ncRNA expression associated with cancer (e.g., *HOTAIR, MALAT1, HULC, T-UCRs*) can take place by epigenetic, transcriptional, or post-transcriptional processes[117,118].

Furthermore, the careful examination of the tumoral cell transcriptome and its relationship to cancer progression has become critical in elucidating the functional basis of tumor invasion and spread. The most widely accepted hypothesis of tumor progression postulates that mutations are acquired over time, enriching mutations that confer metastatic capabilities. However, mutational drivers of metastasis have not been identified, and rather emerging evidence suggests an aberrant transcriptome as a driver of metastatic capacity[119–121]. Specific phenotypic traits, resulted from the expression of particular molecules in a coordinated fashion, thus mediate metastatic cancer progression. In fact, over the last decade, these transcriptional programs required for metastasis have been the focus of several studies. For example, it has been reported that in the migratory and invasive steps of the metastatic process, activation of embryonic programs occurs through the up-regulation of transcription factors[122,123]. Different studies have also subjected post-

transcriptional controls of the cancer cell transcriptome by microRNAs and other determinants of RNA stability[124,125]. Nevertheless, due to its complexity, how cancer cells arrange and sustain the transcriptional programs involved in metastasis remains incompletely understood.

Given that the majority of the human cancer-related mortality is caused by the effects of metastasis, such as in metastatic breast cancer, the detailed analysis of the transcriptome is considered of particular relevance.

1.3.3 Breast cancer and metastasis

Breast cancer is the most common malignant disease in women worldwide. According to the last global cancer statistics, about 2.1 million new cases annually diagnosed with breast cancer[126]. However, in these patients, it is not the primary tumor, but its metastases that are the leading cause of death. In the disease progression, the primary tumor spreads throughout the body to distant sites, most commonly in lung, liver, bone and brain. Approximately 30% of women diagnosed with breast cancer have an aggressive disease and develop distant metastases within 3 years after the initial detection or even 10 years or more[127].

Breast cancer generally starts with an atypical ductal hyperplasia which develops into in situ and invasive carcinomas, and finally into metastatic cancer. The carcinogenic factors involved in its development include molecular alterations, obesity, and the use of hormone therapies (progestin and estrogen)[128–131]. However, breast cancer is not a single disease, but rather encompasses distinct subtypes on the basis of gene expression profiles, associated with different clinical outcomes. Understanding this heterogeneity is crucial for the development of diagnostic markers and therapeutic interventions.

1.3.3.1 Distinguishing tumor subclasses of breast carcinomas

Given that multiple molecular alterations drive breast carcinogenesis, several classifications have been developed to group tumors accordingly. Initially, based on gene expression profiles, Perou, Sorlie and collaborators distinguished four major subtypes of breast cancer in the *intrinsic classification*[132,133]: Luminal A, Luminal B (with Oestrogen Receptor (ER) expression), Basal-like, and Human Epidermal growth factor Receptor 2 (HER2)-enriched (without ER expression) (

Figure **6**). With this classification, the clinical management of breast cancer shift from, focusing on tumor burden, to the biology of the disease.

Currently, with a *surrogate classification*, tumors can be classified into five subtypes, based on histological characteristics and immunohistochemistry expression of key proteins: ER, Progesterone Receptor (PR), HER2, and the proliferation marker Ki67. Tumors expressing ER and/or PR, are termed 'hormone receptor-positive', whereas tumors that do not express ER, PR and HER2 are called 'triple-negative' breast cancer (TNBC) (

Figure **6**).

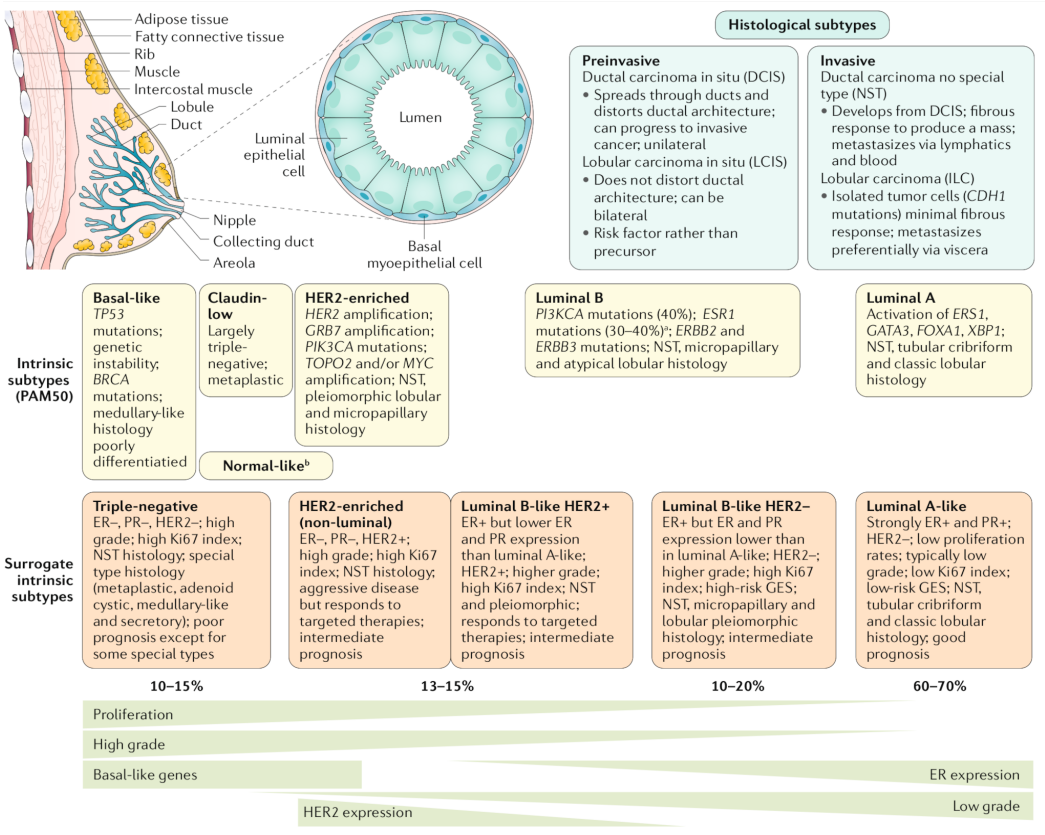


Figure 6. Different classifications of breast cancer subtypes on the basis of molecular and histological characteristics. Breast cancer arises in the terminal duct lobular units of the collecting duct. Based on the histological characteristics (top right), the most abundant subtypes are ductal carcinoma no special type (NST), which is the invasive form of ductal carcinoma in situ, and lobular carcinoma, which has its preinvasive counterpart as lobular carcinoma in situ. Based on gene expression profiles, the intrinsic classification distinguishes four different subtypes, Basal-like, Luminal A, Luminal B, and HER2-enriched (yellow boxes). Currently, the surrogate intrinsic subtypes are five, based on both, histological and molecular characteristics (orange boxes). The green arrows (in the bottom) indicate the intensity of some characteristics of the different subtypes, including proliferation, tumor grade, and expression of ER, HER2 and Basal-like genes. GES: gene expression signature. [a] ESR1mutations induced by aromatase inhibitor targeted therapy. [b]Artifact; expression of normal breast components due to low tumor cellularity. Image taken from reference[127].

These molecular differences between subtypes result in distinct clinical outcomes. In general, HER2-positive breast cancers are associated with higher mortality, followed by TNBC, while luminal A-type tumors show better prognosis[134]. Similarly, clinical therapies are based on these molecular subtypes

and their features. For instance, HER2-positive patients receive HER2-targeted antibodies combined with chemotherapy, whereas patients with TNBC receive only chemotherapy[135]. Nevertheless, despite recent progress, this systemic treatment has poor efficacy for some patients due to drug resistance caused by the heterogeneity of breast cancer cells[136].

1.3.3.2 Tumor progression: invasion and metastasis

Two current models explaining the inter- and intratumoral diversity of breast tumor cells are the *clonal evolution* and the *cancer stem cell (CSC)* models[137]. The clonal evolution model proposes that every transformed cell within a tumor has tumorigenic potential for originating a new tumor, where mutations accumulate, epigenetic changes occur, and the 'fittest' cells survive. The *cancer stem cell (CSC)* model proposes that only a small subset of cells, the precursor cancer cells, can trigger tumor progression. The CSC model has a fundamental role in therapy and drug resistance. After treatment, the survival of CSCs (or cancer cells with stem-cell-like properties) represents a high risk of disease relapse and metastasis.

Although the exact etiology of breast cancer is unknown, family history is one of the strongest determinants of risk. Approximately 10% of breast cancers are inherited, involving germline mutations in the high-penetrance tumor suppressor genes, *BRCA1* and *BRCA2*, which encode proteins involved in DNA repair through homologous repair[138,139]. Additionally, several studies have screened panels of genes to determine the inherited breast cancer risk, revealing allele variants in *ATM, CHEK2, PALB2, PTEN, STK11,* and *TP53*, as being associated with breast cancer risk[140–142]. These findings have allowed the development of cancer-preventative therapies and surveillance for early diagnosis and prevention.

The natural history of breast cancer encompasses progression across different pathological and clinical stages. As mentioned above, it starts with ductal hyperproliferation, followed by in situ and invasive carcinomas, and finally, metastatic disease (Figure 7). Several studies based on molecular, epidemiological, and pathological features have pointed out that the ductal carcinoma in situ (DCIS) stage is a precursor of invasive ductal carcinoma and metastatic spread[143,144].

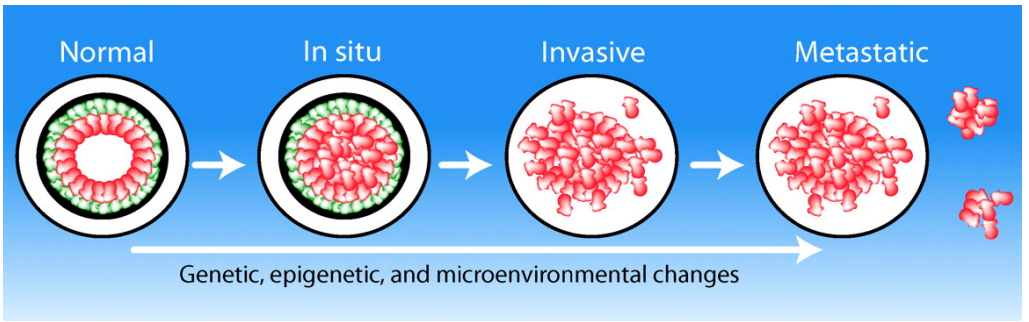


Figure 7. Model of breast cancer progression. Breast tumorigenesis evolved through different histologic and clinical stages, where genetic, epigenetic, and microenvironmental alterations drive the progression. Thin and thick black circles represent the breast and the basement membrane surrounding the ducts, respectively. Myoepithelial/basal cells (green cells) synthesize and are adjacent to the basement membrane, while luminal epithelial cells (red cells) lay on top of the myoepithelium. A characteristic of the transition from in situ to invasive carcinoma, is the absence of the basement membrane and the myoepithelial cell layer. Image taken from reference[145].

However, the factors involved in tumor progression are poorly understood. Despite tumor initiation and progression are mainly driven by acquired genetic alterations, in vivo and in vitro studies have demonstrated that epigenetic alterations and tumor microenvironment (TME) have an essential role in breast cancer progression. In this regard, cells that compose the TME, including myoepithelial and endothelial cells, fibroblasts, myofibroblasts and cells of leukocyte lineage, and the extracellular matrix (ECM) components, modulate the growth, survival, and invasive behavior of breast cancer cells[146]. The breast cancer cells exploit this microenvironment, 'hacking' the normal programs to

proliferate and escape from cancer immunoediting[147]. Immunogenicity of breast cancer varies across the different molecular subtypes, where TNBC and HER2-positive tumors have the highest and luminal A and luminal B subtypes the lowest[148,149].

Regarding epigenetic determinants in tumor spread, gene activation, upregulation of oncogenes, and chromosomal instability can occur through global hypomethylation in breast cancer. Also, genes can be focally (locus-specific) hypermethylated, resulting in gene repression from the silencing of tumor suppressors[150,151]. Other epigenetic mechanisms include histone tail modifications through DNA methylation, which lead to chromatin structure alterations, silencing gene expression and nucleosomal remodeling[152–154].

Together, the mechanisms contributing to tumor invasion and spread in breast cancer, are reflected in the altered transcriptomic profiles in metastatic disease. Thus, the characterization of cellular phenotypes and transcriptomes of metastatic breast cancer cells lead us to identify clinical biomarkers that are present in metastasis disease. In this regard, the exploration of new transcripts involved in tumor progression is understudied. The discovery of fusion transcripts significantly contributes to the comprehensive characterization of cellular transcriptomes.

1.3.4 Fusion genes and transcripts

Fusion genes, also called chimeric genes, are hybrid genes formed by joining portions of two different genes. Although fusion genes are mostly known to result from chromosomal rearrangements at the DNA level, fusion events can also occur by transcription read-through of adjacent genes[155,156], where consecutive genes on a genome strand are spliced together, or by cis- and trans-splicing of pre-mRNAs[157,158] (Figure 8).

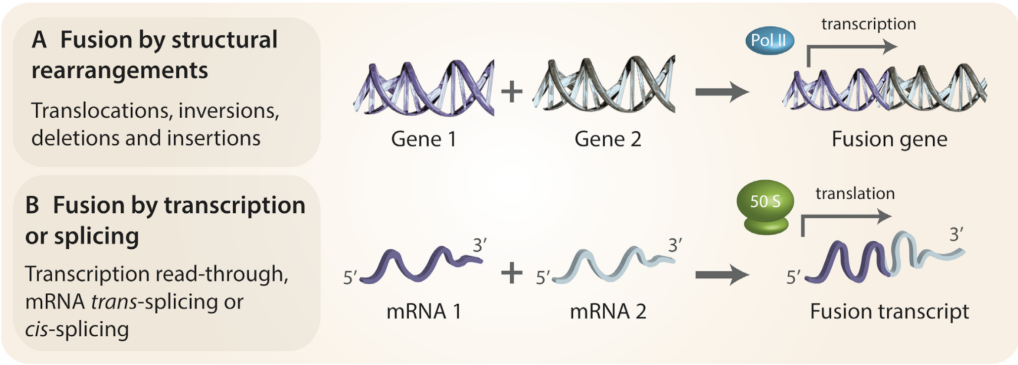


Figure 8. Mechanisms of fusion gene formation. A. Chromosomal rearrangements, such as translocations, inversions, deletions and insertions, can result into fusion genes. **B.** Non-structural rearrangement mechanisms, such as transcription read-through of neighboring genes and non-canonical splicing of mRNA molecules can also lead to the formation of fusion transcripts. Image taken and adapted from reference[159]

Fusion genes are a prototypical example of a pathognomonic alterations. Therefore, they have gained great importance in clinical management and in the understanding of tumorigenesis. The oncogenic properties of fusion genes comprise deregulation of genes by juxtapositioning promoter or enhancer regions[160] (e.g., a fusion of a strong promoter and a proto-oncogene), chimeric transcripts that produce fusion proteins with aberrant functionality[161], and induction of loss of function[162] (e.g., truncation of a tumor suppressor gene).

1.3.4.1 From cytogenetic to RNA-seq data: fusion gene discovery

The discovery of fusion genes involved in cancer diseases began with the characterization of structural rearrangements in the genomes of neoplastic cells. Through cytogenetic analyses, in which each chromosome and chromosome region could be identified based on its unique banding pattern, several studies described balanced translocations in hematological malignancies[163–167]. The first example was the balanced translocation between chromosomes 9 and 22 (t(9;22)(q34;q11)), present in the Philadelphia

chromosome in patients with chronic myeloid leukemia (CML)[163]. Further studies determined that the fusion gene *BCR-ABL1* resulting from such translocation produces a chimeric protein with constitutively activated tyrosine kinase activity, transforming benign tissue into malignant one[168,169].

The introduction of fluorescence in situ hybridization (FISH) allowed to characterize already known fusion genes, revealing that some genes were promiscuous in the sense that they recombine with multiple partners, providing evidence of the critical contribution of particular genes in tumorigenesis, such as the mixed lineage leukemia (*MLL*; also called *ALL-1*, *HRX* and *HTRX1*) gene[170]. The *MLL* gene was found to be involved in a variety of balanced translocations associated with both, lymphoid and myeloid acute leukemias. Furthermore, different *MLL* fusion proteins were associated to a particular tumor. For example, *MLL–AF9* prevailed in acute myeloid leukemias, whereas the fusion gene *MLL–AF4* was only found in B-cell linage tumors[171,172].

The development in the 1990s of Genome-Wide array platforms for gene expression and copy number profiling, enabled a guided fusion gene detection at much higher levels of resolution than cytogenetic analysis. The prime example was the discovery by Tomlins *et al.* studying prostate cancer, of prevalent fusions between the transmembrane protease serine 2 gene (*TMPRSS2*) and two genes encoding ETS transcription factors, *ERG* and *ETV1*[173]*.* This was the first time that a specific fusion gene was a distinct trait found in solid tumors.

Although, the identification of fusion genes in both, hematological and solid tumors, increased substantially by using the approaches described above, it was not until the introduction of NGS that the complex landscape of fusion genes could be studied at the DNA or RNA levels without any prior information on the cytogenetic features of the tumoral cells. Thus, a plethora of novel fusion

genes were identified in previously uncharacterized tumour types, as well as some fusions were re-discovered"[174–180].

Furthermore, NGS enabled the detection of fusion events that cannot be evidenced at the DNA level, such as the transcription-induced gene fusions (TIGFs), which result from alternative splicing involving more than one gene. For example, different studies detected in both normal and neoplastic prostate tissue, *SLC45A3–ELK4* and *MSMB–NCOA4* fusions, which are two fusions forming through alternative splicing involving neighboring partners genes located on the same DNA strand (cis-TIGFs; also known as read-through fusion transcripts)[174,181,182].

Additionally, transcriptomic profiling using NGS and the development of computational tools have resulted in an enormous increase of studies reporting fusion events in individual tumors[183]. In the majority of cases, the identified fusions seem to be non-recurrent and, therefore, have not received particular attention[184]. Nevertheless, the accurate detection of fusion events represents a challenge to identify clinically relevant fusion transcripts in cancer.

1.3.4.2 Fusion genes detection methods

As mentioned above, NGS has promoted the discovery of novel fusion genes in cancer at the DNA and RNA levels, dramatically changing the gene fusion landscape[185,186]. However, while deep sequencing is highly sensitive, it is also error-prone. Additionally, the introduction of errors might occur either, prior to sequencing, during library preparation, or later, at the bioinformatic level, during the analysis of the reads obtained, due to extensive sequence similarities between genes. Several bioinformatics tools have been developed to identify fusion events, which differ considerably in terms of sensitivity and specificity[187,188]. These differences lie in the criteria for predicting and the data used for the detection of fusions.

For example, many tools combine RNA-seq with WGS data, detecting fusion genes evidenced by expressed chimeras and associated structural variants[189–193]. These tools could fail in identifying the fusion transcripts without any signs of underlying rearrangements at the genomic level, such as the fusions resulting from alternative splicing[194].

Furthermore, these approaches are based on aligning the RNA and DNA reads to the human reference genome. Short read sequences have lower alignment specificity, particularly under the presence of SNPs, sequencing errors, and repeat regions. These short reads mapping incorrectly lead to false predictions of fusion genes. To overcome this, some algorithms have been designed for paired-end libraries to look for supporting information such as read pairs, while often require restrictive filtering that affects the sensitivity[195].

Together, these different drawbacks and limitations, on the basis of the types of sequencing data and the computational strategies implemented in gene fusion detection tools, have an impact on the quality assessment of the fusion gene landscape in cancer.

1.3.4.3 Unexploited fusion genes associated with metastasis

While the vast majority of clinically relevant fusion genes have been discovered in primary tumors, e.g., *BCR-ABL* in leukemia, *EML4-ALK* in non-small cell lung cancer, *EWS-FLI1* and *EWS-ERG* in Ewing sarcoma, fusion events have not been well characterized in metastasis[196]. The lack of known fusion genes involved in metastasis has been attributed to their intrinsic heterogeneity and the limitations of obtaining samples.

On one hand, according to the cellular processes regulated by the genes involved in the fusions, some fusions were sought to be associated with metastasis. For instance, the fusion gene *TMPRSS2-ERG* in prostate cancer

has been associated with metastasis due to the increased aberrant expression of *ERG*[197]. Similarly, the translocation t(15;19), commonly fusing the testis-specific nuclear gene *NUTM1* to the *BRD4* gene in Nuclear protein in testis (NUT)-midline carcinoma, is associated with poor prognosis and development of metastases[198,199].

On the other hand, regarding the detection of fusion events present in metastatic samples, Yu *et al.* reported the *TRMT11-GRIK2* and *CCNH-C5orf30* as frequent fusions across both primary tumor and matched lymph node metastases, from patients with breast, ovary, and colon cancer[200].

In metastatic breast cancers, some fusion events have been detected, involving *FGFR* family members fused to various proteins genes (e.g., *FGFR3-AF4/FMR2* Family, member 3 (*AFF3*), *FGFR2*-caspase 7 (*CASP7*), *FGFR2*-coiled-coil domain containing 6 (*CCDC6*), and *FGFR1*-endoplasmic reticulum lipid raft-associated 2 (*ERLIN2*)); however, these fusions are not recurrent[201].

In summary, the identification of fusion genes associated with metastasis remains underexplored, leading an open field to study the relevance and the role of such events related to tumor progression and development.

1.4 Final considerations

The identification and characterization of molecular alterations associated with cancer is fundamental to understanding the biology behind tumor development. It is necessary to discover diagnostic and prognostic markers to detect and estimate cancerogenic risk. Next-generation sequencing has enabled the detection of thousands of mutations in single samples and in large cohorts, such as TCGA[202], the CGP[203], and the ICGC[204]. These projects have provided comprehensive catalogues of driver mutations in several cancer types, mainly

focused on Single Nucleotide Variants (SNVs), enabling the development of new treatments and the improvement of diagnosis and prognosis protocols for cancer care. Despite the efforts of the community in generating a molecular map of the biology of the tumor, many aspects remain underexplored, usually due to the methodological challenges that imply. For example, a few groups have studied the complexity of structural variation in cancer genomes, leaving this problem partially unsolved. Similarly, at the level of structural changes of the transcriptome, e.g., transcript fusions, previous studies have applied simplified protocols and identified a fraction of all the events present in a tumoral cell.

Current research is still missing two fundamental topics regarding the study of somatic mutations and the study of tumor progression:

1. Structural variation (SV) remains underexplored. Although there have been different efforts to characterize these types of somatic mutations, the landscape of complex SV remains uncompleted due to the lack of systematic approaches to identify and classify complex patterns of chromosomal reorganizations.

2. The events contributing to metastatic spread are unknown. While genomic alterations are accepted drivers of neoplastic transformation, the metastatic process has been attributed to dynamic changes in the transcriptome of tumor cells through post-transcriptional and epigenetic factors. Given the low availability of metastatic tumor samples, the transcriptional features linked to metastasis, including transcriptional profiling and the presence of new biomarkers, such as fusion transcripts, have been poorly explored.

In this thesis, we address these two limitations through two collaborations, and study genomic and transcriptomic structural alterations that contribute to tumor development. As the first chapter of this thesis, our participation in the PCAWG

consortium[56] allowed us to analyze about 2,586 tumor genomes (with matched normal genomes) across 40 cancer types with the aim of characterizing and classifying a particular potential pattern of structural variation that our group uncovered within a study of SVs in more than 500 patients with Chronic Lymphocytic Leukaemia (CLL)[205]. That structural pattern consisted of chromosomal rearrangements involving three mutually connected regions of the genome (Figure 9). To explain these events, we hypothesized a structural scenario, in which different chromatin regions interacted, bringing together distant DNA regions, and prompting multiple all-against-all rearrangements.

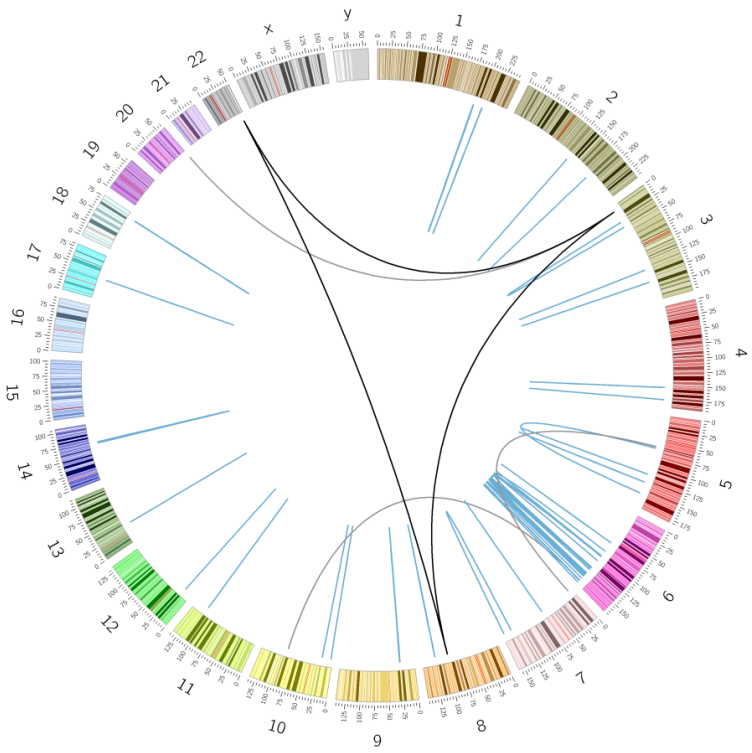


Figure 9. Pattern of chromosomal rearrangements involving three regions of the genome. Circular representation of the human genome using the Circos software[206]. The lines inside the circle represent the rearranges in one chromosome (blue lines) or between two chromosomes (grey lines). In black are the rearrangements involved in the potential pattern.

Furthermore, in the context of machine learning (ML), we collaborated with the Data-Centric Computing group from the Computer Sciences Department at BSC, developing a framework to detect complex patterns of structural variation in the PCAWG cohort.

The second part of this thesis deals with the impact of transcriptomic structural changes in lethal metastatic breast cancer, specifically regarding differential gene expression and fusion transcripts. We search for evidence that could determine the metastasis in different tissues across different RNA-seq datasets, coming from our collaborator Dr. Leticia De Mattos (IrsiCaixa, Neoantígenos y Vacunas Terapéuticas Personalizadas – NeoVaCan, Barcelona, Spain) and covering up to 82 independent metastases.

2 OBJECTIVES

The main objective of this thesis is to understand cancer development, under the light of DNA and mRNA alterations that shape cancer genomes and transcriptomes, respectively. Therefore, this thesis comprises two separated conceptual aims around the study of genomic and transcriptomic events with potential roles in cancer offset and progression, particularly towards metastasis. These objectives are:

1. To design and apply a methodology capable of classifying and isolating complex somatic rearrangements in the cancer genome, with a potential molecular mechanism behind, through the study of somatic structural variants (SVs) within the PCAWG cohort.

2. To identify transcriptomic alterations associated with metastasis in patients with lethal breast cancer from the analysis of RNA-seq data. This aim is divided into two more specific objectives: 1) Identification of irregular expression patterns associated with metastasis, and 2) identification and characterization of potential fusion transcripts.

3 METHODS

The method section has two major sections according to the two main questions developed in this thesis. The first part corresponds to the methods to isolate and classify complex rearrangements in about 2,586 tumor genomes across 40 cancer types from the PCAWG data. The second main section refers to the methodology to study the RNA alterations from 82 metastatic samples obtained from 10 different patients with metastatic breast cancer, focusing on identifying fusion transcripts involved in metastasis disease.

3.1 Classification and characterization of complex chromosomal rearrangements in cancer

This method section has two major blocks, describing the approaches we used and the complex chromosomal rearrangements we identified in 2,586 tumor genomes across 40 cancer types from the PCAWG data. The first part corresponds to the methods to identify and characterize a particular pattern of three SVs through different tumor types. The second part refers to the development of a robust methodology to study complex patterns that could include up to six SVs.

3.1.1 Identification and characterization of 3-SV pattern across different cancer types

We aimed to identify and characterize a specific SV encompassing three different genomic regions, which we here refer as 'the trisomy pattern'. We used data of structural variants encoded in Variant Call Format files (VCF files) obtained from three official PCAWG variant calling pipelines (Sanger, Broad and EMBL/DKFZ)[56]. This dataset corresponds to 2,586 samples, with matched tumor and normal whole genomes, from 40 different tumor types. The BAM (Binary Alignment) files from these samples included all reads aligned to the human reference build hs37d5.

In a VCF file, a structural variant is described as a novel adjacency of two genomic loci or *breakends*. The breakends appear when a chromosome is broken at a given locus (*breakpoint*). The adjacency refers to the SV junction that ties together two breakends (Figure 10).

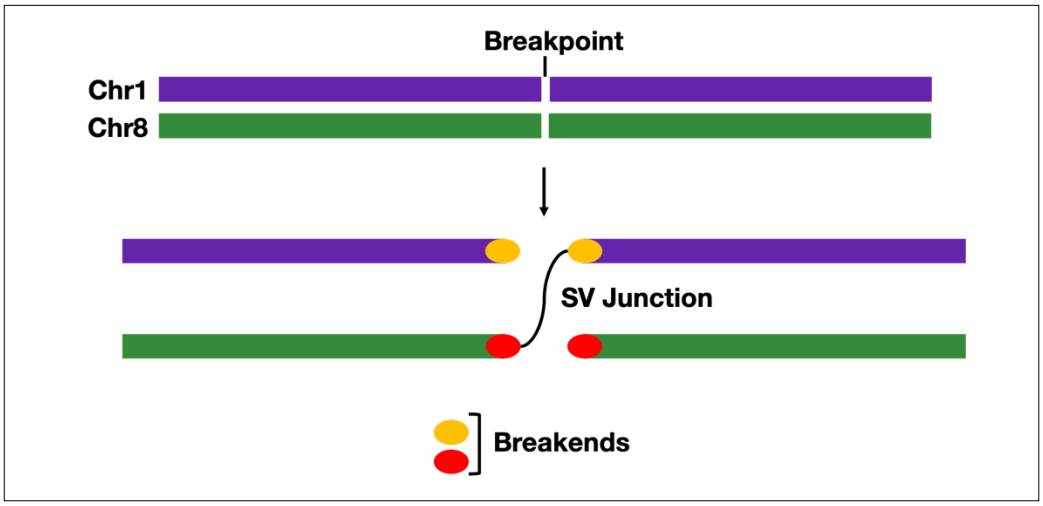


Figure 10. Schematic representation of Breakpoints, breakends and SV junction. In this example, chromosome 1 (Chr1) and chromosome 8 (Chr8) are broken at two genomic positions or breakpoints, generating two breakends per chromosome, shown in yellow and red spots, respectively. The SV indicates that the Chr1 is broken from one breakend of the SV and joined to the locus of one of the breakends from Chr8. The black line shows the SV junction.

3.1.1.1 Surveying the 3-SV pattern

With the final aim of isolating the breakpoints corresponding to interdependent events that could explain the trisomy pattern, we preliminary defined the trisomy pattern as a set of structural rearrangements that comprised three distinct DNA regions in 2 or 3 chromosomes (Figure 11). To identify the event in each cancer sample, we looked for clustered breakpoints, which were close together, involving three genomic regions (thus forming a triangle shape structure in a Circos graph).

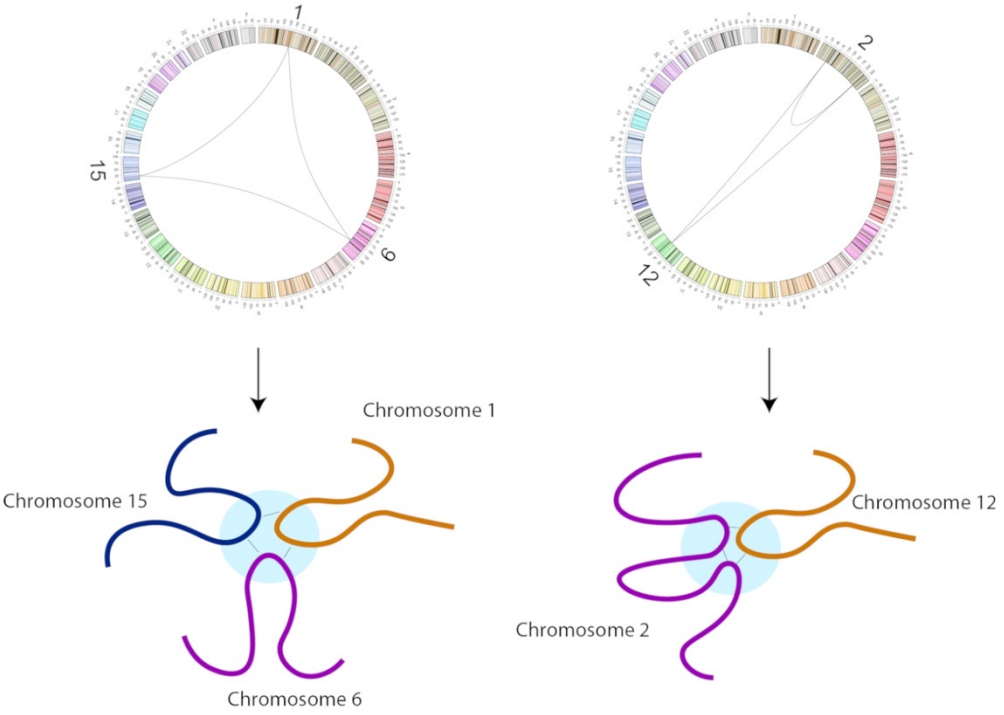


Figure 11. Two main trisomy configurations and hypothesized scenarios of chromatin conformations involving 3 genomic regions. At the top, circular representation of 3 and 2 chromosomes with rearrangements represented by lines. At the bottom, schema from hypothesized scenarios where the genomic regions interact.

First, to set the distance (in base pairs) between the breakpoints of the three SVs, we evaluated two specific features: the *window* and *loop sizes*. The distance between two adjacent breakpoints in a chromosome was called window size. If the breakpoint was intrachromosomal, the distance between the two breakends of the SV was named loop size (Figure 12).

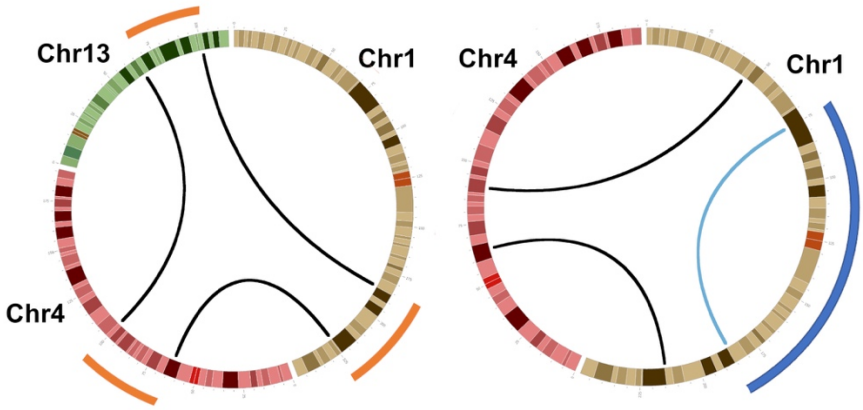


Figure 12. Window and loop sizes representation. Two Circos plots of 3 or 2 chromosomes involved in a trisomy pattern composed of 3 SVs, represented by the inner lines. The distance between two adjacent breakpoints called Window size, and the distance between the two breakends of the intrachromosomal SV named Loop Size, both represented by the orange and blue lines around the Circos plots, respectively.

Then, we developed an in-house algorithm to search for three SV patterns involving at least two chromosomes. Different values of window (500b, 1kb, 2kb, 5kb, 10kb, 15kb, 20kb, 30kb, 50kb), and loop (15kb, 50kb, 100kb, 200kb) sizes were used to test how many instances of the pattern could be found. The values of minimum window and loop sizes that included the most significant number of cases of the trisomy pattern were selected. We calculated consensus events (in every sample) by clustering the adjacent breakpoints (2Kbp) that belonged to the same instance to count the trisomy patterns per sample.

3.1.1.2 Assessment of the trisomy pattern

Once the trisomy pattern was defined in terms of the number of chromosomes, breakpoints, window and loop sizes, we considered whether the pattern might arise by chance. To test this, the breakpoints of the samples with three or more breakpoints were all pooled together to create simulated datasets 1,000 times. The number of samples with at least one trisomy pattern (positive samples) was

calculated and compared with the observed number of positives from the real dataset in each permutation. Then, after 1,000 permutations were done, a *p*-value was calculated.

Furthermore, we determined whether the number of breakpoints per sample was correlated (Pearson correlation) with the number of patterns found in the sample. We also evaluated whether the distribution of the breakpoints involved in trisomy events through the genome was uniform, by taking all of the breakpoints of trisomy patterns from all samples and running a One-Sample Kolmogorov-Smirnov test for every chromosome. We also compared the distribution of the breakpoints involved in the patterns in a given chromosome and the distribution of all breakpoints from the samples using a Two-Sample Kolmogorov-Smirnov test.

3.1.1.3 Functional characterization of the trisomy pattern

To evaluate the potential role or impact of this pattern in cancer, the encompassed genomic regions were examined using different approaches of functional inference. First, to explore the effect of the pattern on any biological process, we performed a gene enrichment analysis of the genes affected by the SVs involved in the patterns using the Enrichment Analysis tool[207–209]. To determine which genes were affected by the breakpoints of the trisomy pattern, we annotated the regions where the breakends were at the exact loci and also at 2Kbp up and downstream of the locus of a given breakend.

We performed a further gene enrichment analysis using a list of differentially expressed genes in all samples with trisomy patterns. For this, we analyzed the available raw counts from PCAWG RNA-seq dataset[210] using Limma-Voom R/Bioconductor software packages[211] to overcome the problem of having different cancer types. Then, we used DESeq[212] to compare the gene

expression profiles of samples with the pattern against samples without. The results were filtered by the log-fold-change parameter (LFC) set to 1.

3.1.1.4 Genomic Characterization of the pattern

To study the mechanism by which the trisomy pattern likely arose, we used the breakpoints reported in the VCF files, the mapped reads from the BAM files, and the copy number of the involved regions, performing a manual reconstruction of the derivative chromosomes from several examples.

We hypothesized that the pattern arose from a mutational process occurring during cell replication. We expected the three involved regions to have replication times more similar between them than three random regions as a first approach. As each region (named A, B, C) was defined by two adjacent breakpoints, the replication time was assigned to the average position between these two loci. Given that there were different cancer types from several tissues, we used the Wavelet-smoothed signal value averaged across three cell lines to designate the replication times: NHEK (normal skin, ectoderm), GM12878 (normal blood, mesoderm), and IMR90 (normal lung, endoderm) (Repli-Seq data from ENCODE / University of Washington). We calculated the absolute difference of replication times for each event as: $d = |A-B|+|A-C|+|B-C|$. Thus, higher *d* values represented higher differences in replication times between the three genomic regions. Then, we performed a random test of 10,000 permutations taking three replication times randomly from the observed replication time values, to conserve each permutation data distribution. Finally, to assess the significance of the obtained empirical *d* values, we compared them with the median *d* values from these permutations.

3.1.2 A strategy to isolate chromosomal rearrangements that derive from specific molecular mechanisms

In the context of machine learning (ML), we collaborated with the Data-Centric Computing group from the Computer Sciences Department at BSC, to design a strategy to detect complex patterns of structural variation in the PCAWG cohort. Through this collaboration, we studied different methods to identify rearrangements arising under specific molecular mechanisms, which were likely to have a functional implication in tumor progression.

As a result of this collaboration, we developed a framework on the basis of statistical inference, which allowed us to isolate events that deviate from random expectation, and thus could derive from a single catastrophic event.

3.1.2.1 Classifying the SVs that belong to complex chromosomal rearrangements

We aimed at identifying the SVs that were likely derived from the same molecular mechanism, as they share some topological properties. For this, a first step of identifying clustered breakpoints that belong to the same single rearrangement event is crucial. We proposed the use of Kernel Density Estimation (KDE), a non-parametric statistical method, to estimate the probability density function of the breakpoints in each chromosome. Thus, we could rely both on the closeness of the breakpoints and their density within the entire cohort. The clustering of the breakpoints depended on a hyperparameter called Bandwidth: high values lead to bigger clusters, whereas low values generate smaller and sparser clusters (Figure 13). Then, we selected the bandwidth value in which the distance between the clusters was the highest (inter-cluster distance), and the distance between two breakpoints within the same cluster (intra-cluster distance) was the lowest. Therefore, to optimize these two distances, we performed a recursive 2-step clustering inside every

cluster, making breakpoint clusters to be, as separate from each other as possible and, as small as possible.

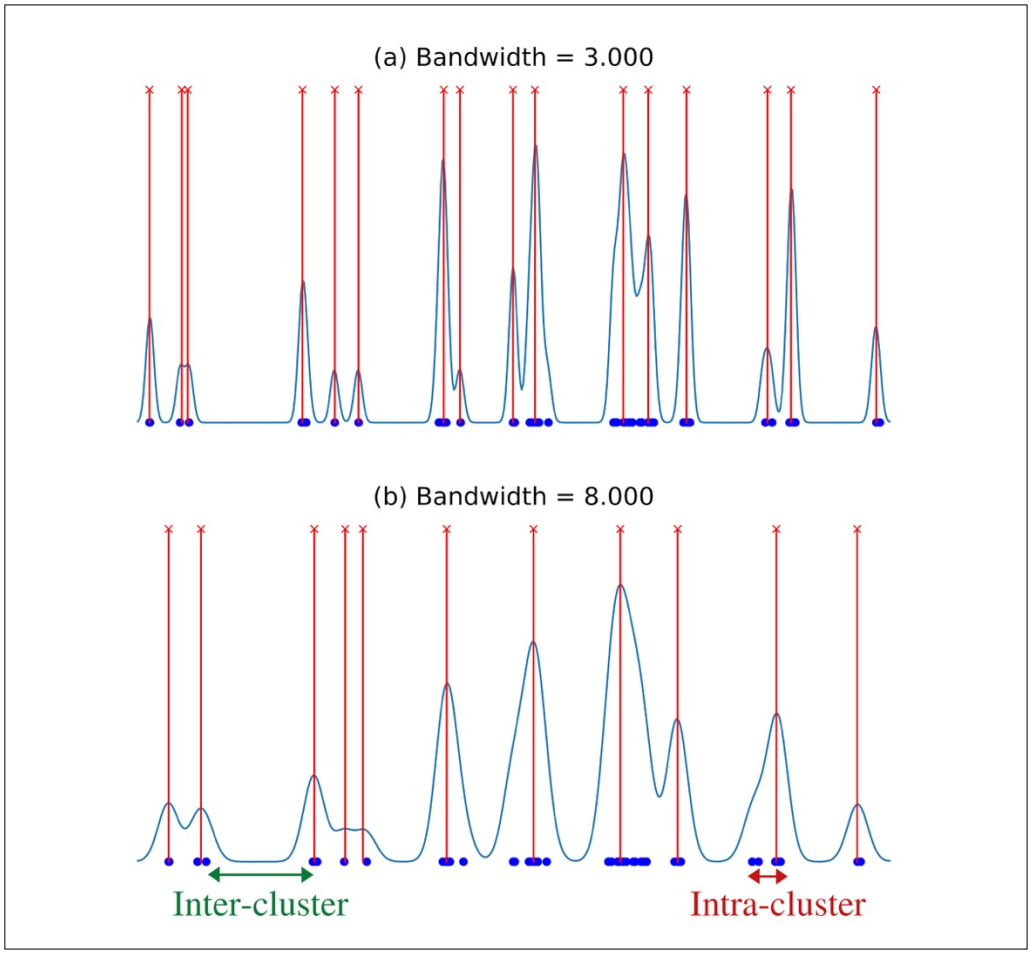


Figure 13. Representation of the intra and inter-cluster distances. Kernel Density Estimation of breakpoint clusters from chromosome 3 setting bandwidth values of (a) 3000 and (b) 8000. Blue dots represent the locations of the breakpoints, the blue line is the kernel density estimation, and the red lines are the obtained cluster peaks. The inter and intra-cluster distances are shown in green and red, respectively.

Once we had the clusters of breakpoints, we validated that our method was not considering random clustered SVs by performing two tests. In these two tests, we used 100 simulated datasets generated as previously described in section 3.1.1.2. In the first test, we estimated the average dispersion of breakpoints in

each simulated dataset. We used as a dispersion measure the standard deviation of the difference of base pair between adjacent breakpoints in a chromosome. Next, we performed a one-sample Z-Test to compare the average dispersion distribution from the simulated datasets against the average dispersion of breakpoints in the original dataset. In the second test, we applied the KDE clustering method to each simulated dataset. We compared (assessing a one sample Z-Test) the average number of breakpoints per cluster and the average number of breakpoints per cluster in the original dataset for each permutation.

3.1.2.2 Searching patterns of SVs by graph mining

In the previous clustering step, every sample was set out as a graph in which the breakpoint clusters were represented as vertices, and the edges connecting these vertices corresponded to SV junctions. Given that the vertices could encompass several breakpoints from different SVs, different graphs could be generated. To narrow down the survey of graphs, we focused only on Hamiltonian cycles (mentioned further only as “cycles”), where every vertex is connected to two other vertices (Figure 14). We used a search approach method based on the VSIGRAM method to find and count the events of chromosomal rearrangements. Moreover, since the subgraph mining problem became computationally hard (NP-hard), we performed a pruned search with max size = 6, searching for patterns up to 6 SVs.

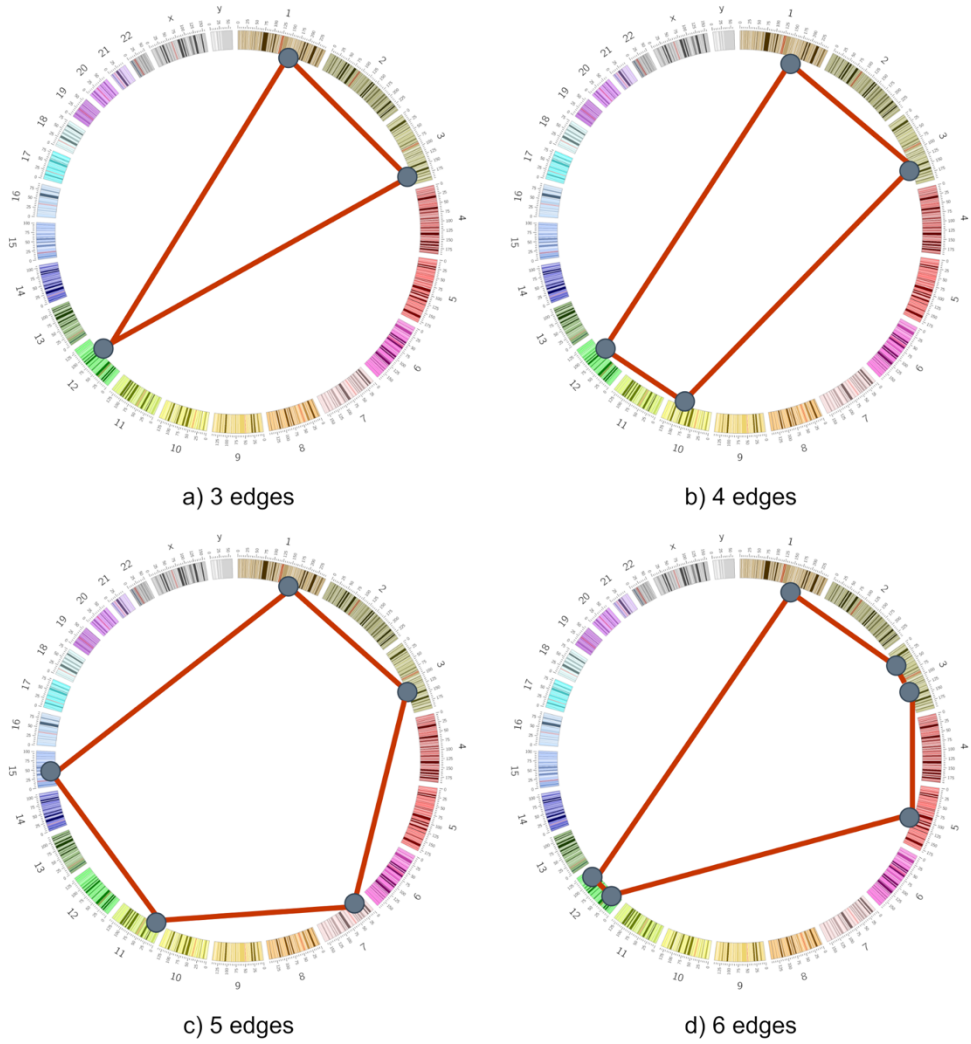


Figure 14. Circular representation of the human genome with cycles of different sizes

The method for subgraph mining visited the graph through depth-first search, allowing parallelism, e.g., by splitting each starting vertex to be processed simultaneously. At every vertex, we looked for all the possible connected paths of size 1. Then, these subgraphs were the candidates for looking for all the possible connected paths of size 2. The process was repeated for the paths of sizes 3, 4, 5, and 6.

3.1.2.3 Isolating statistically significant patterns

Here, we proposed the use of Abundance measure[213] to discern statistically significant SV patterns from random distributions. This metric compared frequencies between the patterns found in the graph mining step and random observations from simulated datasets. Then, for a given cycle, we calculated the abundance as follows:

$$\Delta = \frac{f_{input} - f^{-}_{random}}{f_{input} + f^{-}_{random} + \varepsilon}$$

f_{input} was defined as the frequency of a cycle in the original dataset and, f^{-}_{random}, the mean of the frequencies of a cycle in N simulated random datasets. ε is a pseudo-count (Laplace smoothing) to prevent the ratio from exploding when frequencies are small. Δ could take values between -1, under-represented and +1, overrepresented, being 0 the value for a pattern with the same representation in the original data than in the random datasets.

The simulated datasets were performed by randomizing the SV junctions (edges) between the breakpoint clusters. This method consisted of repeatedly selecting two random edges A-B and C-D, and exchanging the ends to form two new edges, e.g., A-D and B-C. Therefore, we generated 100 simulated datasets as follows: we removed the original edges of every sample and randomly assigned the same number of edges to each sample every time. Then, the resulting graph keeps the same vertices and edges count.

3.1.2.4 Reconstructing genomic configurations of the SV patterns

As previously described in section 3.1.1.4, we used the orientation of chromosomal segments at the breakpoints and their associated copy-number alterations to categorize the most overrepresented and recurrent complex chromosomal rearrangements found across the PCAWG cohort.

3.2 Identification and characterization of mRNA alterations in metastatic breast cancer

As mentioned in section 2, the main goal of this second part of the thesis is to identify characteristic features or patterns within the transcriptome of breast cancer cells that can be related with metastasis. Originally, we plan to cover this goal through two different fronts: 1) the analysis of the transcriptome of metastatic cells with the aim of identifying specific expression patterns associated to this process; and 2) the analysis of fusion transcripts, also associated to the metastatic process. To address these questions, we analyzed RNA-Seq data from 82 independent metastases from 10 patients with breast cancer, coming from our collaborator Dr. De Mattos (IrsiCaixa).

3.2.1 RNA Sequencing datasets from metastatic breast cancer

In this study, we analyzed metastatic breast cancer samples from two datasets that differed in terms of sequencing approach and number of patients. The first dataset consisted of 64 different metastatic samples from 9 cases, assayed using Single-End (SE) RNA sequencing libraries with a read length of 50bp (Figure 15, SE dataset). The second dataset included 18 metastatic samples from five different tissues of a single patient (ID code: 302), sequenced using paired-end (PE), strand-specific RNA libraries with a read length of 150bp (Figure 15, PE dataset).

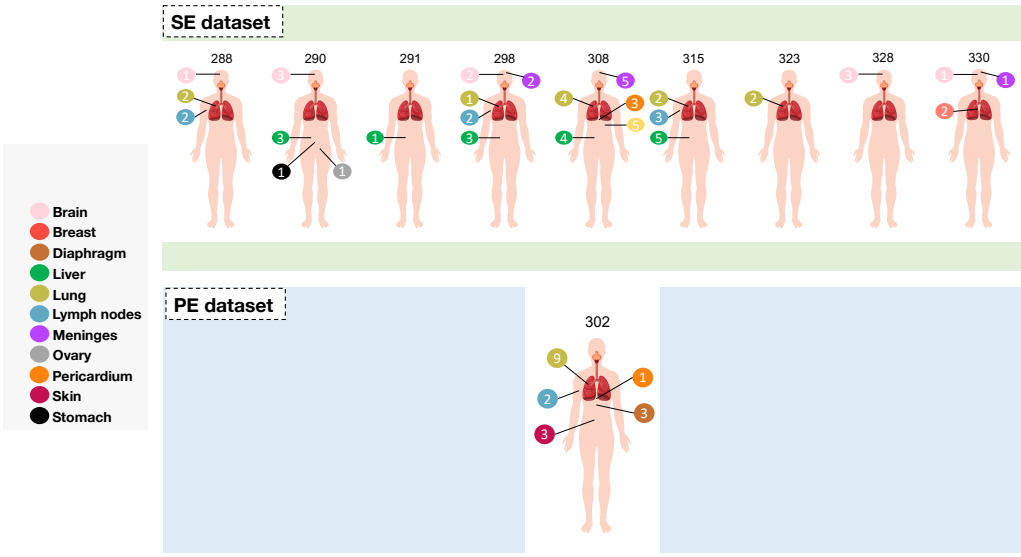


Figure 15. RNA-seq samples from Lethal Metastatic Breast Cancer. Representation of patients with number and type of samples according to tissue. There are 9 patients from SE dataset in the green box, and in the blue box, the case 302 with PE dataset.

As shown in Table 1, the PE dataset from patient 302 contained RNA-seq of multiple samples from the same metastasis, identified with IDs with the same number (e.g., 302_005_A, 302_005_B, 302_005_C were from the same metastasis in the diaphragm). There were also samples from distinct metastasis in the same tissue (e.g., 302_015, 302_016, 302_018 were from different metastasis in the chest wall skin).

Table 1. Tumor sampling from patient 302 (PE dataset).

Sample ID	Tissue	Location
302_005_A	Diaphragm	Left diaphragm
302_005_B	Diaphragm	Left diaphragm
302_005_C	Diaphragm	Left diaphragm
302_006_A	Lung	Medium right lung
302_006_B	Lung	Medium right lung
302_007	Lung	Up right lung
302_008_A	Lung	Down right lung
302_008_B	Lung	Down right lung
302_009	Lung	Up left lung
302_010_C	Lung	Down left lung
302_010_D	Lung	Down left lung
302_010_E	Lung	Down left lung
302_011_A	Lymph Nodes	Lymph Nodes
302_011_B	Lymph Nodes	Lymph Nodes
302_012_A	Pericardium	Pericardium
302_015	Skin	Chest wall skin
302_016	Skin	Chest wall skin
302_018	Skin	Chest wall skin

3.2.2 Gene expression analysis

The reference genome used to align the RNA-seq reads from the metastatic samples was based on the Broad Institute's GRCh38/hg38 reference genome for gene expression analysis (ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg38/). From this genome reference, the ALT, HLA, and Decoy contigs were excluded.

Once we had the RNA-seq reads aligned to the reference genome (using STAR[214] v.2.5.3a), the GENCODE 30 annotation (https://www.gencodegenes.org/human/release_30.html) was used to

annotate transcripts and get the gene-level and transcript-level expression quantifications in terms of Transcripts Per Kilobase Million (TPM), using RSEM[215] v.1.3.0.

To determine how similar the samples across patients and metastases were, we performed Multidimensional Scaling Analyses (MDS)[216] for each dataset (SE and PE), based on the TPMs.

3.2.3 Detection of fusion transcript candidates

The raw sequencing reads from each metastatic sample were mapped to the GRCh37 assembly of the human genome (hg19) using STAR v2.5.2b, specifying the corresponding length of reads (150bp for PE reads and 50bp for SE reads).

The derived BAM files were the inputs in both fusion transcript detection approaches described in the following 3.2.3.1 and 3.2.3.2 sections.

3.2.3.1 Design of a manually curated pipeline to identify fusion transcript candidates

We developed a pipeline to detect fusion transcripts (Figure 16) which was employed as follows. First, we identified RNA-seq reads that could imply fusion transcript events by overlapping a chimeric junction. These reads, defined as split reads, were characterized by having two portions mapping in two distinct locations in the genome.

Then, we used the stand-alone BLAST[217] v2.6.0 tool to compare these split reads with a human transcriptomic database built from coding DNA (cDNA) and ncRNA, and taken from the EMBL databases (release 93). Here, we selected the reads that aligned on one side to one transcript and on the other side to another transcript from a different gene (Figure 17.I.), generating a list of

candidate fusion transcripts as output. In the selection, we set four criteria: i) Reads could not have better hits apart from the two transcript partners; ii) The percentage of identity of a read with both reference transcripts from the database was 100; iii) An overlapping of 7bp maximum was allowed between the two aligned portions of the read to the two partner transcripts; iv) The minimum alignment length of a read to each of the two transcript partners was set to 25bp.

Next, we determined the most likely fusion transcripts filtering out unlikely candidates from the preliminary list as follows. In the first filtering step, we excluded candidates that were reported as pseudogenes. Similarly, fusion candidates A—B, where A and B were transcripts from overlapping annotated genes, were excluded as well. After, we filtered out fusion candidates A—B when A and B were sequence-similar fusion pairs. The sequence similarity between any two reference isoforms of genes was calculated from an all-vs-all BLAST search. If reference transcript A and reference transcript B had a significant sequence alignment (E-value <= 10^{-3}), the fusion candidates A—B were excluded from the preliminary list of candidates. Finally, the fusion candidates that were composed of transcript partners found to have more than three fusion partners were also filtered out. We named these transcript partners fused with multiple partners as *promiscuous transcripts*.

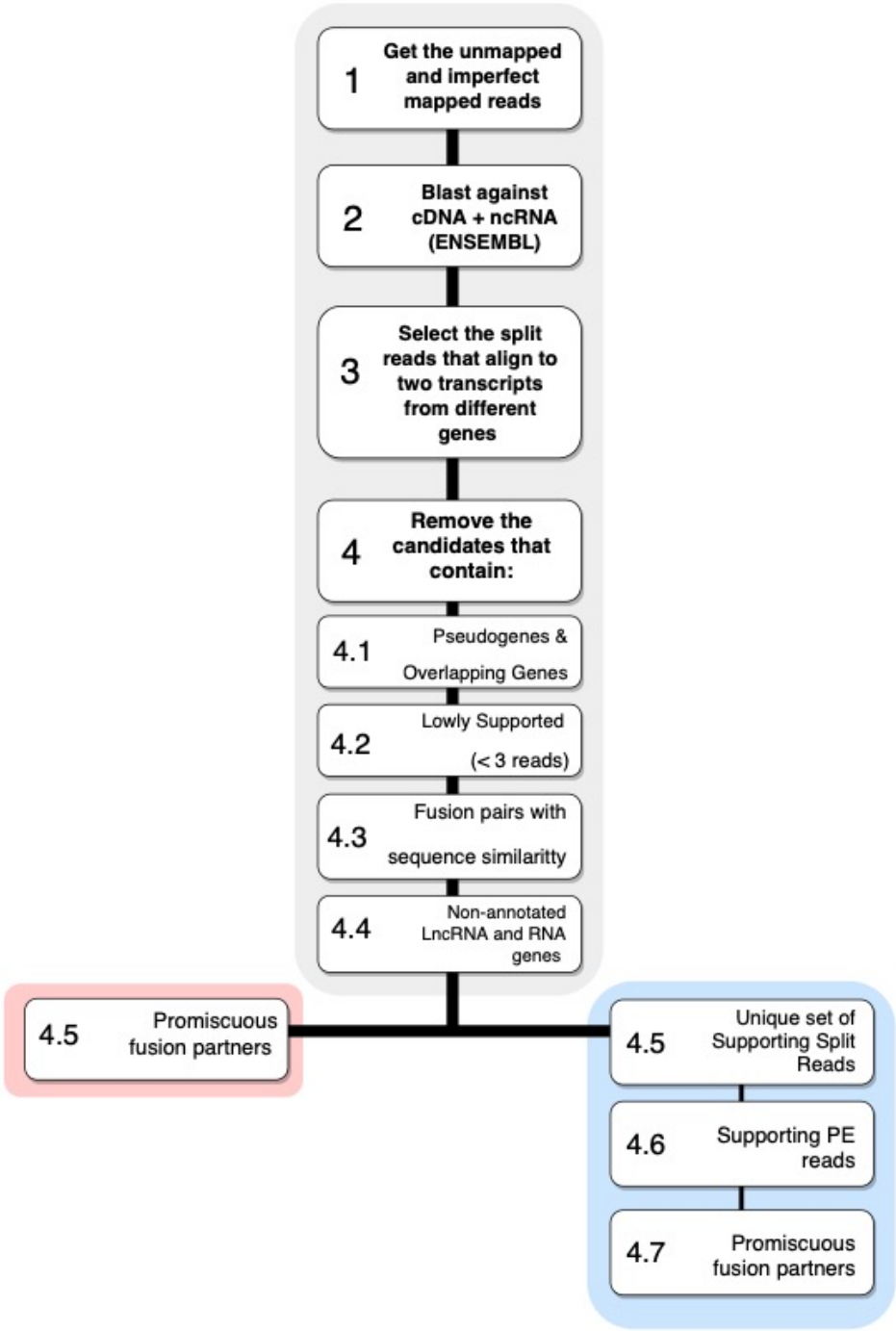


Figure 16. Pipeline to predict and identify fusion transcripts using RNA-seq data. First, we identified the informative reads from the RNA-seq alignment files that could inform fusion transcript events. Second, we performed a BLAST of these reads against

the human transcriptome database. Then, we selected the split reads which aligned to two transcripts from distinct genes. Finally, to filter out the likely false positives, different filters were applied. The last filters were different according to the dataset type: those applied to SE dataset are displayed in the red box, while filters for PE dataset are shown in blue.

To further prioritize the top candidates analyzing the PE dataset, we added a stringent parameter. We only selected the candidates supported by reads with unique mapping in fusion transcripts. Therefore, each read could support only one fusion transcript.

Figure 17. Schematic representation of different scenarios in which supporting reads could indicate the presence of a fusion transcript. I. The supported reads were split reads (in bold) and **II.** The supported reads were split reads (in bold and their matching pair reads supporting any region of the fused transcripts, in grey), and paired-end reads.

For the PE dataset, we could additionally use the paired-end reads information for the detection of fusion transcript candidates. We thus searched for PE reads aligning to the opposite sides of a chimeric junction, and thus supporting the candidates previously identified by split reads (Figure 17.II.).

3.2.3.2 Using available bioinformatics tools for fusion transcript detection

As a second approach, STAR-Fusion[218] v1.4.0 (with default parameters) was used to identify fusion transcripts in the two datasets. For samples from SE data, the parameter *--require_LDAS* was set to 0 in order to allow the detection of fusion transcripts only through split reads.

Since STAR-Fusion requires several genomic resources already built on GRCh37/hg19, we mapped again the reads from our samples to this reference genome (3.2.3.1 section).

3.2.3.3 In-silico validation of multi-fusion partners

In order to validate the fusions with transcripts found to be fused to several partners (so-called promiscuous transcripts), we reconstructed *in-silico* three transcript fusion candidates predicted in the 302_010_D sample (from lung metastasis). The promiscuous partner was an isoform of the *CD24* gene (NM_013230.3), and the fusion candidates were (i) *CD24—HSP90AB1*, (ii) *CD24—MALAT1* and (iii) *CD24—PEG10.*

The *in-silico* reconstruction was performed as follows. First, we identified the split reads and the paired-end reads that supported the fusions to determine the fusion point and the coordinates of the fused transcripts, respectively. Then, to assemble the sequences of the fusions, all the PE reads supporting each of the transcript partners were aligned with the corresponding Refseq transcript isoform (using the multiple sequence alignment Clustal Omega[219]).

3.2.3.4 Experimental validation of promiscuous partners in fusion transcripts

For experimental validation, the reconstructed fusion transcript candidates were amplified by PCR and sequenced via Sanger sequencing. First, 200 ng of total RNA from sample 302_010_D was retro-transcribed using the iScript cDNA Synthesis Kit (Bio-Rad) according to the manufacturer's protocol.

Fusion transcript primer pairs were designed to confirm three pairs of transcript fusion candidates and their different detected fusion forms (same partner pair but different fusion points) with a total of 13 predicted transcript fusion candidates (Supplementary Table 1). KAPA Taq DNA PCR Kit (Roche) was used for PCR reactions. Each PCR was performed in 25 µL containing 2 µL cDNA, 12.5 µL KAPA Taq mix, 8.5 µL Water, and 1 µL (10 µM) each primer. The PCR thermal conditions were as follows: a pre-denaturation of 95°C for 3 min, 42 cycles of amplification (95°C for 30 s, 55°C for 30 s and 72°C for 40 s), and a final extension reaction was performed at 72°C for 5 min. PCR products were detected on a Labchip 2100 Bioanalyzer (Agilent Technologies). 20 ng of PCR products were purified (CleanPCR, CleanNA) and sequenced with a 3500 Genetic Analyzer (Life Technologies).

Given the fact that we could not obtain the full length sequence corresponding to the best candidate validated, the fusion transcript with ID HSP90AB1_III (5′ *CD24*−*HSP90AB1* 3′), we decided to confirm the junction between *CD24* and *HSP90AB1,* by designing a new set of primers (Nested_HSP90AB1_III_FW 5′-GCTTGAGAAATATGGACACTTAATACT-3′; Nested_HSP90AB1_III_RW 5′-ACATGAGTTGGGCAATTTCT-3′) to amplify the junction point (Supplementary Table 1, Nested primers,) and thus, sequenced. The PCR was performed using the above-described reaction mix, and the protocol as follows: a pre-denaturation of 95°C for 3 min, 30 cycles of amplification (95°C for 30 s, 55°C

for 30 s and 72°C for 40 s) and a final extension reaction was performed at 72°C for 5 min.

4 RESULTS

4.1 Classification and characterization of complex chromosomal rearrangements in cancer

With the final aim of studying potential chromosomal interactions across distant regions that could have functional consequences for the biology of the tumor, we initially assessed the features that defined a particular pattern of three SVs, developing methods for identifying complex patterns that likely arise in a single-hit event. To that end, we used the largest and most comprehensive cancer genomic dataset so far, from the PCAWG Consortium of ICGC and TCGA. Finally, we applied all our findings to propose a framework to classify complex rearrangements, with high levels of complexity regarding the number of SV involved.

4.1.1 Identification and characterization of the trisomy pattern

Despite multiple efforts towards the classification of structural variants in cancer, there is currently no consensus or comprehensive approaches to classify complex genomic rearrangements. Given the previous evidence found by our group, we set out to survey a specific complex rearrangement through different cancer types, which encompassed three genomic regions, which we call here the trisomy pattern. We hypothesize that there must be a molecular mechanism by which a likely conformation of the chromatin brings together those three regions.

4.1.1.1 Defining the features of trisomy pattern across cancer types

First, we empirically determined the features that characterized the pattern across the different cancer types in the PCAWG dataset. Given the SVs found in each tumoral sample, we looked for 3 SVs close together that might result from interdependent alterations. We measured that closeness between the clustered SVs as the distance in base pairs between their breakpoints. Our search focused on three non-locally clustered breakpoints in three or two different chromosomes, to distinguish the trisomy pattern from previously described patterns such as chromothripsis. We expected that if there were a complex pattern of SVs, all the events that belonged to that pattern would share some characteristics that reflected the mechanism by which they might arise. In this regard, we explored those characteristics in terms of the distance between the breakpoints that composed each event, and selected the most common values across all cancer types. As shown in Figure 18, the vast majority of the instances of the trisomy pattern across the entire cohort, had SVs at 2Kbp of distance (Window size) and 200Kbp to 14Mbp in the case of intrachromosomal SVs (Loop size).

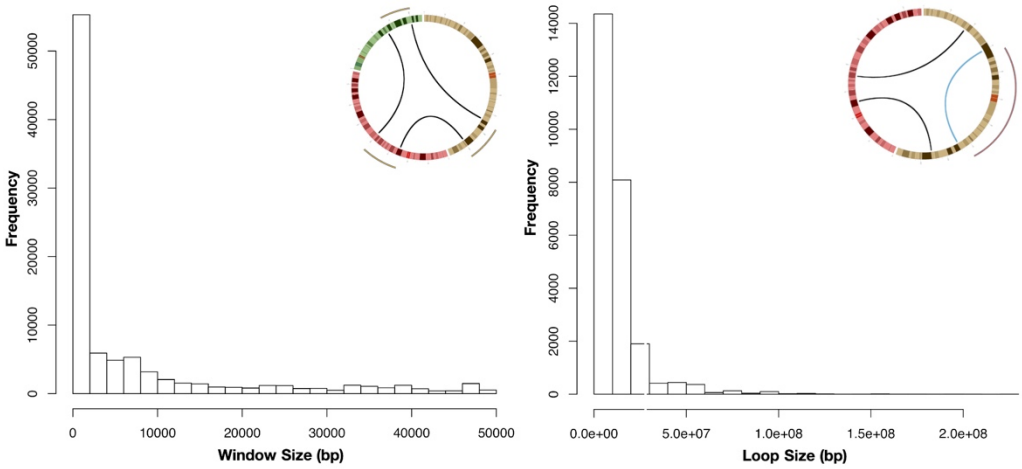


Figure 18. Window and loop sizes distribution in the trisomy patterns found across the PCAWG cohort. The Histograms shows the frequencies of the distances, both window and loop sizes illustrated in the Circos plots.

We developed an algorithm to search for trisomy patterns on the basis of (1) the number of SVs (three interchromosomal rearrangements or two inter- and one intrachromosomal), (2) the above-defined window and loop parameters, and (3) the number of the involved chromosomes (three or two). The trisomy pattern was identified in 537 out of 2,586 samples, approximately 21% of the whole dataset. On the one hand, the distribution of the pattern across cancer types was quite heterogeneous, but recurrent across the different cancers (Figure 19): it was found in 36 out of 40 cancer types. Precisely, Soft Tissue Liposarcoma (SARC) followed by Uterine Corpus Endometrial Carcinoma (UCEC), Ovarian cancer (OV), Lung Squamous Cell Carcinoma (LUSC), and Bone cancer (BOCA) were the types with more samples carrying the pattern. At least around 50% of the samples in those tumor types presented the trisomy pattern.

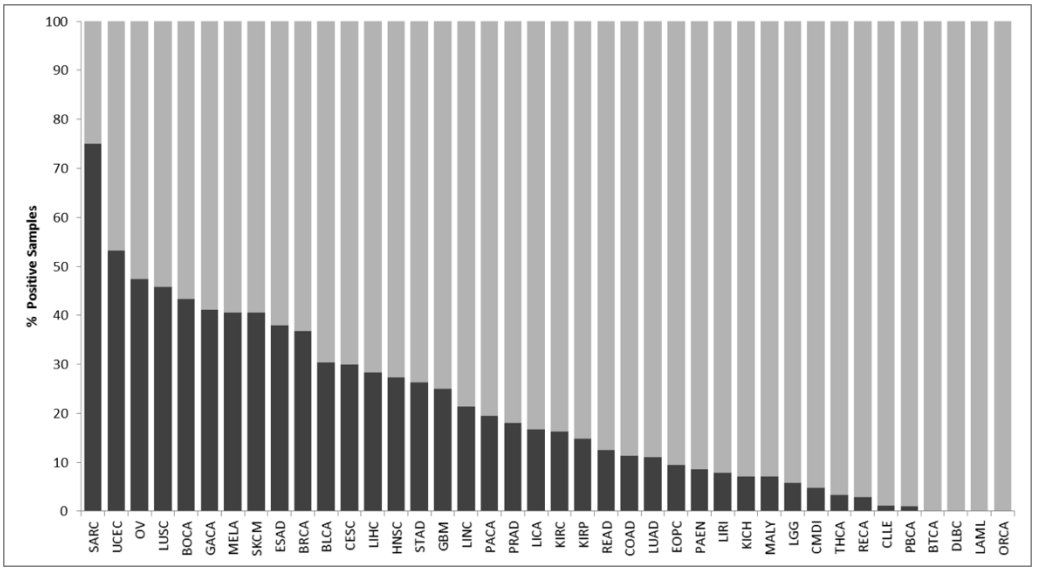


Figure 19. Relative frequencies of samples carrying the trisomy pattern across the entire PCAWG cohort.

Conversely, Biliary Tract Cancer (BTCA), Lymphoid Neoplasm Diffuse Large B-cell Lymphoma (DLBC), Acute Myeloid Leukemia (LAML) and Oral cancer (ORCA) did not have samples with such events. The rest of the cancers varied in their proportion of samples with trisomy patterns between 20 to 40%.

On the other hand, the number of instances of the trisomy pattern per sample varied, ranging from 1 to over 40, but this number was not related to the cancer types with the highest frequency of samples with the pattern (Figure 20).

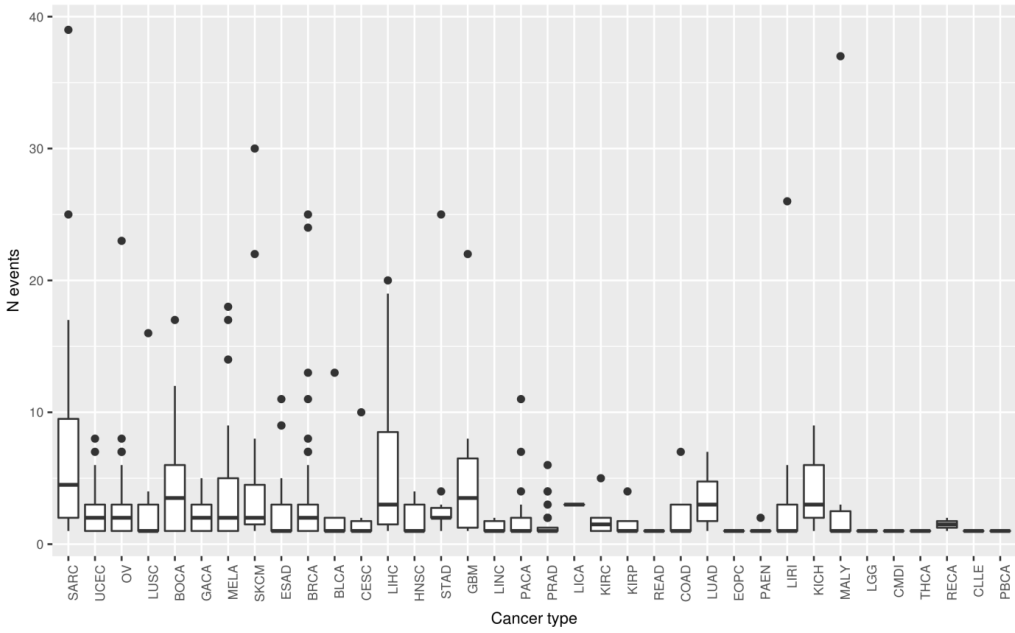


Figure 20. Boxplot diagram showing the number of trisomy patterns per sample across cancer types.

Then, not in all cases where the cancer type with a high number of samples carrying the trisomy pattern, had high numbers of instances of the pattern per sample. For example, although the trisomy pattern was detected in only about 10% of the Malignant Lymphoma (MALY) samples, one of those samples had the second-highest number of events (36 instances of the trisomy pattern) in the whole PCAWG dataset. Furthermore, in samples from a given cancer, the number of events per cancer also differed. In SARC, we found one sample with 39 events even though the median was 4. These results suggested that the trisomy pattern was heterogeneous and did not show the same behavior within samples from a given cancer or across all cancer types.

4.1.1.2 The trisomy pattern derives likely from a single-hit event

Once we set out the features that defined the trisomy pattern, we next asked whether the involved rearrangements arose independently of one another, by performing a permutation test. For this, the breakpoints of the samples were all pooled together to create simulated datasets 1,000 times. Comparing the observed number of samples containing the trisomy pattern with the numbers of samples with the pattern in the simulated samples, a *p*-value less than 0.001 was obtained. This result indicated that the trisomy pattern occurred significantly more often than expected by chance. Therefore, there should be a mechanism behind the generation of these chromosomal rearrangements by which the tumoral cell acquired the three chromosomal rearrangements in a coordinated and simultaneous fashion.

Additionally, there was no correlation between the number of breakpoints per sample and its number of events (Figure 21, Pearson Correlation coefficient r: 0.40). This further supported the hypothesis of an existence of a coordinated process behind this type of chromosomal rearrangements, rather than independent SVs randomly occurring in adjacent genomic loci.

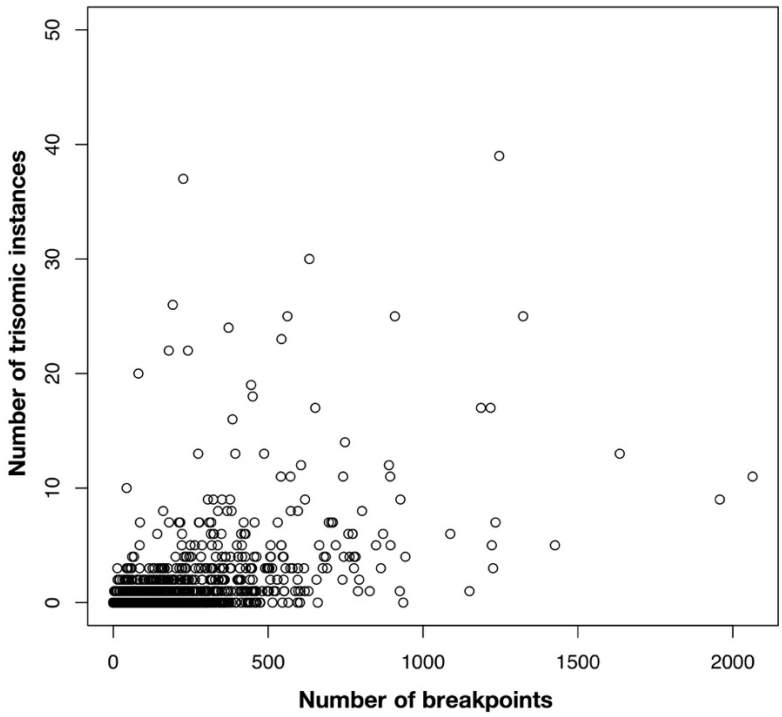


Figure 21. Correlation between the number of breakpoints and number of trisomic events in each sample. We calculated the Pearson Correlation coefficient to evaluate whether the co-occurrence of adjacent breakpoints is random due to genomic instability. r: 0.40.

Furthermore, using the Kolmogorov-Smirnov test, we evaluated whether the distribution of the breakpoints involved in the trisomy patterns in a given chromosome differed from the distribution of all breakpoints in all the 537 tumor samples. We found that they had different distributions across the chromosomes, with *p*-values < 0.01, except for chr3 (*p*-value 0.059) and chrY (*p*-value 0.093). In fact, the breakpoints did not have even distributions in the genome (*p*-value/chromosome < 2.2×10^{-16}), indicating that they were not randomly distributed through the entire genome.

4.1.1.3 Neuronal pathways are being affected by the trisomy pattern

Given that the trisomy pattern was not composed of stochastic SVs, we analyzed the functional relationship between the genomic regions implicated, in order to infer the origin of the trisomy pattern. Then, to assess the role of the trisomy pattern in the tumoral cells, we first analyzed the genomic regions affected by the breakpoints of the pattern. We expected these trisomy events could be derived from a specific biological mechanism that would bias them towards some particular regions of the genome, even if the patterns were not the same across samples.

We performed a gene enrichment analysis of the genes disrupted by the SVs (at breakpoint loci). Surprisingly, the genomic regions affected by the rearrangements were enriched in genes of biological processes related to the regulation of synapsis, such as glutamate receptor signaling pathway (Fold Enrichment - FE 5.06 and *p*-value 2.54×10^{-4}) and regulation of dendritic spine morphogenesis and development (FE 4.74 and *p*-value 5.07×10^{-3}). Moreover, when the analyzed breakend regions were expanded to 2Kbp up and downstream, the biological processes of glutamate receptor signaling pathway (FE 4.64 and *p*-value 6.72×10^{-4}), regulation of postsynapse organization (FE 3.79 and *p*-value 3.95×10^{-3}) and dendrite development (FE 3.30 and *p*-value 5.43×10^{-4}) were also overrepresented.

We further performed a differential gene expression analysis between the samples with a trisomy pattern and samples without. Similarly, we found an enrichment of the dopaminergic neuron differentiation process.

Considering the types of cancer with most trisomy patterns, we did not expect such as neuronal related terms in the enrichment analyses. Our work puts

forward the hypothesis that glutamate-related processes might play an important role in catastrophic carcinogenic events.

4.1.1.4 Different genomic conformations derive from the trisomy pattern

Due to the challenges of studying and understanding the underlying basis of complex chromosomal rearrangements in cancer, previous analyses have described different patterns of SVs based on their clustering, orientation and associated copy-number changes[91,92,220–222]. Having recovered such information, the hypothesis that such chromosomal rearrangements arose from sequential occurrences of independent events in the cell has been tested in those studies under different approaches. In this regard, some models were built to find the most plausible molecular mechanism that explains how the clustered SVs emerge.

In order to understand the mechanism pushing forward the trisomy pattern, we decided to use the orientation and associated copy-number changes to perform manual reconstructions of the derivative chromosomes. In terms of recurrence and frequency, the above-described heterogeneity of the trisomy pattern suggests a possible scenario in which more than one mechanism could trigger the pattern. Thus, we expected that the genomic conformations derived from the pattern would shed light on the molecular mechanisms giving rise to this trisomy pattern.

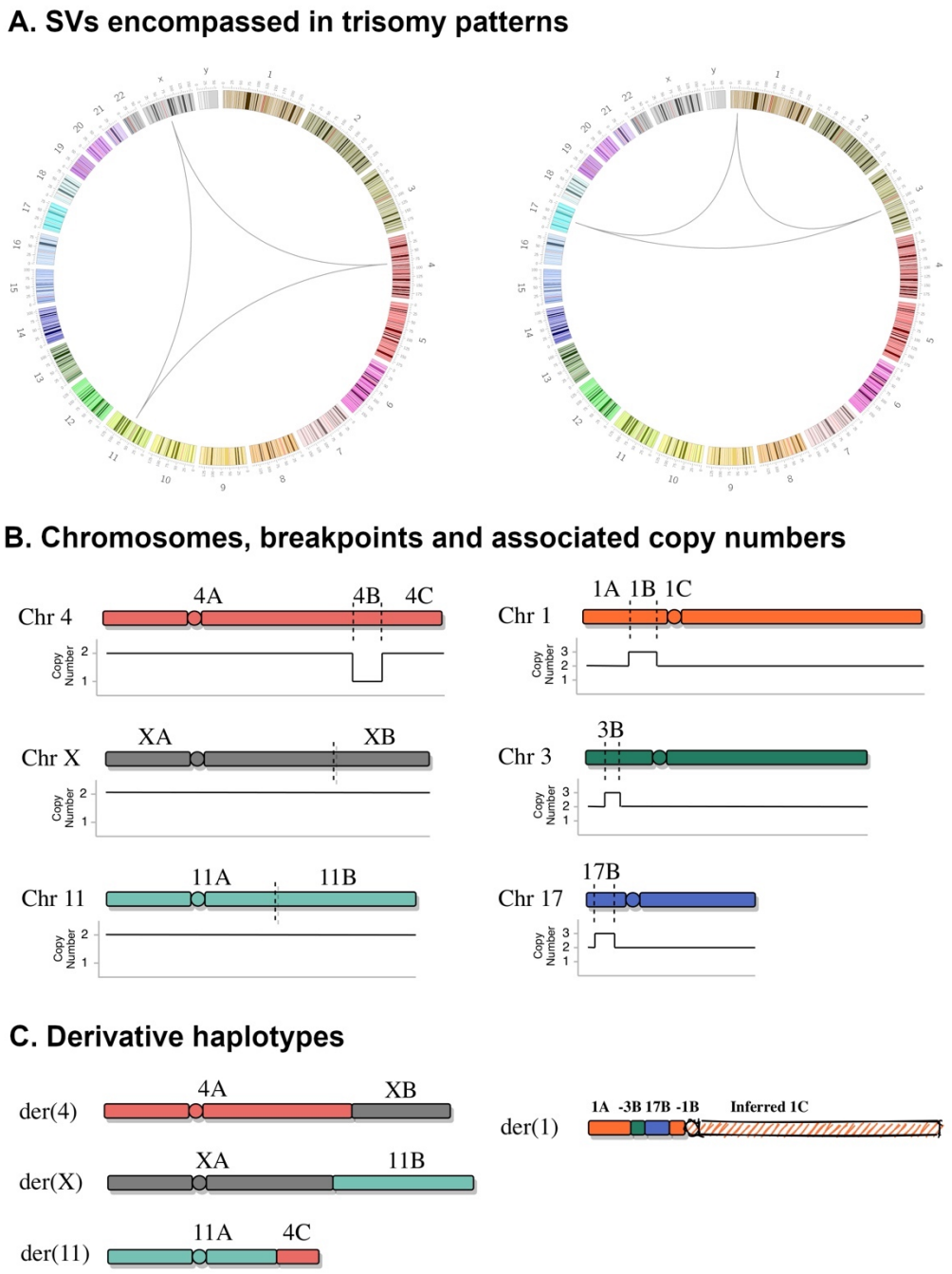


Figure 22. Two main genomic conformations were found from the trisomy pattern. **(A)** Circos plots show two different instances of the trisomy pattern **(B)** Original breakpoints (dashed lines) and associated copy number alterations. **(C)** The corresponding derivative chromosome and haplotypes. The minus sign "-" indicates an inverted genomic region. The "sketched-chromosome arm refers to an inferred region, due to the limitations of using paired-reads to end the total haplotype reconstruction.

Although we reconstructed several patterns from different samples, we found two main conformations regarding the resulting haplotypes, which were utterly distinct (Figure 22). The first one resembles chromoplexy, according to Baca and collaborators[92], the pattern formed by reciprocal translocations between different chromosomes and large deletions in some cases (bottom left part of Figure 22C). The explanation of this complex rearrangement suggests a DSB occurred in three different loci, followed by an incorrect rejoining of the genomic segments, leading to chains of rearrangements.

As shown in the bottom right part of Figure 22C, the second genomic conformation, named here as Cycles of templated insertions, could be phased to a single haplotype (one derivative chromosome) under the evidence of the split (at the junctions) and paired-end reads, by which we inferred the joined chromosome regions. The copy numbers associated with the involved breakpoints were higher than the neighboring genomic regions, resembling gene amplifications instead of translocations. This might implicate a copy-paste mechanism occurring likely during cell replication.

Additionally to these two genomic conformations, we also observed two other different conformations (Complex rearrangement I and Complex rearrangement II) that result challenging to reconstruct manually. Based on these observations, we developed a systematic classification of the trisomy events according to the associated copy number and the orientation of the involved genomic regions (Supplementary Table 2).

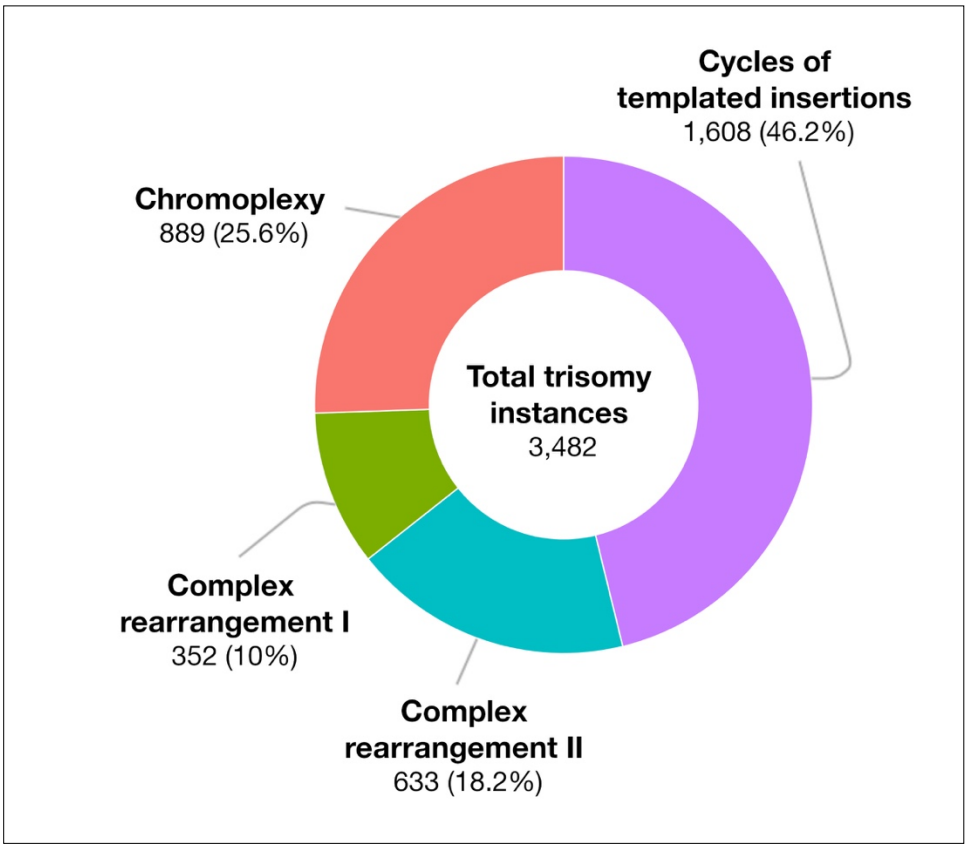


Figure 23. Trisomy pattern configurations found in the PCAWG cohort. In total, we found 3,482 trisomy events in all of the samples. The most abundant configurations were chromoplexy and cycles of templated insertions. The other two configurations were further studied in detail, as described in the following sections.

We observed that the main configurations were the most abundant in the PCAWG cohort, and the Complex rearrangement I and Complex rearrangement II were represented in a minor frequency (Figure 23). We further analyzed these two in detail, as described further in the 4.1.2.2 section.

4.1.1.5 Complex rearrangements arise in cell replication

The new insights in the trisomy pattern raised the hypothesis that it could derive from and during genomic DNA replication during mitosis. In a first instance, we evaluated whether the rearrangements arose simultaneously, in a particular moment of the replication process. We found that the regions involved in the

rearrangements of the trisomy events had an early replication trend, with a median of Wavelet-smoothed Signal value of 55.563 (see section 3.1.1.4). These values were related to regions near replication initiation, which have been significantly associated with high gene density, gene expression, Alu density, GC content, CpG density, and vertebrate nonexonic conservation[223].

Furthermore, given that our results suggest a complex rearrangement encompassing three SVs, we expected that the replication times of these three genomic regions were more similar among them than would be expected by chance, reflecting a spatial and temporal concordance. Indeed, we found that the pattern regions' replication times were significantly closer than those of three loci selected at random (Figure 24).

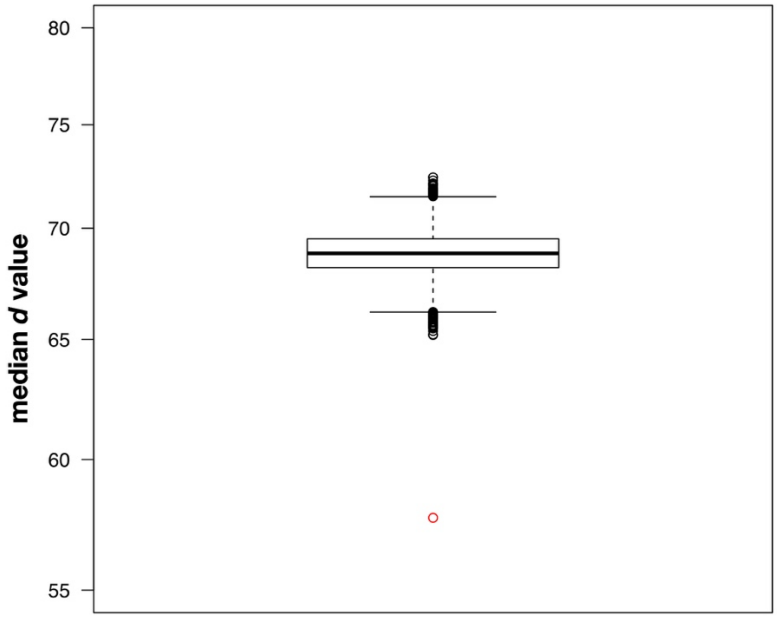


Figure 24. Boxplot diagram of the median *d* values from 10,000 randomizations and the observed value (spot in red).

At this time, an unpublished study by Li and collaborators[224] from the PCAWG consortium described a similar replication-based mechanism of structural variation that encompassed distinct chromosomal structures with copy number gains and usually inverted rearrangements. This work ultimately ended up as part of the PCAWG publications in February 2020[225]. These findings validated our results and added more insight into this new pattern. They referred to the pattern as “Cycles of templated insertions”, which could involve not only 3 SVs but also up to 8 SVs, within eight distinct chromosomes.

In that study, Li and collaborators proposed a strategy to classify complex chromosomal rearrangements based on models of sequential occurrences of simple SVs in order to determine whether the SVs that were close together occurring in the context of a single complex event. Additionally, cycles of templated insertions were confirmed through different approaches. First, analyzing the RNA-seq data, they found transcripts that comprised the joined genomic regions resulting from the rearrangements of the pattern. Second, the junction of the genomic regions from the derivative chromosome was validated experimentally by long-read sequencing. Third, the clonal fraction of the tumoral cells were more similar for the involved SVs than for random SVs. And fourth, copy number gains were almost the same for all the regions that encompassed the patterns. Together, this suggests that the SVs involved in cycles of templated insertions were probably acquired concurrently through a copy-and-paste mechanism.

Cycles of templated insertions were identified primarily on Uterine Corpus Endometrial Carcinoma, Ovarian cancer and Lung Squamous Cell Carcinoma, also coinciding with our findings. Strikingly, this study revealed that the vast majority of the SVs from the PCAWG cohort remained under the category of “Complex unclassified”.

At this point, we noted that: (i) there were still unclassified SVs that belonged to the undescribed complex rearrangements category; (ii) there could be different genomic configurations resulting from a single category of complex rearrangements, i.e., the trisomy pattern; and (iii) there were shortcoming of methodological approaches addressing the identification and classification of complex patterns of SVs. Given these points, we decided to pivot and propose a new strategy for the identification of patterns of complex chromosomal rearrangements in cancer.

4.1.2 Clustering and Graph Mining Techniques for Classification of Complex Structural Variations in Cancer Genomes

In collaboration with the Data-Centric Computing group at the BSC, we conceived a statistical framework to classify SVs and identify complex rearrangement patterns. This new framework is composed of unbiased statistical solutions that identified recurrent and significant complex rearrangements in PCAWG data (Figure 25).

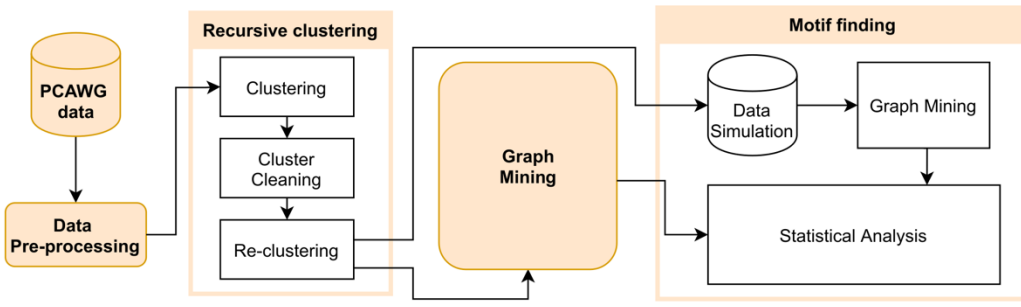


Figure 25. Workflow developed to identify complex rearrangements in PCAWG genomes. Simple data pre-processing was performed before implementing the recursive clustering. Then, the graph mining method was applied to find patterns.

Finally, the motif finding strategy was used to determine the statistically significant patterns.

We applied all the insights we gained while studying the trisomy pattern, towards the characterization of additional patterns we discovered. The manuscript is currently under review in the journal BMC Biodata Mining (DOI: 10.21203/rs.3.rs-476852/v1) and is included in the appendix of this thesis.

4.1.2.1 Development of a statistical framework to identify complex structural variation

We expect that the recurrent mutational patterns that deviate from random might have some underlying biological processes, somehow associated with tumor progression. The search for these complex rearrangements in such large amounts of genomic data without an a priori knowledge is a challenge, even if we use AI approaches. After a year trying to solve such a problem and given the characteristics of the trisomy pattern in PCAWG data, we resolved to narrow down the search to complex rearrangements encompassing from 3 to 6 clustered SVs (SVs that were significantly close together) that return to the original chromosome (the above-mentioned “cycles”), a generalization of the trisomy pattern.

To achieve this, we analyzed the whole set of SVs from all samples that belonged to different cancer types, taking into account the local distribution of SVs in every sample, and optimizing it using the global distribution across the entire dataset. Therefore, using this methodology, we could find recurrent patterns across the different cancer types and define the general features of each of the patterns.

We used the KDE method (see section 3.1.2.1) to consider both the density of clustered breakpoints and their closeness. Here, instead of setting a particular distance in base pairs between the breakpoints in a given chromosome, we

calculated the best bandwidth hyperparameter to determine the breakpoint clusters that could belong to the same event. First, we performed different KDE analyses, selecting a first bandwidth of 1000Hz to maximize the breakpoint clustering. Next, we performed a second KDE step to obtain high inter-cluster distances and also small intra-cluster distances, observing an optimal bandwidth value of 400Hz (Figure 26).

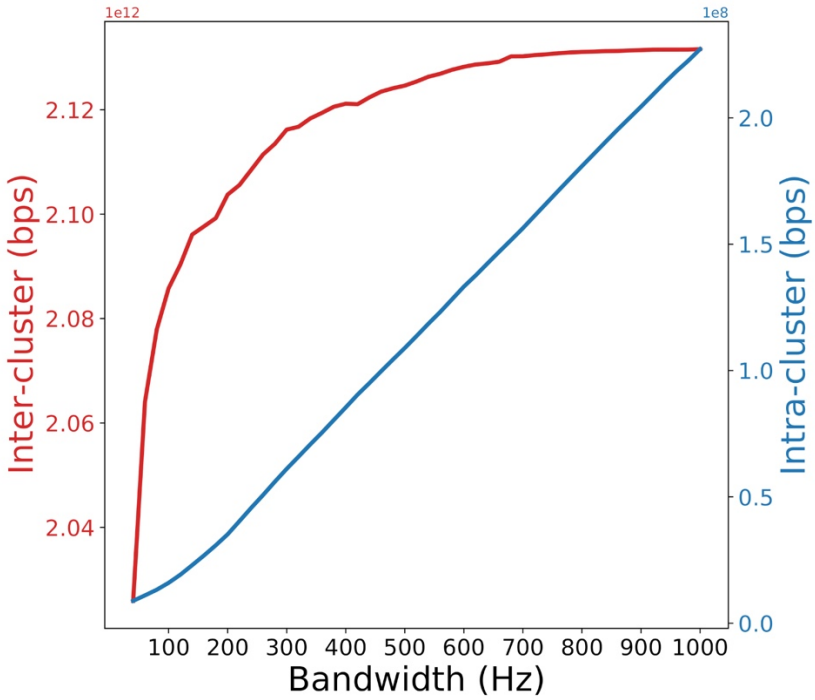


Figure 26. Total inter and intra-cluster distances across the entire PCAWG cohort. We used the 2-step KDE clustering with different bandwidth values to optimize the breakpoint clusters.

Once we had the breakpoint clusters, we determined they were not randomly close together by performing two statistical analyses. We compared the dispersion of breakpoints in simulated datasets with the dispersion from the original dataset, obtaining a *p*-value smaller than $1x10^{-5}$. This result indicated that the locations of the breakpoints did not follow a random distribution across

cancer genomes. Furthermore, we compared the cluster density in the simulated data and the original dataset, finding that the cluster density of the original dataset was unlikely obtained in a random simulation (p-value < $1x10^{-5}$). Therefore, the clusters we obtained implementing the 2-step KDE clustering contain SVs that are likely mechanically linked and not just random occurrences.

After setting the basis for defining breakpoint clusters, we were able to convert our pattern search across all the genome of every sample in a more straightforward graph search, using graph mining approaches. Next, we determined which SV pattern did not come from stochastic rearrangements by using the measure Abundance, where we evaluate the significance of the patterns against a random scenario. As shown in Figure 27, all the cycles assessed in this study were overrepresented (with positive values of Abundance). However, as the number of rearrangements of the cycle increased, the Abundance decreased, being the 3-SV pattern (trisomy-like pattern), the most overrepresented pattern.

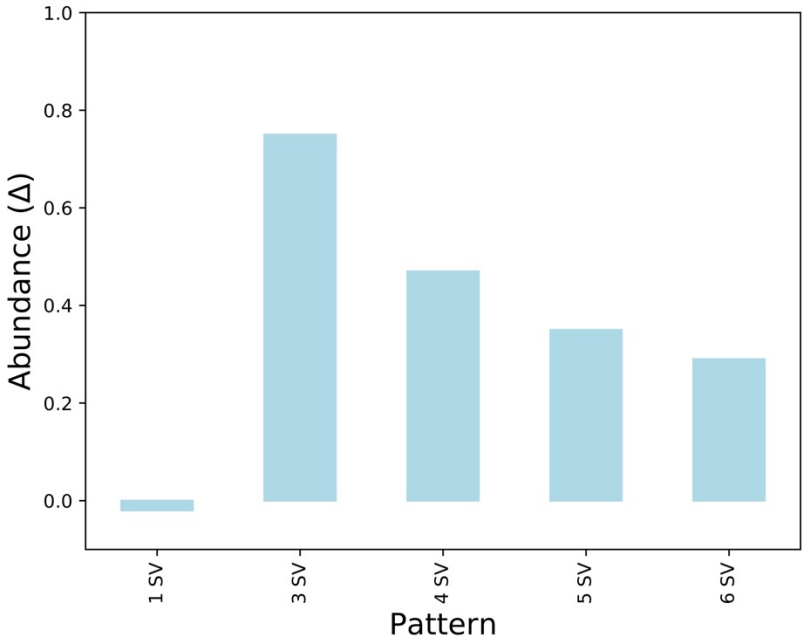


Figure 27. Abundance values for the analyzed cycles. Its value can go from -1, under-represented, to +1, overrepresented. The Abundance of a single rearrangement (1 SV) is also shown as a control value. Its value is 0 since we fix the rearrangements during the simulation of the random datasets, which means that its representation is the same in every dataset.

Furthermore, we confirmed that the most significant recurrent pattern in the PCAWG cohort was this 3-SV pattern, called the *triangle pattern* in this thesis. Its confidence was almost twice the confidence of the next simplest cycle, composed of only 1 SV more (Table 2).

Table 2. Statistical values for the evaluated cycles. The values obtained are defined as follows: Confidence, which provides the number of samples with at least one cycle occurrence. Average refers to the average number of cycles happening in all the samples. Finally, frequency, which is the sum of all the occurrences of the cycle across the whole dataset

Cycle size	Confidence	Average	Frequency
3	814	4.68	3,817
4	417	6.75	2,817
5	260	4.04	1,051
6	188	44.43	8,354

4.1.2.2 Characterization of the most significant recurrent 3-SV pattern across the PCAWG cohort

Given that the triangle pattern was the most overrepresented and recurrent event across the PCAWG cohort, we aimed at fully characterizing such complex rearrangement. Based on the previous trisomy pattern analysis results, we knew they could be observed in different genomic configurations. Therefore, we categorized the triangle patterns on the basis of the orientation of the two genomic regions at the breakend and associated copy number alterations, as described previously in sections 3.1.1.4 and 4.1.1.4. Here we uncovered the two additional configurations apart from chromoplexy and cycles of templated insertions. We found that the above-described Complex rearrangement I coincided with non-canonical chromothripsis, recently reported by Cortés-Ciriano and collaborators[102]. The Complex rearrangement II was a new pattern that did not correspond to any other pattern previously described.

We named this new pattern *Chromotrikona* (from the Sanskrit *trikona,* meaning triangle), characterized by frequent inverted rearrangements with no significant gains or losses of DNA. Unlike cycles of templated insertions and non-canonical chromothripsis, chromotrikona did not end up in a single derivative chromosome. Instead, it was characterized by a "cut and paste" mechanism derived into more than one host chromosome. Moreover, it also differed from chromoplexy, since chromotrikona did not encompass only balanced translocations but also inversions. Indeed, we found that chromotrikona and chromoplexy shared features such as the involved genomic regions, suggesting likely common molecular mechanisms (Figure 28).

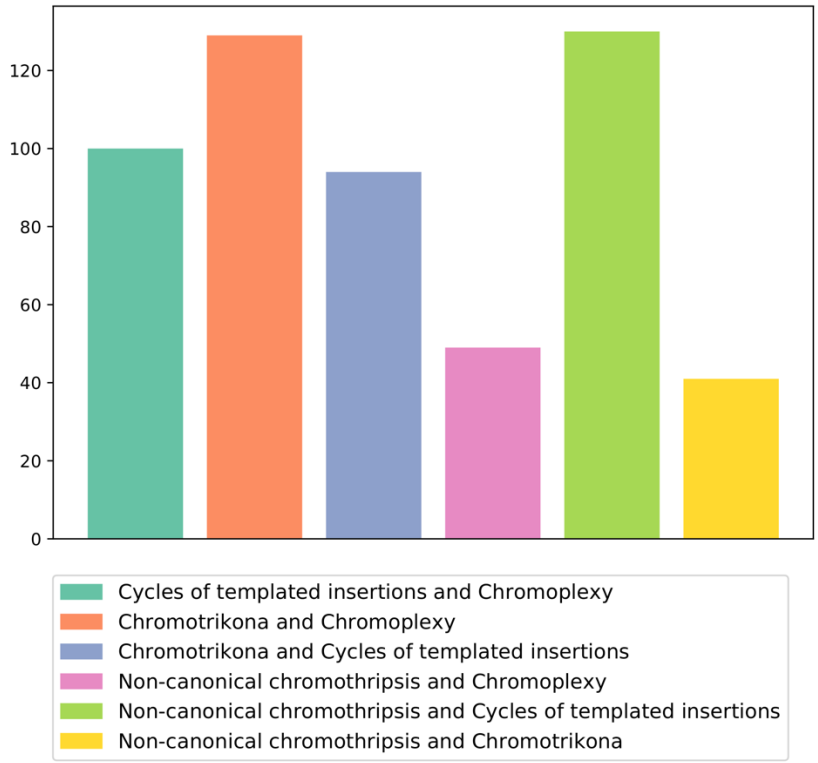


Figure 28. Common clusters between every pair of triangle types.

Strikingly, we found that the chromotrikona events were surrounded by several other SVs, showing a high complexity around this new rearrangement (Figure 29).

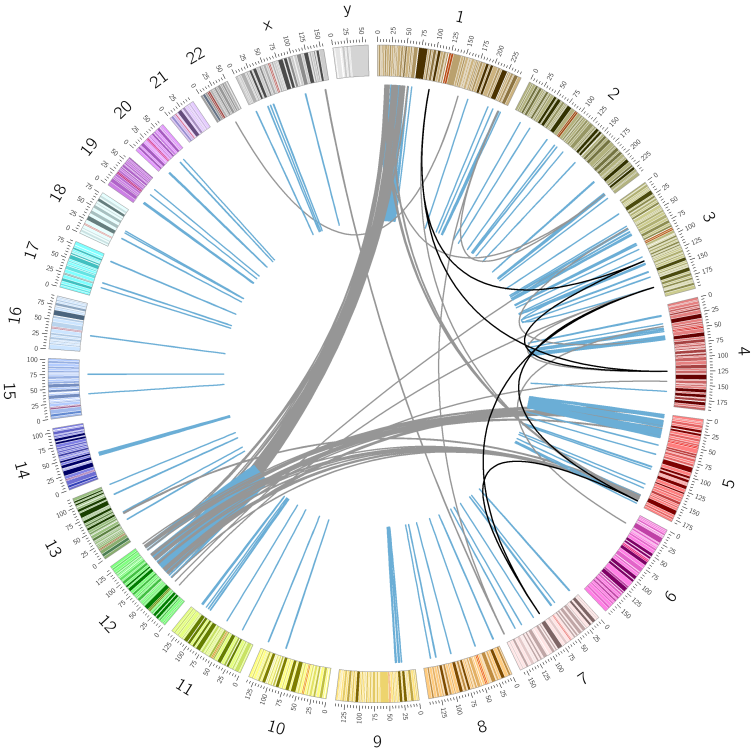


Figure 29. The complexity of complex rearrangements. Circos plot of a sample showing chromotrikona events (black lines). Chromotrikonic rearrangements co-occur with other SVs, reflecting the complexity of catastrophic events.

However, the number of instances of chromoplexy and chromotrikona over the distinct cancer types were significantly different (Figure 30). Compared to the three categories, chromotrikona predominated only in Kidney Renal Clear Cell Carcinoma (KIRC), and was the less represented pattern in Bone cancer (BOCA), Liver Hepatocellular carcinoma (LIHC), Head and Neck Squamous Cell Carcinoma (HNSC), Skin Melanoma (MELA) and Soft Tissue Liposarcoma (SARC).

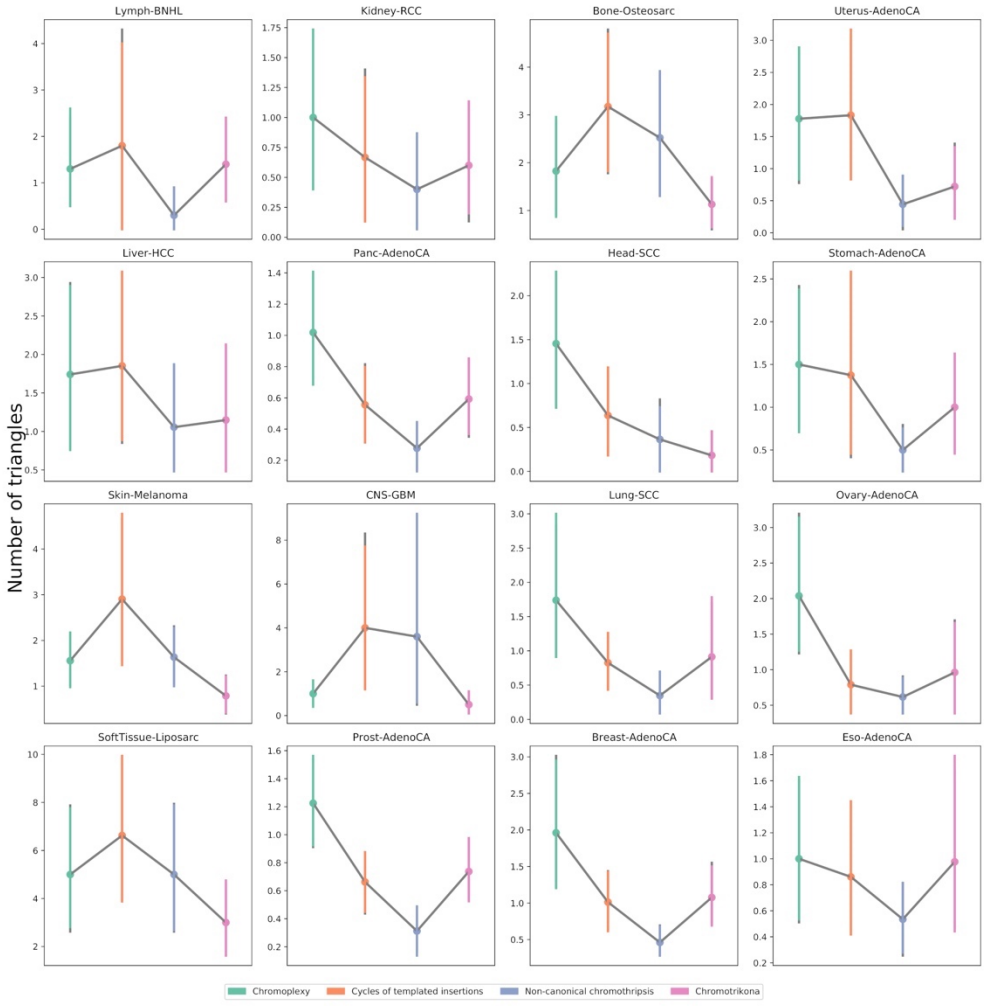


Figure 30. Confidence intervals of the mean of the frequency for each triangle type throughout cancer types. Only cancers with more than 10 samples with triangles were showed.

Taken together, our results highlighted the heterogeneity and plasticity of complex rearrangements in terms of frequency and recurrence across different tumors. Interestingly, we found evidence suggesting that catastrophic events play major roles in sculpting the genome structure in cancer genomes and possibly contributing to tumor evolution.

4.2 Identification and characterization of mRNA alterations in metastatic breast cancer

Given the key role of transcriptomics in the molecular characterization of metastasis (see section 1.3.3.2), we studied different types of mRNA alterations that could be associated with metastasis derived from breast tumors. For that, we focused on characterizing the expression profiles in a total of 82 metastatic RNA-seq samples, and in parallel, the identification of potential new transcripts in the form of mRNA fusions.

4.2.1 Searching for gene expression patterns associated to metastasis

To evaluate the expression patterns related to the metastatic process, a first methodological approach involved the analysis of changes in gene expression across metastases from different patients with lethal metastatic breast cancer. We relied on two datasets: the SE dataset formed by 64 samples from 10 different tissues from 9 patients and, the PE dataset composed by 18 samples from 5 tissues from patient 302.

With the final aim of identifying specific expression patterns related with the metastasis, we first evaluated how the 64 metastatic samples from 9 patients and 10 different tissues clustered according to their gene expression profiles. Here, the gene-level and transcript-level expression quantifications were calculated (with RSEM[215] v.1.3.0), and the transcripts per million (TPMs) were used to perform a Multidimensional Scaling Analysis (MDS). Opposite to what occurs in gene expression in samples from healthy tissues[106], the gene

expression landscape was more similar across metastases of the same patient than across metastases in the same tissue from different patients (Figure 31).

In other words, all metastases of a given patient shared more common gene expression patterns than all samples of a given tissue from the nine patients, even if they were metastases in the same tissue. These results indicate that metastases likely carry common alterations from the primary tumor conserved in the metastatic cells, resulting in shared gene expression patterns across different tissues in the same individual.

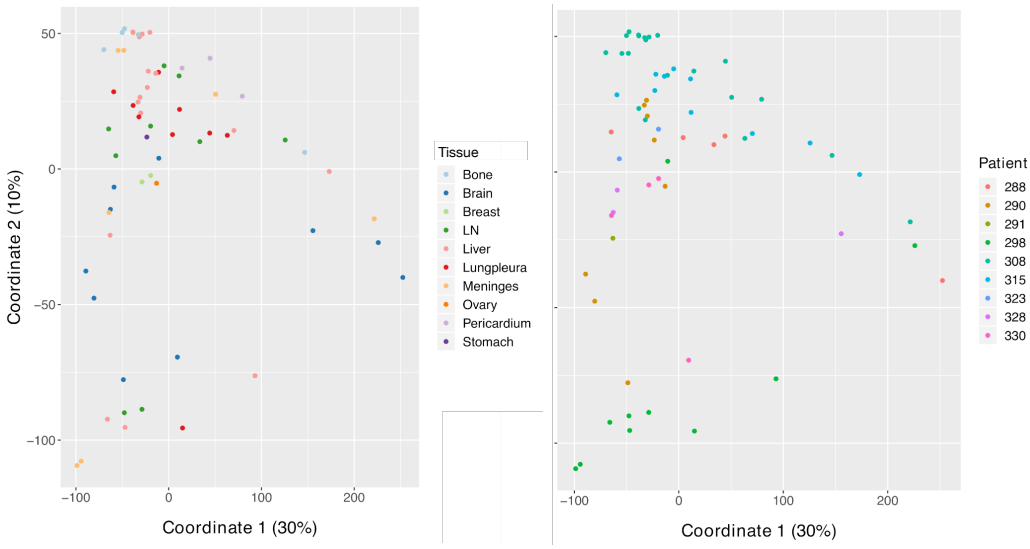


Figure 31. Multidimensional Scaling Analysis of gene expression profiling from 64 metastatic samples from 9 patients. Each point represents a metastatic sample, colored according to the tissue of origin (left plot) or to the patient (right plot). Samples clustered by patient, while no evident cluster appeared by tissue, indicating that the gene expression profiles were more similar between metastases from a patient than between metastasis from a given tissue. LN: Lymph Nodes

We further analyzed the metastatic samples from patient 302 (PE dataset) and observed that the gene expression profiles from colocalized metastases, such as lung, diaphragm, and pericardium showed higher similarity (Figure 32). This observation indicates that, within the same patient, metastases showed

similarity by tissue. This result was further integrated into the fusion transcript downstream analyses in section 4.2.2.5.

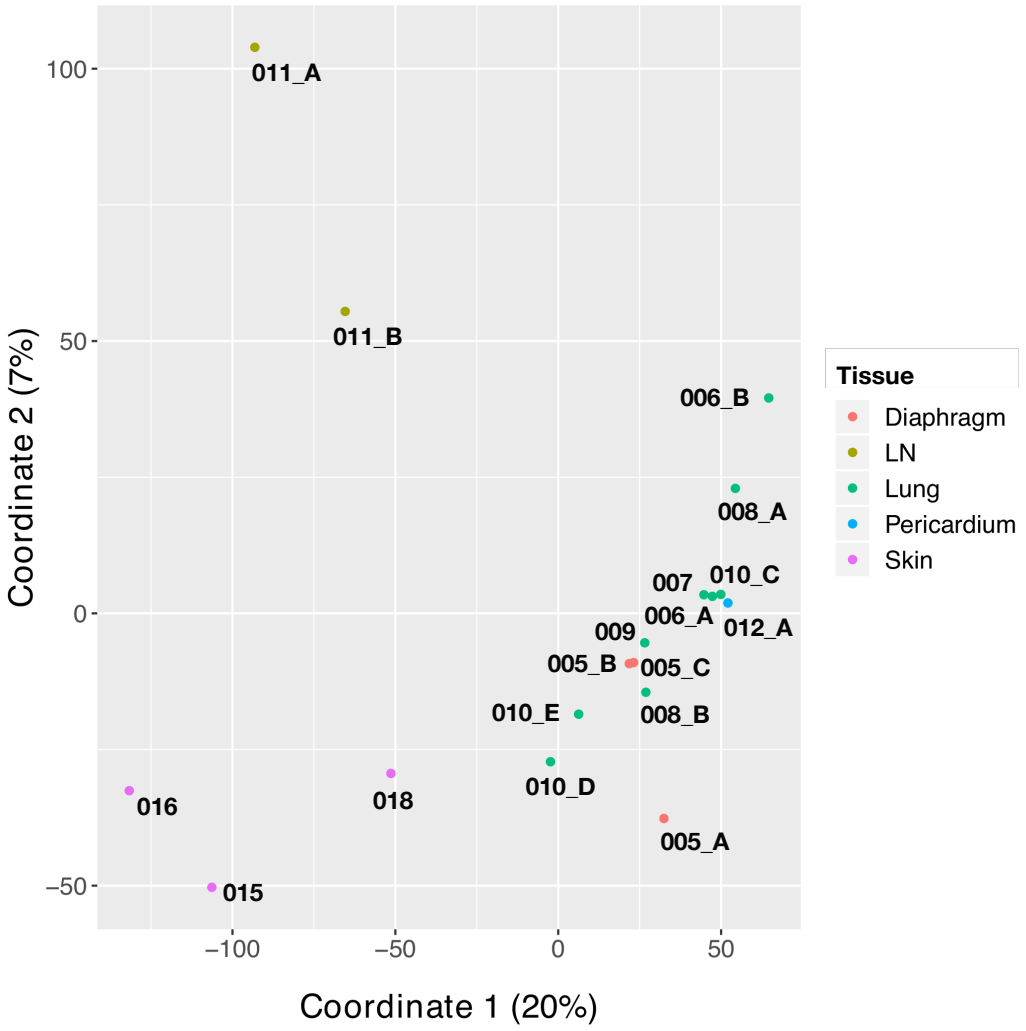


Figure 32. Multidimensional Scaling Analysis of gene expression profiling from the 18 metastatic samples from patient 302. The samples were discriminated by tissue (color dots). LN: lymph nodes.

Having this general outline of gene expression profiles across metastases, we aimed at addressing the following questions: what do the metastases share regarding the gene expression through the same patient? Which are the genes

differentially expressed across metastases from different tissues? Nevertheless, since we studied fusion transcripts in parallel, which showed more relevant results regarding our scientific question, we continued this project in the context of fusion transcripts in metastases. The study of gene expression patterns associated with metastasis is currently on hiatus.

4.2.2 Detection of fusion transcripts in metastatic breast cancer samples across patients

In order to detect new transcripts in metastatic breast cancer samples, we focused on the identification of fusion transcripts in different patients, how they were recurrent and their likely relationship to metastasis.

4.2.2.1 Generation of a manually curated pipeline to detect fusion transcripts

Several algorithms and bioinformatics tools have been developed to identify fusion transcripts from RNA-seq data[195]. However, their strategies generally show two main issues: first, users are unable to control every single step in the identification process, and second, these programs favor specificity at the expense of sensitivity[159]. Consequently, our first goal was to design an accurate pipeline that we could control and select filters affecting specificity and sensibility when convenient.

Initially, we analyzed the SE dataset, which was composed by reads with a length of 50 bp. In the first step, we identified informative reads from RNA-seq alignment files (BAM files) that could inform fusion transcript events. Since SE RNA-seq data were available, we based our analyses on split reads, which did not map contiguously in the genome, but spanned two different genomic

regions. Therefore, we identified and extracted all split reads and also the unmapped reads from the BAM file. We then performed a BLAST search against the human transcriptomic database, to obtain all the reads that aligned to two different transcripts.

From the BLAST results, we selected the reads that aligned one portion to one transcript and the other portion to another transcript from a different gene. In this regard, we faced the fact that besides being single-read sequenced, the reads from these libraries were short (50bp). This represents a challenge regarding the amount of accurate information from a given portion of the read that aligned to a gene transcript. We performed several analyses, setting different cutoffs for the length of the two parts of a read to predict a fusion transcript candidate, in combination with cutoffs of the percent of identity between the two portions of the read with the two gene transcripts. Here, we observed that the number of predicted fusion transcripts did not vary when we set values lower than half of the read (25bp) as the minimum alignment length of a read to a gene transcript. Similarly, the number of fusion transcripts did not vary when we set identity values lower than 100% in those alignments (Table 3).

Table 3. Number of fusion transcript candidates identified using different parameters from BLAST results in one sample from SE data.

N Fusion Transcripts	N Supported Reads	% Identity (>=)	Alignment length (>=)
10,686	45,422	98	15
10,686	45,422	100	20
10,686	45,422	100	25
6,307	*10,029*	*100*	*25*

N: number. >=: more or equal than. The last row in *italic* (highlighted in grey) indicates the results obtained restricting the analysis to those that did not have better alignments than those corresponding to the two transcript partners of the fusion.

Since a given read could align to multiple reference transcripts with the same percentage of identity, but with varying lengths of alignment, we removed reads whose best alignment was to a single transcript, rather than two transcript partners (see Table 3, last row). Furthermore, we allowed an overlap of a maximum of 7 bp between the two partners of the fusion (corresponding to the fusion point, see Figure 33).

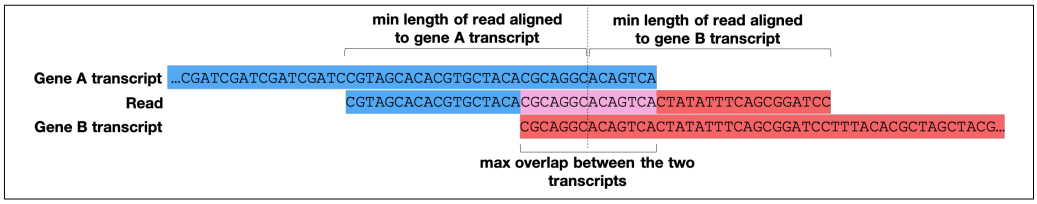


Figure 33. Parameters to identify the split reads that support a putative GenA-GenB fusion transcript.

Thus, to obtain a preliminary list of fusion transcript candidates supported by split reads across all SE samples, we set all these four parameters into the selection step of our pipeline: (1) the split reads that did not have better hits, and (2) had a minimum of alignment length of 25bp to each gene transcript partner, (3) both with 100% of identity, (4) allowing an overlapping of maximum 7 bp between the two transcripts of the fusion.

Next, having the preliminary list of fusion transcript candidates, we designed different filters, applied consecutively to filter out likely false positives (see Figure 16), in the following order:

1. *Pseudogenes and overlapping genes*

First, given the short length of alignment threshold (25 bp), forced by experimental constraints, it became hard to distinguish between a transcript from a pseudogene and sequencing errors on the corresponding gene transcript. As a result, we decided to discard the candidates that had

pseudogenes as transcript partners. Second, we excluded fusion transcript candidates composed by genes overlapping in the genome. In that situation, it was not possible to determine whether the split reads represented one single transcript or the other, or whether it was a true positive fusion transcript.

2. *Lowly supported fusion transcript candidates*

To have supporting evidence to predict fusion transcripts, we set a minimum of 3 split reads to report a candidate.

3. *Fusion pairs with sequence similarity*

Next, we assessed another filter regarding the approach applied here. Since we identified the transcript partners through sequence similarity to a human transcriptomic database, the sequence similarity between each pair of partners also had to be considered. To measure this, we performed an all-vs-all BLAST in the transcriptomic database, searching all significant sequence alignments between any two reference transcripts of the genes. The significance was evaluated in two ways. On one hand, every isoform aligned with another isoform from a different gene with an E-value <= 10^{-3} was reported. On the other hand, since the reads were 50bp long, we evaluated whether any two reference isoforms from different genes had 100% identical 50 bp segments. Then, if a fusion transcript candidate had transcript partners that were similar in sequence based on any of these two parameters, it was filtered out.

4. *Promiscuous fusion transcripts*

Finally, fusion candidates with *promiscuous transcripts*, defined as those transcripts found to fuse to multiple partners (>3) within a single sample, were discarded from the putative fusion transcripts list.

4.2.2.2 The repertoire of fusion transcripts across patients identified by SE data

As mentioned in the section above, we initially analyzed the 64 SE samples from 9 patients and manually inspected the outcomes after every single filter step (Figure 34A).

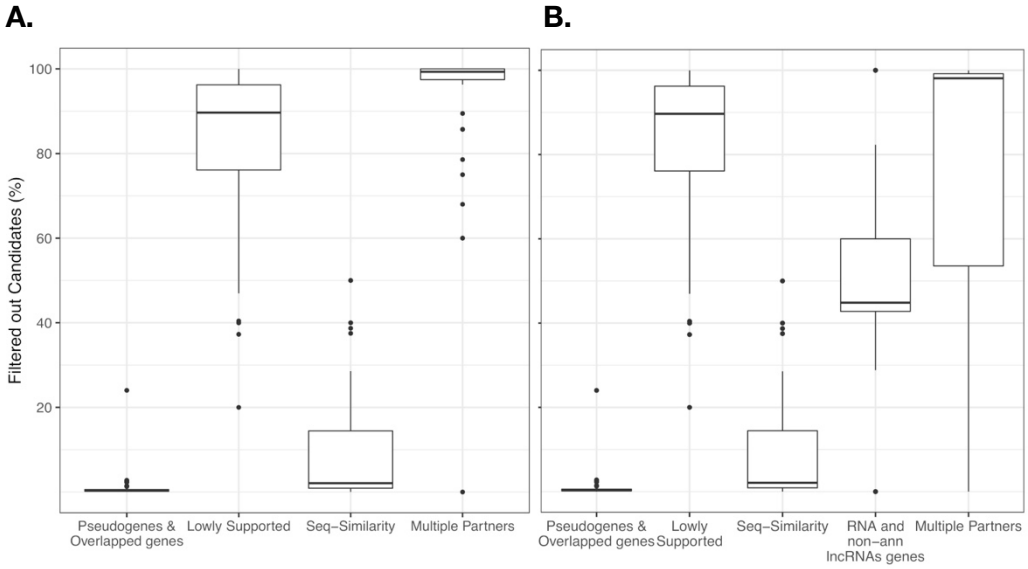


Figure 34. Percentage of candidates filtered out in each filter step of our pipeline from 64 metastatic samples from 9 patients (SE dataset). A. Filter 1: Pseudogenes & Overlapped genes. In this filter, the candidates that were pseudogenes or the fusion candidates A-B, where A and B were transcripts from overlapping reference annotated genes, were filtered out. Filter 2: Lowly Supported. The fusion candidates that had less than 3 supporting reads were excluded. Filter 3: Seq-Similarity (Sequence similarity). Fusion candidates A-B, where A and B were transcripts similar in sequence, were removed. **B** Filter 4: RNA and non-ann lncRNAs genes (RNA and non-annotated long non-coding RNAs genes). In a second round, we filtered out the candidates whose transcript partners belonged to RNA genes or non-annotated lncRNAs. Filter 5: Multiple Partners. The fusion candidates composed of transcript partners found to have multiple fusion partners (promiscuous transcripts) were finally filtered out.

We observed that a high percentage of the fusions contained transcripts that belonged to RNA genes and non-annotated long non-coding RNAs (lncRNAs). In fact, these RNA genes and non-annotated lncRNAs corresponded to the most expressed genes in the metastatic samples. Since those genes are rich in repetitive and low complexity sequences, we could not rely on them as true positives using the evidence we had. Therefore, we decided to filter out such elements in a final step (Figure 34B).

Interestingly, we further observed that the majority of the candidates were being discarded due to have promiscuous transcripts (Figure 34). Strikingly, the pair of fused partners from each of these sequences did not share common sequences, suggesting that those candidates were not potential artifacts derived through multiple matching of these common sequences during BLAST. Therefore, to obtain better insights, we manually explored these candidates, analyzing their supporting reads. We observed three different scenarios in a given sample: i) every fusion transcript with promiscuous transcript partners was supported by a unique set of reads; ii) an identical set of reads supported all the candidates with a given promiscuous transcript; iii) different candidates with different promiscuous transcripts were all supported by the same set of reads. This last scenario likely reflects false-positive fusion transcripts, and therefore, we decided to filter them out (Table 4). In the other two scenarios, we did not have enough evidence to determine the credibility of those fusion transcripts.

Table 4. Number of samples and patients from the SE dataset with fusion transcript candidates after applying filters. These numbers were calculated by sequentially applying the filtering steps indicated in the first column.

Parameters to filter out fusion transcript candidates	Default*		0.001*	
	N samples	N patients	N samples	N patients
SEQ-SIMILARITY	63	9	27	9
RNA GENES AND NON-ANN. LNC-RNAS	60	9	22	7
WITH PROMISCUOUS PARTNERS	45	9	7	6
WITH REAL PROMISCUOUS PARTNERS**	16	5	4	4

*Two parameters were tested to filter out fusion candidates that had fusion pairs with sequence similarity. We excluded fusion candidates A-B, when the transcript partners had significant sequence alignments in terms of: A and B had a 50bp alignment length with 100% identity (Default), or A and B had an alignment with an e-value < 10^{-3} (0.001).
**This filter refers to excluding fusion candidates with promiscuous transcripts that were supported by an identical set of reads.

In search of an additional independent validation of our findings, we further analyzed the 64 samples using the STAR-fusion algorithm[218]. However, the detection of candidates with promiscuous transcripts was hampered because STAR-fusion, as most bioinformatic tools, discard these kind of fusions as false positives. In fact, we did not find any fusion with promiscuous transcripts using STAR-fusion.

4.2.2.3 Identification of massive fusion transcripts in metastatic breast cancer

To further evaluate the robustness of our findings, we analyzed an additional dataset of 18 metastatic samples from 5 different tissues from a single patient (ID: 302). RNA-seq data for these samples were obtained with PE and longer reads (150 bp, see Figure 15), offering the advantage of exploring additional evidence to determine whether these promiscuous transcripts really exist.

We applied the pipeline described above for SE dataset, finding the same pattern in this new dataset: the majority of the initial predictions comprised

fusions with promiscuous transcripts. To ensure the false discovery rate was controlled, we only kept fusion transcripts supported by unique sets of reads. Therefore, we ensured that each of the fusion candidates that had a given promiscuous transcript was supported by different split reads. Surprisingly, applying the filter of multiple partners, we confirmed that fusions with promiscuous transcripts still prevailed (Figure 35).

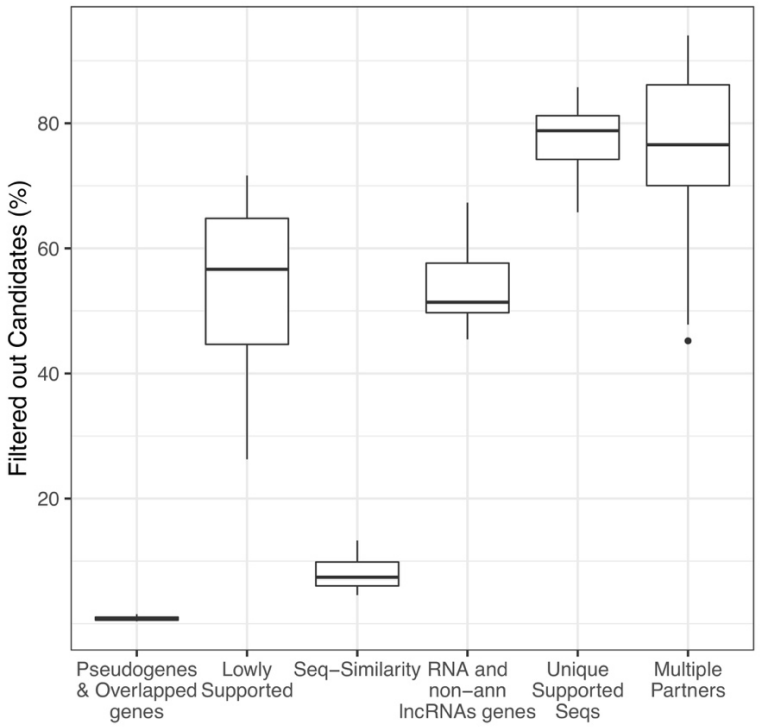


Figure 35. Filtering of fusion transcript candidates by applying different filter steps. Boxplots represent the percentage of filtered candidates in the 18 metastatic samples from case 302 for each filtering step used to discard false-positive candidates. Seq-Similarity: Sequence similarity. Seqs: Sequences. RNA and non-ann lncRNAs genes: RNA and non-annotated long non-coding RNAs.

To further improve the prediction of fusion transcripts, we took advantage of the paired-end reads information. We kept only the candidates supported by split reads (covering the candidate fusion point) and by spanning read pairs,

where the two reads of the same pair mapped each in one of the two partners composing the fusion transcript (Figure 17.II.). In case of split reads, their matching pair read could match any region of the fused transcripts, as is shown in Figure 17.II.

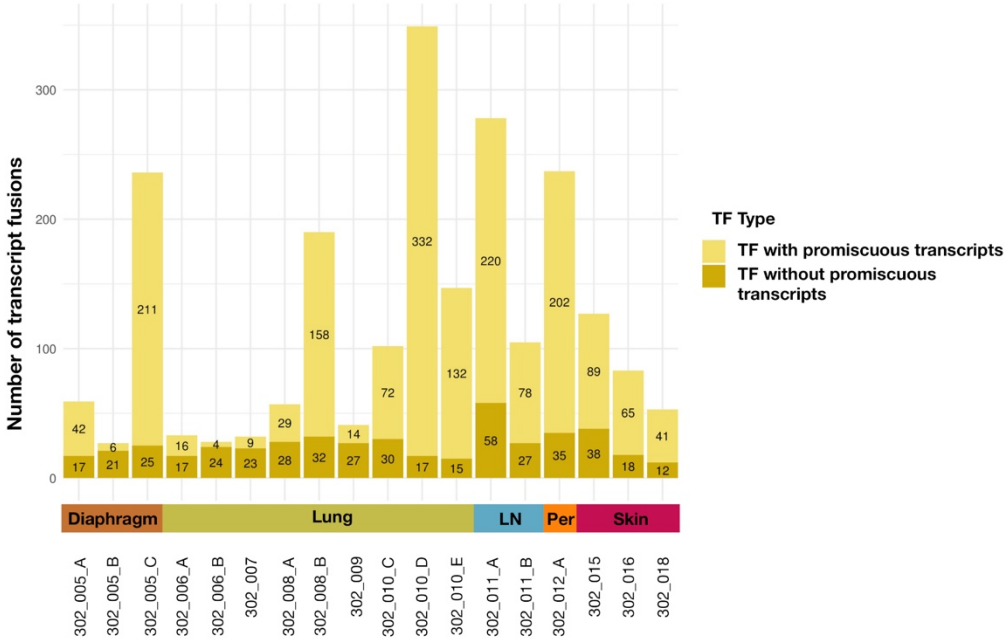


Figure 36. Number of fusion transcripts by sample, classified by the presence of promiscuous transcripts. Each bar represents a metastatic sample from case 302; the samples were sorted and plotted according to their tissues as indicated at the bottom bar. TF: Transcript Fusion. LN: Lymph Nodes. Per: Pericardium.

Interestingly, these promiscuous transcripts were present in the vast majority of the identified fusions in almost all the samples (Figure 36) and were found fused with up to 145 different transcripts in a given sample (Figure 37). Most of the promiscuous transcripts, however, had less than 20 partners, while 19 were fused with 20 to 60 different partners, and only 5 transcripts showed more than 70 partners (Figure 38).

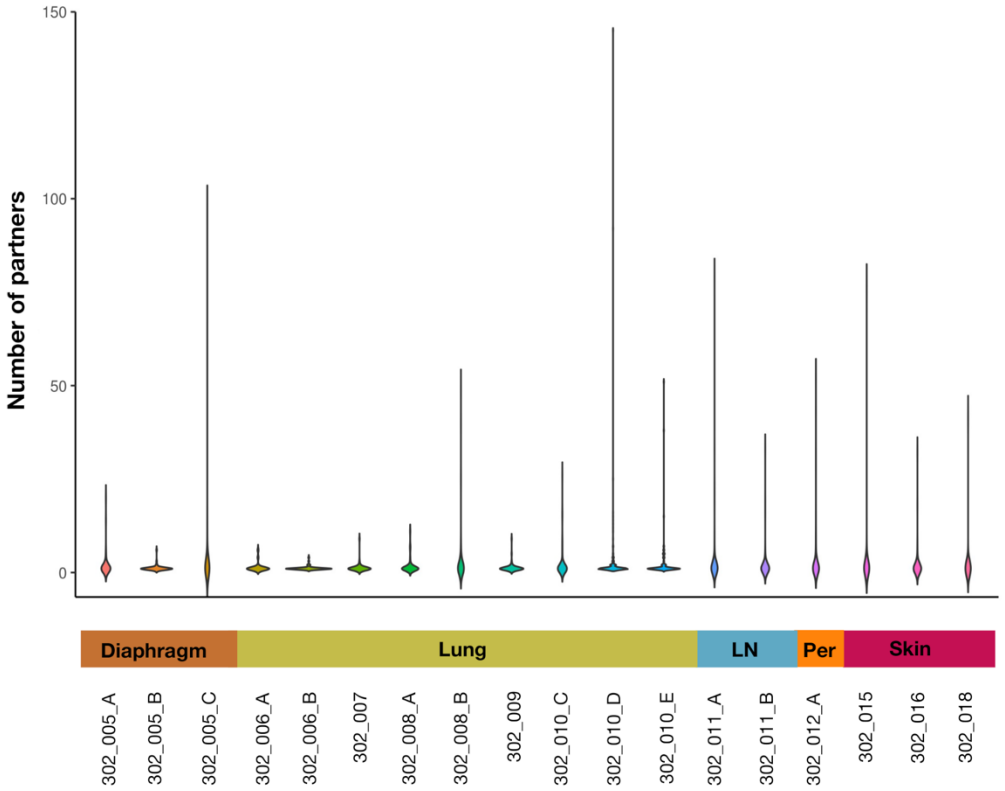


Figure 37. Distribution of the number of partners per transcript across the PE samples. Violin plots were made with the number of partners per transcript from the fusion transcript candidates predicted in each metastatic sample. In 8 out of 18 samples, transcripts with more than 50 fusion partners were found. LN: Lymph Nodes. Per: Pericardium.

The promiscuous transcripts did not share the same transcript partners within a given metastatic sample, and consequently, we observed more than 2,000 transcripts fused only once (Figure 38).

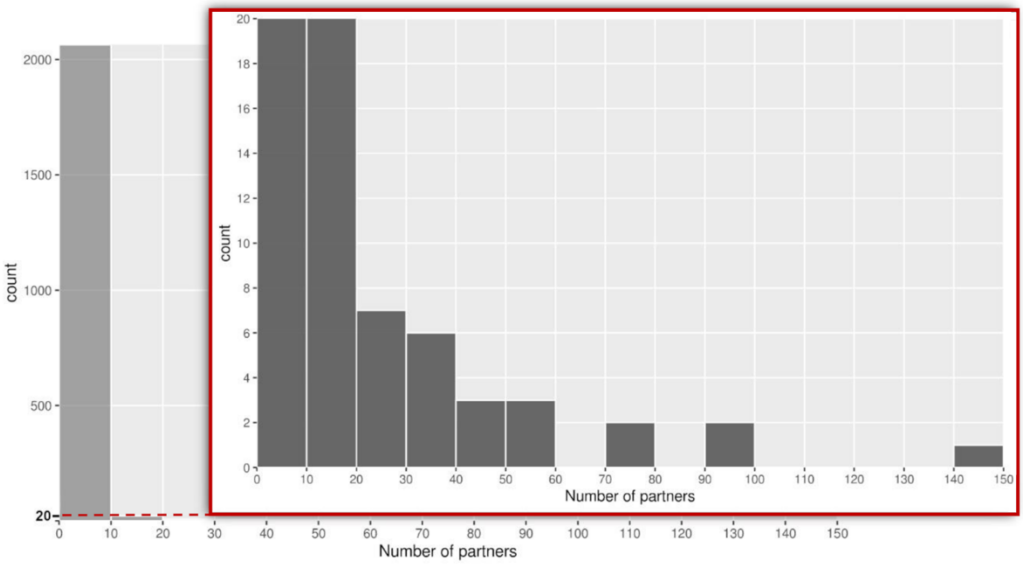


Figure 38. Histograms of the number of partners per transcript fused in the 18 metastatic samples from case 302. The histogram at the front is a zoom-in of the histogram from the background to appreciate the values under 20.

Considering the novelty of these striking findings, we sought to investigate whether the fusions with promiscuous partners were artefacts due to the sample quality. To address this issue, we used the RNA Integrity Numbers (RIN scores), a measure of the quality of extracted RNA samples. As shown in Figure 39, there was no correlation between the sample quality and the proportion of fusions with promiscuous transcripts (Pearson Correlation coefficient r: 0.37). Indeed, good quality samples (with RIN scores above 8) had more than 70% of the predicted fusions with promiscuous transcripts. This result suggests that the fusions with promiscuous transcripts were not a consequence of methodological artefacts, but instead, they were possible alterations observed in the RNA of these metastatic samples.

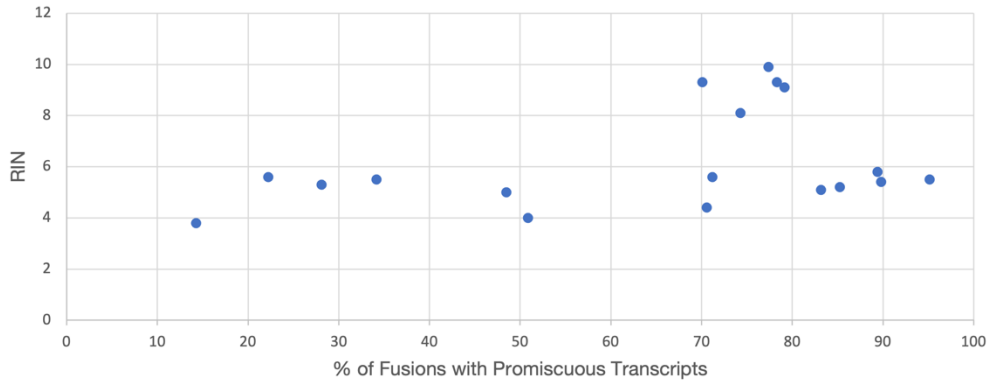


Figure 39. Scatter plot of the percentage of fusions with promiscuous transcripts and RIN values from 18 PE samples. Each blue dot represents a sample. Pearson Correlation coefficient r: 0.37.

Under this evidence, we decided to validate the fusions with promiscuous partners by both a manual reconstruction of the fusions (in-silico) through the supporting reads and an experimental procedure, as described in the next section.

4.2.2.4 *In-silico* and Experimental Validation of promiscuous partners

To validate the fusions, we selected the metastatic sample with the highest percentage of fusions with promiscuous transcripts from the PE dataset (sample ID: 302_010_D, lung sample). Then, we used the reference transcript sequences from the transcriptomic database to manually map the supporting reads and reconstruct the predicted fused regions. Given the fact that the transcript NM_013230.3 from the *CD24* gene was the promiscuous transcript with the highest number of partners in sample 302_010_D, we aimed at assembling the predicted fusion transcripts with *CD24* and the partners (i) *HSP90AB1*, (ii) *MALAT1* and (iii) *PEG10*.

Strikingly, under the evidence of supporting split reads that pointed out various junctions of the two transcripts, we found different fusion points between each pair. This suggested that a given specific pair of transcripts was prone to suffer multiple independent fusion events, resulting in different fusion products. Similarly, further paired-end reads supported distinct fused regions (see Figure 40 as an example). We found cases in which several PE reads supported a putative fusion between two transcripts, but no split reads agreed with the fusion. We also found the opposite, there were split reads not in concordance with the transcript regions supported by the PE reads.

Consequently, we determined the most likely fusions between pairs of transcripts, choosing only those with the highest number of overall supporting reads (only split reads, split and PE reads, or only PE reads). Interestingly, in all fusions analyzed, both partners were fused in their 5′–3′ orientation. This result could suggest that the fusion happened at the transcript level, where only 5′–3′ molecules exist.

Our findings also showed that *CD24* could be found both at the 5′ and at 3′ end in fusions with *HSP90AB1* and *PEG10* (Supplementary Table 1). For instance, there were two fusions between *CD24 and HSP90AB1* in which *CD24* was the 5′ partner, and three fusions located in 3′. Similarly, we found one 5′ *CD24*–*PEG10* 3′ fusion and another one in the form 5′ *PEG10*–*CD24* 3′. Instead, *CD24* was the 3′ partner in all the fusions with *MALAT1*.

Overall, we reconstructed 13 fusions: 5 fusions between the *CD24* and *HSP90AB1* genes*,* 6 fusions between *CD24* and *MALAT,* and 2 between the *CD24* and *PEG10* genes*.*

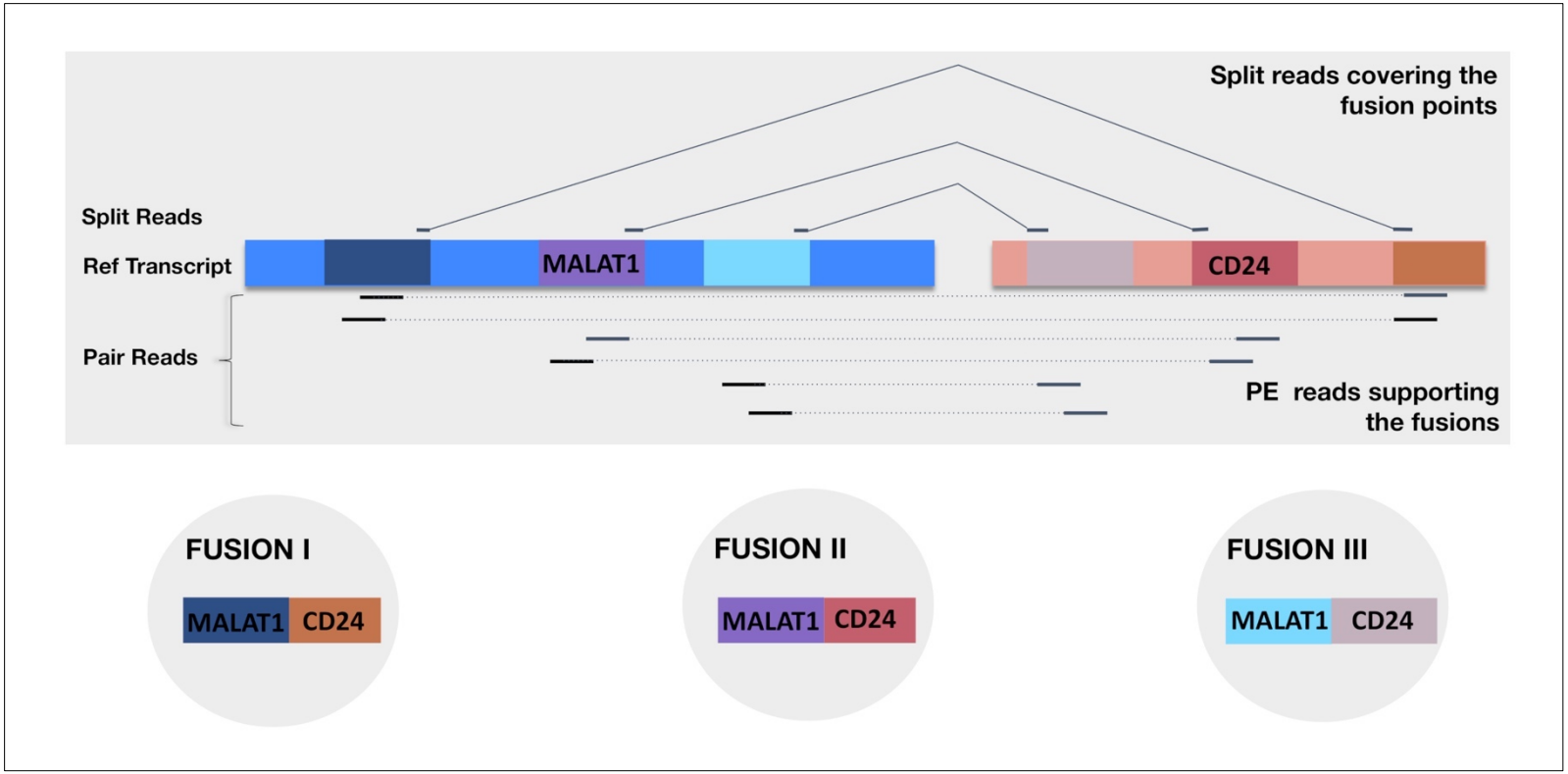


Figure 40. Different fusion points between the promiscuous transcript and its partners. Schematic scenario where transcripts from the *MALAT1* and *CD24* genes were found to be fused at different points. Split reads (at the top of the figure) were used to identify the fusion points, and PE reads to predict the flanking regions of each partner. The three fusions that could be identified combining all read information are represented at the bottom.

Once we had reconstructed all these 13 different fusions between the promiscuous transcript *CD24* and the three *HSP90AB1, MALAT1* and *PEG10* partners, we designed primers to confirm the presence of the fusion transcripts through RT-PCR and Sanger sequencing (Supplementary Table 1). The low quantity and quality of the available RNA sample allowed only to test 9 out of the 13 fusions. We observed different bands in the electrophoresis gel using a given set of primers, and no bands with the expected sizes in some cases (Figure 41).

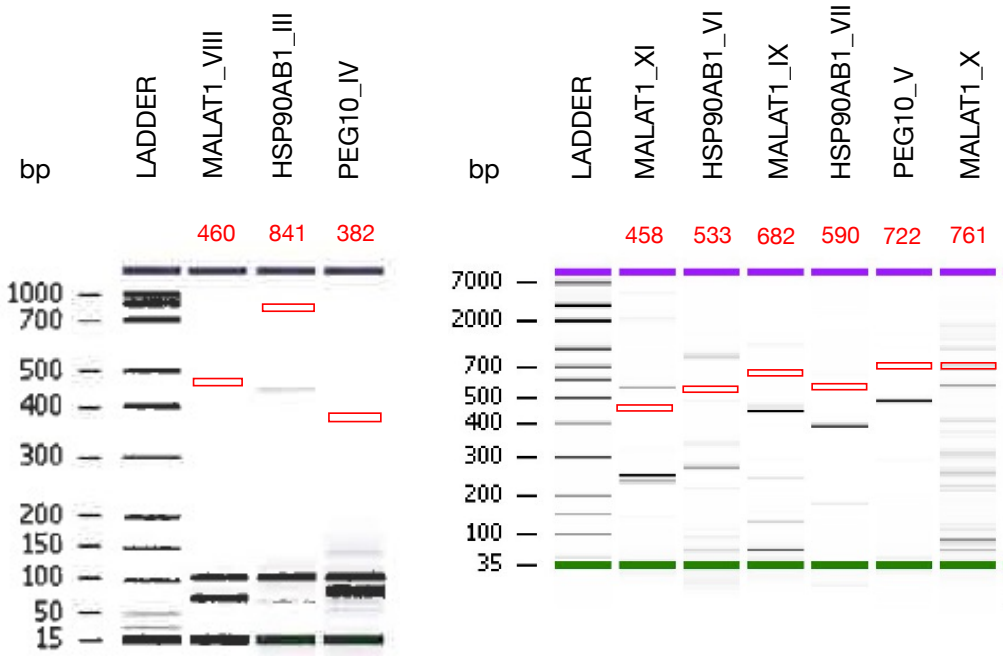


Figure 41. Electrophoresis gels from the nine RT-PCRs performed in this study. The expected sizes (values specified in red at the top) are shown in red squares in the gel images.

We validated by Sanger sequencing the PCR products from two different fusions between the *CD24* and *HSP90AB1* genes in the form of 5′ *CD24*–*HSP90AB1* 3′ (Supplementary Table 1and Figure 41, Fusion IDs: HSP90AB1_III and HSP90AB1_VII). The two fusions were amplified and sequenced. Although, we sequenced the partners that were predicted *in-silico*, we could not assemble

the sequences of the entire fusions since the length of the forward and reverse sequences did not allow overlapping.

Then, we designed another set of primers to confirm the fusion point of the HSP90AB1_III fusion (Supplementary Table 1, Nested_HSP90AB1_III). In this Sanger assay, the *HSP90AB1* transcript partner was sequenced using the forward primer designed on the *CD24* transcript, indicating that the two transcripts were fused and thus, being part of the same cDNA molecule. Nevertheless, the fused transcript regions predicted using the RNA-seq data represented only a part of the fusion sequenced with Sanger. We sequenced an extended region of the transcript from the *HSP90AB1* gene near the junction point (Supplementary Figure 1). Remarkably, we observed that the sequence belonging to the fusion point was a polyA tail upstream of a 3´UTR-like sequence.

We confirmed this finding by analyzing *in-silico* the fused regions predicted by the RNA-seq data. Indeed, as we had the assemblies of the fusions from the reference transcripts, we extracted the transcript annotations and found that a region from the 3´ UTR of the promiscuous transcript was fused with the *HSP90AB1* transcript partner. The inspection of such regions through PCR and Sanger sequencing was challenging because they contained low-complexity sequences. In fact, these polyA tail sequences were not sequenced through Illumina, and they were absent in the RNA dataset, but only detected by Sanger sequencing using the Nested primers.

The 5´ *MALAT1*– *CD24* 3´ fusion, with ID MALAT1_XI, was validated by sequencing the *CD24* transcript with the forward primer (designed on the *MALAT1* transcript) and *MALAT1* with the reverse (created on the *CD24* transcript). As observed in the HSP90AB1_III fusion, the obtained sequence of the fusion included a region of the transcript from the *CD24* gene not previously

predicted through the RNA-seq reads. Likewise, the *CD24* gene was fused at its 3´ UTR with *MALAT1*.

Overall, these *in-silico* and experimental validations confirmed the fusions with promiscuous transcripts. These findings suggest a more complex event of fusions: First, there were multiple fusions between a given pair of transcripts; Second, the 3´ UTR of the promiscuous partner was present at the junction point of the fusions; Third, Sanger sequencing of the fusions revealed more extended fused regions not predicted by the RNA-seq approach.

4.2.2.5 Characterization of fusion promiscuous transcripts in metastatic breast cancer

A possible event of massive fusion transcripts in metastatic breast cancer, not previously described, emerged from the evidence described above. We began the characterization of this novel event by studying the recurrence of the promiscuous partner across tissues. We found 44 promiscuous transcripts across the 18 metastatic samples, which varied across samples and even across tissues (Figure 42). The most recurrent promiscuous transcripts across the different tissues were from the *EEF1A1*, *CD24* and *MALAT1* genes.

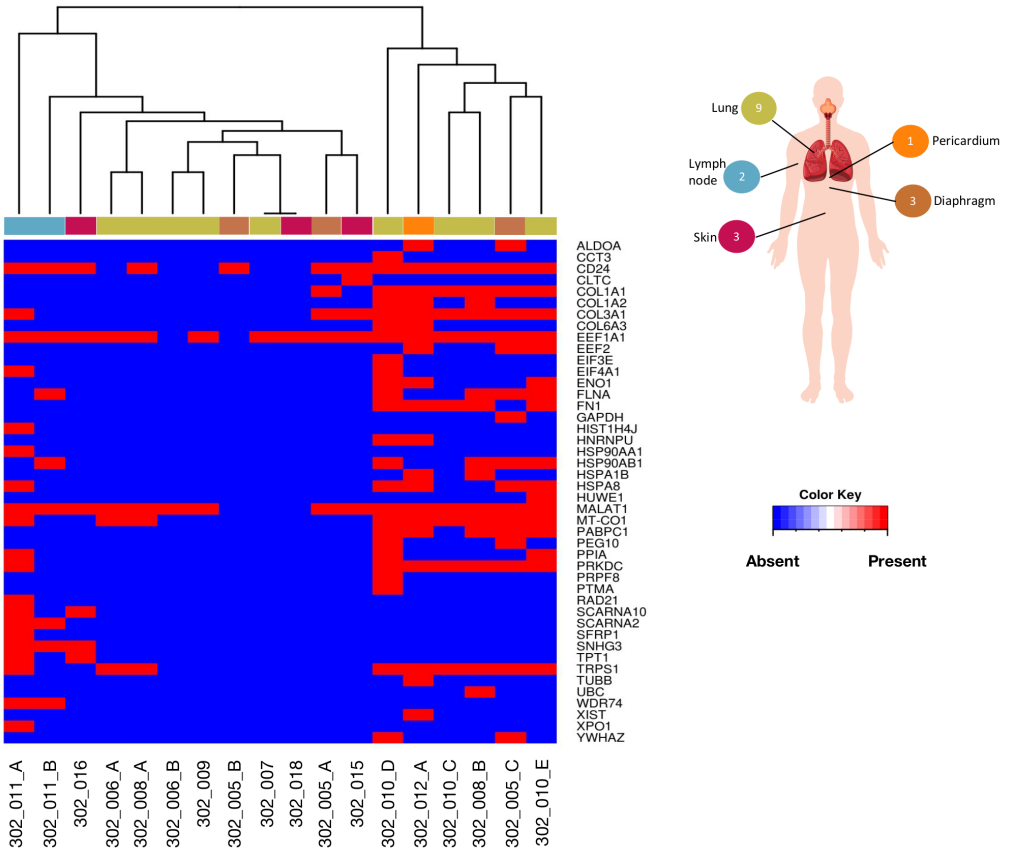


Figure 42. The 18 metastatic samples from patient 302 contained different promiscuous transcripts. Heat map of presence (in red) and absence (in blue) of the 44 promiscuous transcripts found across the entire PE dataset. The top color bar indicates the sample tissue using the color key in the human figure to the right. Samples clustered according to promiscuous similarity profile.

Likewise, the level of promiscuity of the transcript partners changed across samples. For instance, *CD24* and *MALAT1* were promiscuous genes present in most samples but had a different number of partners in each metastasis (Figure 43).

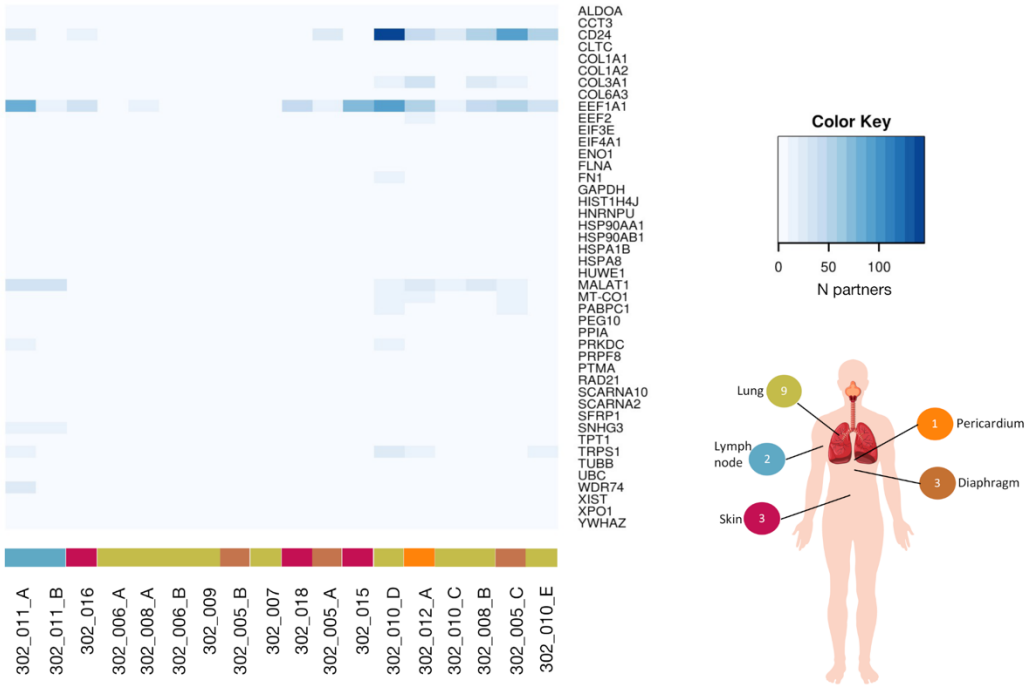


Figure 43. The number of partners per promiscuous transcript varied across metastases of case 302. Heat map of the number of partners that were found for a given promiscuous transcript in each sample.

To determine whether the similarities between samples that we observed at the gene expression levels, were also found at the transcript fusion level, we compared the sample clusters according to gene expression, with the clusters from the promiscuous transcript profiles. As mentioned in section 4.2.1, the metastatic samples from tissues closed together, like lung, pericardium and diaphragm, had similar gene expression profiles (Figure 32). However, based on the promiscuous transcript profiles, samples from metastases in the skin also shared promiscuous transcripts with such tissues (Figure 42). These results suggest that the promiscuous transcript profiles do not recapitulate the expression level profiles and, it seems they are independent.

Given these findings, we wondered whether the promiscuous transcripts belonged to a particular expression profile according to the expression level.

Based on the cutoffs defined by the Expression Atlas[226], used to determine if a given transcript has low (0.5 to 10 TPM), medium (11 to 1000 TPM) or high (> 1000 TPM) expression, we determined the expression levels of the transcripts from our dataset, according to the TPMs calculated as described in section 3.2.2. We evaluated a possible scenario in which the promiscuous transcripts belonged to highly expressed genes in the metastatic samples, and thus, they would have more chance to be fused with other transcripts. Thus, we estimated the promiscuous transcripts that were highly expressed in four metastatic samples: 302_010_D from the lung, 302_012_A from the pericardium, 302_011_A from the lymph nodes, and 302_005_C from the diaphragm. We did not include metastases in the skin, given that these samples did not have a significant number of promiscuous transcripts to perform statistical analyses. We found that the proportion of highly expressed promiscuous transcripts was very low in all analyzed samples (Table 5).

Table 5. The promiscuous transcripts were not highly expressed across tissues. The proportion of highly expressed promiscuous transcripts was estimated by counting the ones with TPM values above 1000 per sample. The proportion of highly expressed transcripts was calculated by assessing the number of highly expressed (more than 1000 TPM) from the total expressed transcripts in the sample (cutoff: 0.5 TPM). The *p*-value was obtained through a *t*-test between the two proportions.

Metastasis Tissue	Sample ID	Total Pt	Highly expressed Pt	Proportion of highly expressed Pt	Proportion of highly expressed transcripts	*p*-value
LUNG	302_010_D	25	1	0.04	0.0075	0.47
PERICARDIUM	302_012_A	20	0	0	0.0055	1
LYMPH NODES	302_011_A	20	1	0.05	0.0074	0.36
DIAPHRAGM	302_005_C	17	1	0.06	0.0067	0.25

Pt: promiscuous transcripts.

Then, we performed a test for proportions for each of these samples, comparing the proportion of highly expressed promiscuous transcripts with the proportion of the highly expressed transcripts in the sample. The *p*-values were above 0.05

in all the samples (Table 5), indicating that the number of promiscuous transcripts highly expressed was similar to the estimated numbers of highly expressed transcripts in the tissues. In fact, we found that the promiscuous transcripts had mainly medium expression levels in the analyzed samples, with TPM values ranging from 55.55 to 327.62.

In summary, the promiscuous transcripts were not the most abundant in the samples. Instead, they had medium expression levels. These findings suggest that the expression levels of the promiscuous transcripts were not related to the massive fusion pattern.

4.2.2.5.1. Searching for the origin of the promiscuous fusions at DNA level

Further, we wondered about the possible mutational process behind the generation of fusions with promiscuous transcripts, which we called promiscuous fusions. Although our findings pointed out that the promiscuous fusions might arise at the transcriptomic level, we investigated whether the promiscuous fusions could arise from complex chromosomal rearrangements encompassing multiple SVs, that would explain the multi-fusion partners of the promiscuous transcripts.

First, we inspected the genomic locations of the genes corresponding to the pair of fused transcripts, given that it could illustrate the possible type of chromosomal rearrangements triggering the promiscuous fusions. We studied the sample 302_010_D, as shown in Figure 44, where we found that the promiscuous transcripts belonged to genes that were localized in distinct chromosomes respected to their fusion partners.

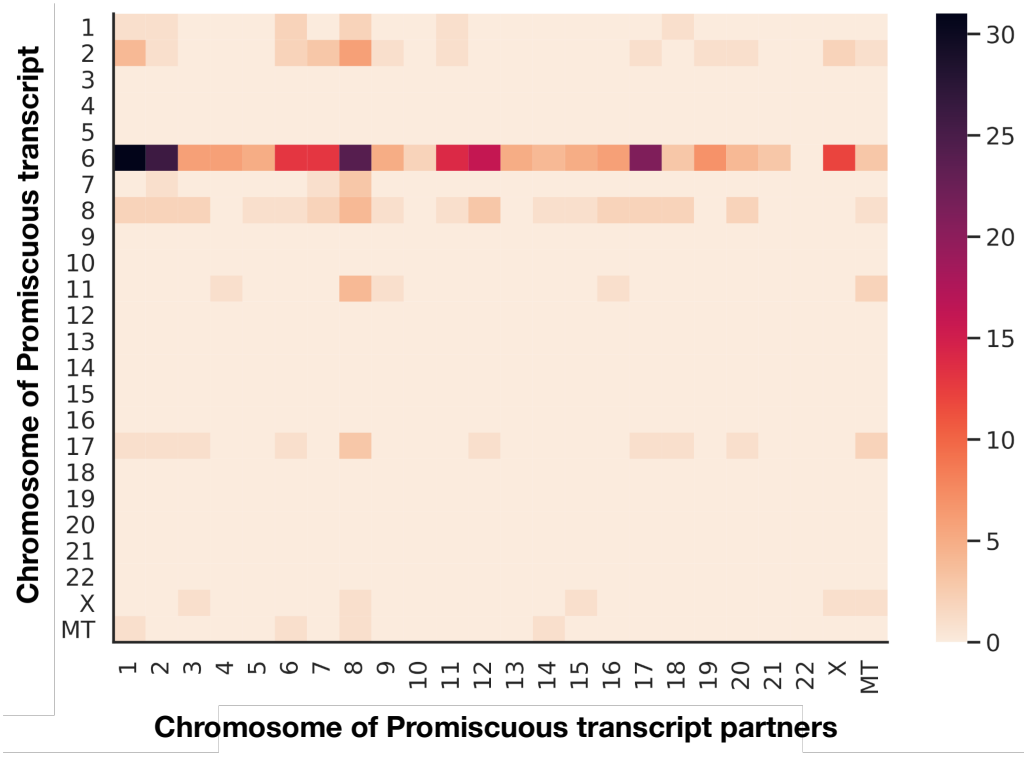


Figure 44. Heat map of the chromosomes where the genes, corresponding to the pair of fused partners, were located. The color key indicates the number of interactions between the chromosomes according to number of promiscuous fusions predicted.

Thus, under the possible scenario where the promiscuous fusions arose at the DNA level, this result would suggest a possible event of multiple copies of different regions of the promiscuous transcripts and further translocations. Then, we further performed a preliminary analysis detecting structural variants on the available WES data from the 302_010_D sample. We did not find any SV that could explain the promiscuous fusions detected through RNA-seq. However, since we analyzed the exome due to WGS data were not available, the SV detection we performed did not reveal possible SVs occurring in intronic regions associated with the fusions.

4.2.2.5.2. Understanding the role of promiscuous fusions

To better understand the role of the promiscuous transcripts in tumor development, we performed an enrichment analysis that revealed they were related to mammary neoplasm pathways (Figure 45). These results might suggest a relationship between the promiscuous event and metastasis process in breast cancer patients.

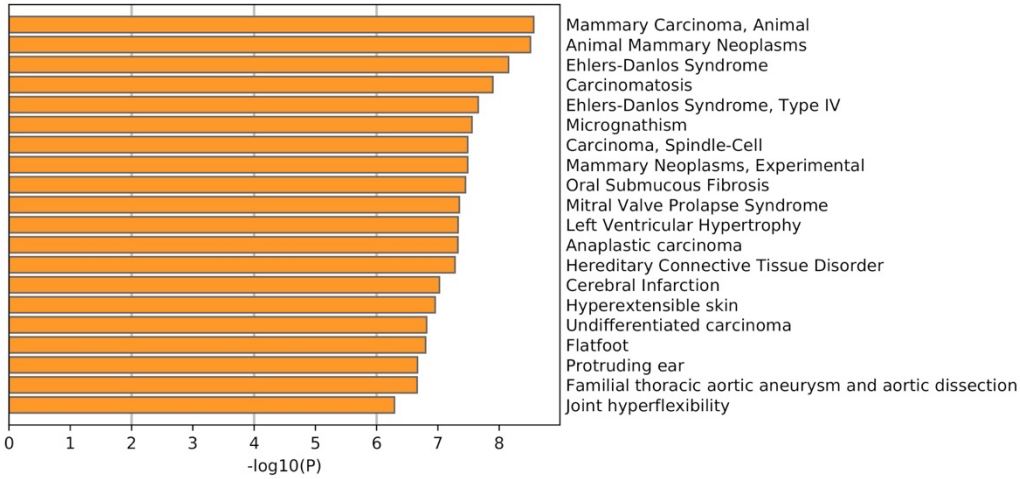


Figure 45. The promiscuous genes were enriched in pathways related to the primary tumor. To determine the possible role of the promiscuous transcripts, we performed a gene enrichment analysis from the promiscuous transcripts found in all metastatic samples from case 302. This graph was obtained using Metascape[227].

Next, we investigated the gene networks that arose from the promiscuous fusions to determine their interconnection. Interestingly, although the promiscuous transcripts did not share fused partners, we found several promiscuous transcripts fused between them (Figure 46). Furthermore, the partners of a given promiscuous transcript were not fused between them.

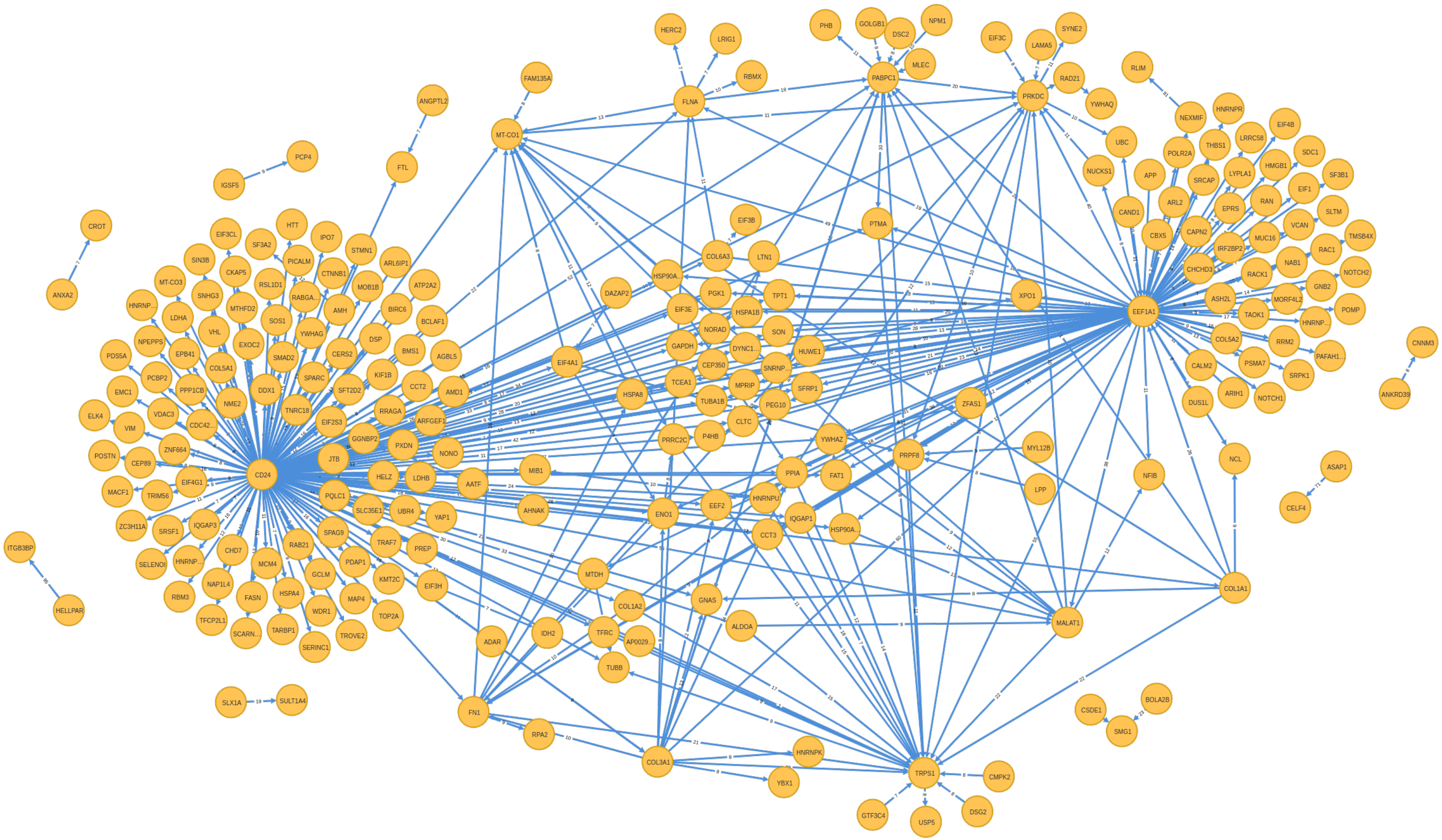


Figure 46. Transcript fusion networks in a single sample of metastatic breast cancer. The network graph from the 302_010_D sample showed the fusions (edges) between the genes displayed as nodes. It was built by using Neo4j®.

Since the partners of the promiscuous transcripts did not seem to interact, we wondered whether they shared any feature that made them to be fused with the promiscuous transcripts. Thus, we analyzed the relationship between the partners of a given promiscuous transcript in terms of the similarity between their sequences (BLAST, E-value <= 10^{-3}). We calculated the proportion of similar partners per promiscuous transcript for every sample. Interestingly, the percentage of similar partners of a given promiscuous transcript was below 20, suggesting that the partners did not belong to the same gene family (Figure 47).

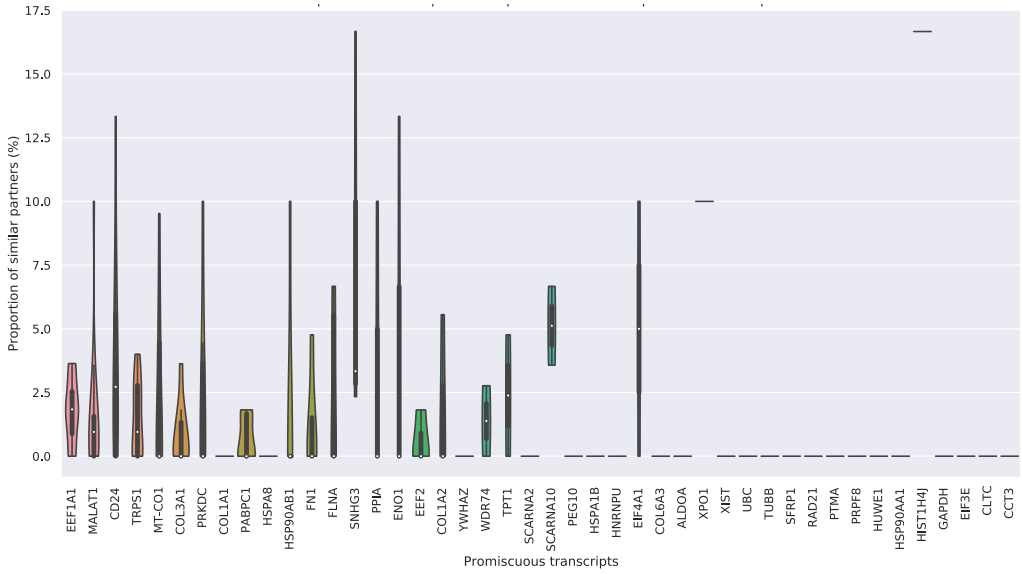


Figure 47. Proportion of similar (E-value <= 10^{-3}) partners per promiscuous transcript. The violin plot displays the proportion of similar partners of the 44 promiscuous transcripts found in the 18 metastatic samples from case 302. The low percentage of similar partners of a given promiscuous transcript suggested that they were not related.

Taken together, these results suggest that the impact of the fusions with promiscuous transcripts possibly lie in the alteration of the function of the promiscuous genes, which in turn might influence tumorigenesis, as they are related to breast tumor development.

4.2.2.6 Beyond metastatic samples: the case of promiscuous transcripts found in healthy RNA samples

Since the analysis of the SE dataset from metastases of nine patients set the evidence of fusions with promiscuous transcripts, and given the complex scenario found in the metastases of case 302, we wondered whether promiscuous fusion transcripts could be a hallmark of metastatic breast cancer. To address this question, the direct approach would be to analyze other datasets from metastatic breast cancer. Nevertheless, as additional samples were not available, we decided to survey the presence of fusions with promiscuous partners in samples from healthy breast tissue (normal samples). Considering that there was no matching normal RNA-seq data for our cancer samples, we analyzed six RNA-seq samples from healthy breast tissue from breast cancer patients (available PCAWG data); this dataset included PE reads of 50 bp.

First, applying the same strategy to detect fusion transcripts as in the metastatic samples from case 302 (PE dataset), we couldn't find fusion transcripts with promiscuous partners in any of the normal breast samples, since we did not evidence any split reads supporting promiscuous fusions. However, the RNA-seq data from these samples had less reads than the metastatic samples analyzed here (Supplementary Table 3). Therefore, if fusions with promiscuous transcripts were present, the chances of detecting them could be reduced. Additionally, knowing the challenges to detect fusions through supporting reads of 50 bp, we decided to apply low-astringent parameters to increase sensitivity for the prediction of fusion transcripts and then, evaluate the presence of promiscuous transcripts.

Strikingly, we found a high number of fusions with promiscuous transcripts in all six normal samples by using evidence only from supporting PE reads (a minimum of 3 supporting PE reads per fusion transcript). The number of

promiscuous transcripts per sample ranged between 22 to 590 (Table 6), of which 28 (out of the 44 promiscuous transcripts detected in the samples from patient 302) were also found as promiscuous transcripts in the 18 metastatic samples from patient 302. Nevertheless, in terms of the percentage of promiscuous fusions, the metastasis samples harbored more promiscuous fusions (median, 73%) than the normal samples (median of 61%).

Table 6. Fusions with promiscuous transcripts in samples from healthy breast tissue. We assessed the detection of fusion transcripts through the supporting PE reads.

PCAWG Normal Sample	Total N of TFs	N of promiscuous fusions	% of promiscuous fusions	N of promiscuous transcripts
1	2,459	1,656	67,34	189
2	834	359	43,05	22
3	2,822	2,022	71,65	191
4	934	362	38,76	38
5	5,859	3,990	68,10	590
6	1,179	648	54,96	52

N: number. TFs: Transcript fusions.

Given these outcomes, we considered the hypothesis that the phenomenon of promiscuous fusions does occur at a basal level in normal cells and is amplified in metastasis. Since our findings from metastatic samples from patient 302 were obtained through the evidence of both supporting split and PE reads, we reanalyzed the RNA-seq dataset from case 302 to make results comparable those from normal samples. We thus predicted fusion transcripts by detecting supporting PE reads only. Additionally, we performed a downsampling of reads in each of the 18 metastases, to reproduce the same experimental conditions as in the 6 normal samples from PCAWG (Supplementary Table 3).

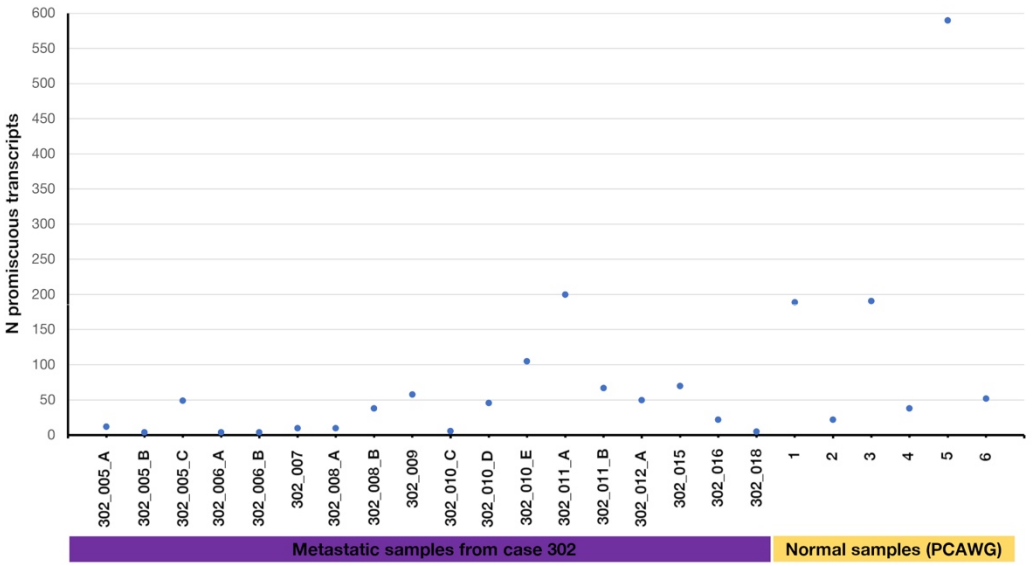


Figure 48. Number of promiscuous transcripts in downsampled metastatic samples from patient 302 (as indicated with the purple bottom bar) and normal breast samples from the PCAWG dataset (indicated with the yellow bottom bar).

As shown in Figure 48, the normal and downsampled metastatic samples had similar number of promiscuous transcripts. Surprisingly, a sample from healthy breast tissue had the highest number of promiscuous transcripts. In this sample, the majority of the promiscuous transcripts found, however, were genes encoding Immunoglobulins. We further evaluated the enriched pathways from these promiscuous genes predicted in this normal sample that revelated neoplasia-associated pathways (Figure 49).

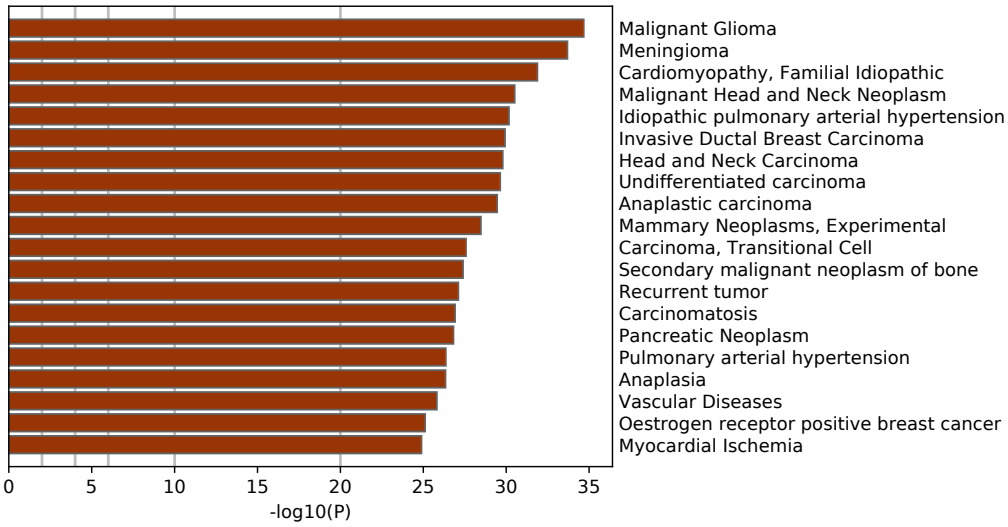


Figure 49. The promiscuous genes detected in normal sample 5 were enriched in neoplasia-associated pathways.

Thus, we wondered whether the promiscuous fused transcript phenomenon found in healthy breast tissue was the same as in metastases. First, we evaluated the gene networks originated from the promiscuous fusions, observing the promiscuous genes as hubs, connected to all their partners, which in turn were not interconnected (Supplementary Figure 2). Next, we analyzed the sequence similarity of the partners that fused with the promiscuous transcripts. We found that the promiscuous transcript partners belonged to non-related genes, as we found in metastases (Supplementary Figure 3).

In summary, we found that the prevalence of the promiscuous phenomenon was much higher than previously appreciated. This event was beyond the metastasis: it was detected in healthy breast cells.

5 DISCUSSION

5.1 Classification and characterization of complex chromosomal rearrangements in cancer

We have characterized complex patterns of structural variation across 2,586 primary tumors of diverse histological origins. By assessing the hallmarks of chromosomal rearrangements encompassing 3 SVs, we identified the trisomy pattern as a common process across different cancer types.

The differential trisomy pattern complexity between the samples across the tumors, and within the samples of a given cancer, might be associated with the clonal status of variants in the cohort. Indeed, although there are shared features and driver mechanisms within tumors, i.e., tandem duplications in breast and ovarian cancers[97,228], there is high intrinsic variability of alterations across the tumoral genomes of patients with a given tumor[229].

Furthermore, such prevalence and heterogeneity across diverse cancers had been observed analyzing other complex rearrangements such as chromothripsis[102]. Interestingly, in agreement with our results, Cortés-Ciriano *et al*. reported SARC (Soft Tissue Liposarcoma), BOCA (Bone Osteosarcoma) and LUSC (Lung Adenocarcinoma) as tumor types with high percentages of samples with the complex rearrangement, and LAML without chromothripsis events. These results suggest a likely co-occurrence of complex events, as reported in the study by Cortés-Ciriano *et al*.

Although the higher frequencies of the trisomy pattern were not in samples with neuronal-related cancers, the pattern affected genes related to neuronal pathways in multiple instances. Especially, using all the approaches to determine the likely functional impact of the pattern, we evidenced that the glutamate receptor signaling pathway was altered. The glutamate was

suggested as a potential growth factor in tumor development. Different studies have reported that glioma cells release higher levels of glutamate (Glu) that allow them to grow more aggressively, which might arise from glutamate uptake/release systems aberrantly expressed and/or activated in these cells[230]. The cells from non-central nervous system cancers may have Glu receptors and secrete Glu as well, such as human breast, mouse melanoma and rat prostate cancer cell lines[231]. In other types of cancer, the Glu has been correlated with cancer aggressiveness stimulating cancer cell proliferation, like in pancreatic cancer[232] and lung cancer[233,234]. This might suggest that the trisomy pattern could arise late in tumor progression. This complex rearrangement likely needs previous alterations that further trigger its emergence. We analyzed here samples mainly from primary tumors. However, the clonal status within the tumor of these complex chromosomal rearrangements should be assessed in order to provide more insights into the association with oncogenesis and progression.

Although the number of chromosomes, the number of SVs, and the closeness of the SVs were unifying features of trisomy-associated alterations, multiple mechanisms may account for the trisomy pattern. We have been able to detect a new pattern of complex chromosomal rearrangements that were previously undescribed. The extensive study from the PCAWG consortium confirmed our findings, describing a new Replication-based mechanism of rearrangement[225]. Our results regarding different genomic conformations align with the observation that there is a high complexity in the landscape of chromosomal rearrangements, which remains under-discovered.

5.1.1 Uncovered complex rearrangements across different cancer types

Despite these efforts towards the classification of structural variants (SVs), the characterization of the complete landscape of chromosomal rearrangements, the mechanisms by which all of them arise, and their relationship with tumorigenesis remain to be discovered.

Our outcomes regarding the trisomy pattern generated a more nuanced set of criteria for complex chromosomal rearrangements, paving the way to assess a new statistical strategy to overcome the challenges in identifying complex patterns of structural variants in cancer. Applying the KDE-based clustering method, we isolated SVs as non-independent events from the SVs that arise over time by genomic instability. The KDE clustering has been proven to be fast and simple and very suitable for distribution based-clustering tasks without setting a priori number of clusters[235,236]. Facing the lack of reference complex patterns of SVs to compare with, we presented a statistical approximation to prove that SVs clusters were not by chance, indicating that they must be related to each other[237,238].

The adaptation of the graph-mining strategy to search for patterns of SVs suggests a high performance compared to previous classification approaches, given that we found known and novel classes of structural variants. The selected measure for analyzing the significance of the patterns, the Abundance, is directly related to the *z*-score of the pattern but normalized, allowing us to compare among different patterns. The use of such graph techniques, combined with statistical analyses, looks like a promising alternative to survey complex rearrangements in cancer genomes. In fact, a recent study implementing graph techniques to infer and classify SV patterns, described novel complex rearrangements in cancer genome graphs[239]. We also added the

new pattern of SVs, chromotrikona, to the shortlist of known complex rearrangements.

Chromotrikona encompasses three SVs that involves balanced inversions between distinct DNA regions in two or three chromosomes. Its genomic configuration regarding different host chromosomes, suggests a mechanism of cut-and-paste, in which some genomic regions are translocated to distant DNA regions without following the 5'-3' orientation. The presence of other co-localized rearrangements in the vicinity of complex rearrangements showcased the high complexity of the catastrophic events, as reported by Cortés-Ciriano *et al*.[102]. In fact, Zhang *et al*. found that shattering and reassembly mechanism and replication error models, are not mutually exclusive and could co-occur[240]. This could explain our findings regarding the process of cell replication, in which we found these kinds of 3-SV patterns. Further analyses will provide more insights into the possible mechanisms by which different categories of complex patterns may arise, their co-occurrence and, their overall functional implications in tumor development.

This study showed that continuous development of methods for studying complex patterns of SVs is needed. All efforts from the recent studies, including this thesis, represent a significant step towards understanding the role of complex rearrangements in carcinogenesis. However, it is very likely we are still missing essential parts of the landscape of mutational processes, due to the experimental and computational limitations that currently exist. This study opened the field to further discoveries of more complex SV patterns, and as such, it represents an important step towards a better understanding of the catastrophic genomic events that lead to cancer development.

Given the ubiquitous of the complex rearrangements, it is feasible that different mutational processes together reshape the human cell genomes, generating a variety of possible structures, which might provide a selective advantage on

tumorigenesis. The detection and comprehensive characterization of these events will constitute the basis for potential diagnostic biomarkers and therapy targets.

5.2 Identification and characterization of mRNA alterations in metastatic breast cancer

With the main goal of identifying transcriptomic alterations associated with metastasis, we analyzed RNA-seq data from 82 samples from 10 patients with lethal breast cancer, through two different approaches. First, we conducted a preliminary analysis of the transcriptional profiles in metastases from different tissues and second, we performed a comprehensive characterization of fusion transcripts present in metastatic samples.

5.2.1 Transcriptional profiles of metastatic breast tumors

We explored the gene expression profiles in metastases from patients with lethal breast cancer, aiming to identify transcriptional programs that characterize the different metastases. The transcriptional patterns were more similar across metastatic sites from the same individual, than between metastases in the same tissue from different patients, indicating a core program of gene expression, probably due to the tumor burden shared between metastases from the same patient. This observation suggests that the molecular program of the primary tumor is retained in its metastases, as reported in other studies on breast cancer[132,241]. Conversely, in other metastatic tumors, such as ovarian carcinoma, gene expression profiles of metastases differ from those of the primary tumors[242].

However, when analyzing an individual patient, such as the case 302, the transcriptional profiles from metastases to the same organ were more similar, indicating differences in the programs of gene expression among tissues. This transcriptional heterogeneity among metastases from different tissues could be

explained by their different origins, where the metastatic tumors in each tissue probably evolve from distinct clones.

Additionally, metastases in lung, diaphragm, and pericardium displayed common gene expression programs. This might illustrate that these metastases occurred in similar moments, probably originated from the same clone due to the closeness of the metastatic sites. Furthermore, an additional explanation for these results lies on the possible adjusted metastatic niche that cancer cells can promote by remodeling the extracellular matrix (ECM), according to the nutrient accessibility and metabolic reactions in the invaded tissue[243]. For instance, Elia *et al.* reported that metastatic breast cancer cells metabolize pyruvate, which is abundant in the lungs, to drive collagen-based ECM remodeling in the lung metastatic niche[244]. Thus, cancer cells might undergo a tissue-specific transcriptional reprograming to promote their own metastatic growth, which converges towards a common transcriptomic landscape of tumors in a given tissue.

Given these findings, we aimed to address downstream analyses to dissect the expression patterns and identify the transcriptional signatures of metastases within an individual patient, as well as within a given tissue. In the beginning, we addressed the transcriptional profiling and the identification of fusion transcripts in parallel. However, as the characterization of fusion transcripts showed the most novel and relevant results, we decided just to continue with the analysis of fusion transcripts.

5.2.2 The promiscuous fusions: A new pattern of multi-fusion partner transcripts in breast cells

With the aim of characterizing the transcriptome of breast cancer metastasis through the identification of fusion transcripts, we first developed a guided

framework that allowed us to control and select filters affecting the specificity and the sensibility. We implemented a double approach to identify the reads that infer fusions among transcripts from two different genes, including the extraction of reads that incorrectly map to the human genome, and then, the alignment of those to the reference transcriptome using BLAST. There are two main advantages in alignment to the transcriptome rather than genome: (i) the complexity of splice site alignment is avoided due to the transcriptome only includes mRNA sequence, and (ii) the reference transcriptome allows more accurate alignment algorithms to be used, such as BLAST[245,246].

Furthermore, the filters applied to evaluate possible false positives led us to discover new insights of the fusion transcript events in breast cancer. As we analyzed datasets that differed in terms of read length and sequencing approach (SE and PE), we implemented rigorous filter steps and criteria needed to end up with the most likely fusion transcript candidates in every case. For instance, short read sequences have lower alignment specificity, particularly in cases such as SNPs, sequencing errors and repeat regions. In that sense, incorrect alignments might lead to false predictions that could be overcome by looking for more supporting information, such as read pairs which evidence the fusion[195].

Our results demonstrate the existence of massive transcript fusion events in metastasis. In this scenario, not only promiscuous transcripts were found to be fused to several others, but also the promiscuous transcripts could be fused with each partner at different points. The high complexity of this pattern limited the RT-PCR validation of the multiple isoforms of the fusions that involve promiscuous transcripts, but even so, we validated different fusion isoforms and various fusions of the gene transcript *CD24*.

Some well-known genes have been shown to be promiscuous, such as *MLL* in leukemias[247,248], *EWS* in sarcomas[249] and *TMPRRS2* and *ETV1* in prostate

cancer[173,250]. More recently, a study from the PCAWG consortium reported promiscuous partners through different cancer types[210]. *MALAT1*, one of the recurrent promiscuous transcripts found in the present study, has been detected fused with several genes in breast cancer[251] and ovarian cancer[252] samples. However, in all above-mentioned studies, promiscuity refers to the diversity of partners that a given fused gene has between patients within a given cancer type or between different cancer types. In this study we found promiscuous transcripts within the same sample and in different isoform fusions (different fusion points between the two transcripts).

Besides, given the heterogeneity of the promiscuous transcripts across metastases and even within each tissue, we hypothesize that this novel multi-fusion event occurs at the RNA level. Conversely, the hypothetical scenario in which chromosomal rearrangements trigger the promiscuous fusions would imply a mechanism of multiple copies of the promiscuous genes and then juxtaposed with the fusion gene partners. Although we cannot exclude this possibility due to we did not have WGS available to analyze, plausible RNA molecular mechanism(s) could parsimoniously account for the multiple fusions and at multiple loci of a given transcript.

For instance, hypothetical promiscuous fusions forming at DNA, hardly explain cases such as the *CD24* gene transcript fused with 145 partners in the sample 302_010_D from lung metastasis. Furthermore, the levels of expression of the promiscuous transcripts did not support the hypothesis that the promiscuous fusions arise from genomic rearrangements, since the promiscuous transcripts should be present in high abundance, which was not the case. Therefore, our findings suggest that the origin of the promiscuous fusions may be associated with non-canonical transcription factories where aberrant splicing processes occur massively.

The presence of promiscuous fusions in the samples from healthy breast tissues raises several questions concerning their role in breast cancer and the mechanism of their formation. One possible explanation would be that the promiscuous fusions are technically created fusion artefacts caused by the template switching capabilities of reverse transcriptase in both RNA-seq library preparation and RT-PCR, where reverse transcriptase (RT) was used for conversion of mRNA into double stranded cDNAs[253,254]. With regard to apparent non-canonical trans-splicing generated by RT, it has been reported that those splicing events occur between non-canonical splice sites, often share short homologous sequences, e.g., repeats; it is though that the RT requires that homology to switch templates during primer extension, leading chimeric artifacts[255]. However, after manual inspection, we did not find evidence of homologous sequences between the promiscuous transcripts and their partners. In fact, if promiscuous transcripts were artifacts from RT, they should be present in all the samples with a similar distribution, showing the homologous regions of the promiscuous transcripts in which RT would switch templates. Instead, we found heterogeneity in both, the promiscuous transcripts, and fusions across metastatic and normal samples. Additionally, the template switching usually implicates intramolecular splicing events (within a same RNA molecule), whereas there is little evidence for its involvement in intermolecular trans-splicing (between two different transcripts)[256].

In contrast, another explanation could be that the promiscuous fusions are tissue-specific, with this multi-fusion event occurring at basal levels in normal breast cells, while exacerbating in tumoral cells. This hypothesis was not fully tested in this study, since paired samples were not available (normal and metastatic samples from a same individual). We rather compared different patients and different datasets in terms of read length and number of reads, which does not allow us to set a definitive conclusion.

It is worth noting that complex events in human transcriptomes were found in healthy tissues. Djebali *et al.* reported chimeric transcripts present in human cells from different lines/tissues, that tend to form "cliques" of different sizes, where sets of genes were all pairwise connected[257,258]. However, the chimeric events observed in that study differ in several features from the promiscuous transcript fusions described in our study. Regarding the transcript networks formed by the fusions, in our study we found that promiscuous transcripts were fused one-to-one to several others, whereas Djebali *et al.* showed that the transcripts were fused all among them, creating the so-called transcriptional network hubs. Furthermore, in that study, the gene-to-gene interactions were not random, observing that the genes involved in the hubs were phylogenetically related, which suggest a functional role of the chimeras as an RNA network. Conversely, we found that the partners of the promiscuous transcripts were from unrelated genes. In fact, the promiscuous involvement of one transcript in such multiple fusions might provide evidence of the contribution of this transcript to the functionality of breast cells.

Overall, the presence of promiscuous fusions as molecular events present in breast tissue was evidenced, although their potential biological function in cancer is unclear. There are still several questions that must be addressed, in order to elucidate the importance and the role of these events in cancer development. Although, these promiscuous fusions might represent stochastic events in vivo, with little or no impact on cellular functions, we speculate that these fusion events could occur at basal levels in breast cells and that, under particular genetic, epigenetic and environmental conditions, could be exacerbated, promoting tumor progression.

Further analysis of paired samples, where RNA-seq data from healthy breast tissue, primary tumor, and metastases from an individual patient can be analyzed, will provide new insights into the role of the multi-fusion event that could be applied in breast cancer prognosis and clinical management.

6 CONCLUSIONS

Classification and characterization of complex chromosomal rearrangements in cancer

I. The characterization of the features that define complex chromosomal rearrangements in cancer enabled to assess a novel statistical strategy to overcome the challenges of identifying recurrent patterns of structural variants and distinguish them from random in large cohorts, such as PCAWG dataset.

II. We identified a novel recurrent pattern of somatic chromosomal rearrangements in the cancer genome. We named this pattern *Chromotrikona*, which consists in the cooccurrence of reciprocal translocations between different chromosomes.

Identification and characterization of mRNA alterations in metastatic breast cancer

I. We have designed and implemented a strategy for the identification of fusion transcripts from RNA-seq data, that overcomes the lack of sensitivity of current available programs, by allowing us to control and tune the different steps and filters involved in the process.

II. The application of this strategy to different metastatic breast cancer transcriptomes evidenced the presence of a novel pattern of fusion transcripts, whereby in each sample, some transcripts were found to be fused with several distinct transcript partners, independently.

III. The promiscuous fusions seem to be tissue-specific, being present in both, healthy and tumoral cells. Further analyses comprising healthy, primary and metastases samples from an individual patient are needed to pinpoint the role of these molecular events in tumor progression.

7 SUPPLEMENTARY INFORMATION

Supplementary Table 1. Primer sequences for validation of thirteen transcript fusions. Different transcript fusion forms were targeted to validate three transcript fusions predicted in a lung sample from case 302.

Fusion	ID	Primer name	Sequence 5'-3'	Expected Amplicon length (bp)	5´Gene Transcript	Coord 5´Ref Gene Transcript	3´Gene Transcript	Coord 3´Ref Gene Transcript
CD24 - PEG10	IV	PEG10_IV_F	CAGATTTATTCCAGTGAAACAACAAC	382	CD24	464 - 660	PEG10	5500 - 5741
		PEG10_IV_R	CAAGTTCACTGTATTACTTCACGA					
	V	PEG10_V_F	CACATTTGGCTGTTTACTAAAGC	722	PEG10	4981 - 5315	CD24	1556 - 2059
		PEG10_V_R	GGCTATTCTGATCCATAGTTGTTT					
CD24 - HSP90AB1	III	HSP90AB1_III_F	GAAAATGTTGAGAATCCCAAATTTGA	841	CD24	819 - 1368	HSP90AB1	270 - 559
		HSP90AB1_III_R	GAATAAAAGCCAACACCAAACT					
	III	Nested_HSP90AB1_III_F	GCTTGAGAAATATGGACACTTAATACT	400	CD24	974-1046	HSP90AB1	227 - 158
		Nested_HSP90AB1_III_R	ACATGAGTTGGGCAATTTCT					
	IV	HSP90AB1_IV_F	GCCAAGTCTGGTACTAAAGC	807	HSP90AB1	495 - 710	CD24	1765 - 2366
		HSP90AB1_IV_R	AGTAGCTTCAAAACTGTTCGATC					
	V	HSP90AB1_V_F	CTATTTATTCCTCGTCGGGC	703	HSP90AB1	1153 - 1659	CD24	1930 - 2202
		HSP90AB1_V_R	GCAGAATCAAGCCCACTTTTA					
	VI	HSP90AB1_VI_F	CACTTTTCTGTAGAAGGTCAGTTG	533	HSP90AB1	1083 - 1453	CD24	754 - 1016
		HSP90AB1_VI_R	GAAATCATGTCTTAACTATTTTGGATGTT					
	VII	HSP90AB1_VII_F	GGCAAAATTGCAAATCTTGAAATTAAG	590	CD24	1884 - 2378	HSP90AB1	1685 - 1846
		HSP90AB1_VII_R	CTCTTCCCATCAAATTCCTTGAG					
		MALAT1_III_F	GTGAAGCTAGGAAAAAGGATTCC					

CD24 - MALAT1	III	MALAT1_III_R	TGAGTCTCTTAAGAGTAGAGATGC	929	MALAT1	2824 - 3623	CD24	551 - 718
	IV	MALAT1_IV_F	GGAAGGAAAGTATTGAACTGGG	822	MALAT1	4406 - 5024	CD24	1805 - 2057
		MALAT1_IV_R	AGAGTATAAAAGTTTGTGAATTTAATGCAAA					
	VIII	MALAT1_VIII_F	TATTTGTGATTGAAGCTGAGTACATT	460	MALAT1	6763 - 6914	CD24	562 - 899
		MALAT1_VIII_R	TTAATATTGGCATCCATCATCTAGTC					
	IX	MALAT1_IX_F	TGGTATTCTTCAGACTATAGAAGGAG	682	MALAT1	7140 - 7603	CD24	635 - 906
		MALAT1_IX_R	AGCAGATTTAATATTGGCATCCAT					
	X	MALAT1_X_F	CTTTCAGATGGTATTCTTCAGACT	761	MALAT1	7140 - 7603	CD24	564 - 906
		MALAT1_X_R	AGCAGATTTAATATTGGCATCCAT					
	XI	MALAT1_XI_F	TAACATTTAAGCAAGCTGTTTTTATAGC	458	MALAT1	7632 - 7837	CD24	2075 - 2327
		MALAT1_XI_R	TGTTTGTTCCCATGTAGTTTTCTAA					

*** Expected Amplicon Length**

† The four reference isoforms IDs from the genes *CD24, PEG10, HSP90AB1* and *MALAT1* were NM_013230.3, NM_001184962.1, NM_001271970.1 and ENST00000534336.1, respectively.

Supplementary Table 2. Classification features of the 3 SVs involved in trisomy patterns according to copy number associated. We distinguished 4 different types of trisomy patterns: chromoplexy (in red), cycles of templated insertions (in purple), complex rearrangement I (in green), complex rearrangement II (in blue).

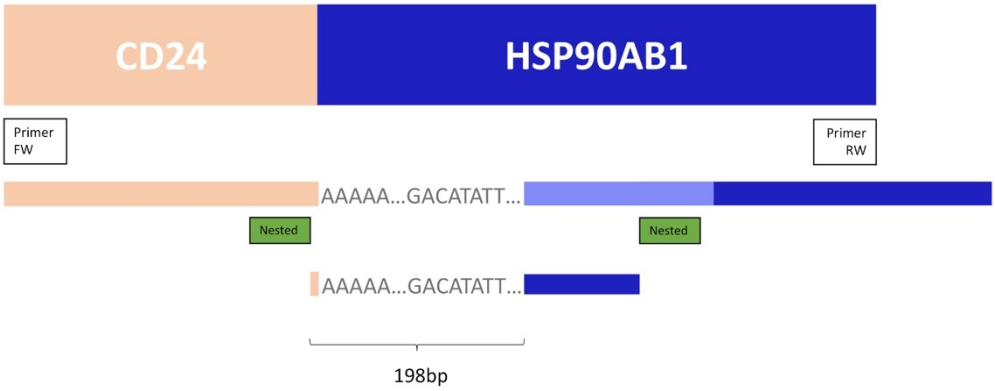
Intrachrom*	INV/NOT**	SV1	SV2	SV3	N patterns	N samples
DEL	INV	DEL	DEL	TRA	1	1
DEL	INV	DEL	TRA	TRA	4	4
DEL	NOT_INV	DEL	TRA	TRA	1	1
DEL	NOT_INV	TRA	TRA	TRA	94	79
DEL	INV	TRA	TRA	TRA	140	58
INV	INV	DEL	DEL	TRA	1	1
INV	INV	DEL	TRA	TRA	8	5
NA	INV	DEL	DEL	TRA	1	1
NA	INV	DEL	TRA	TRA	36	15
NA	NOT_INV	DEL	TRA	TRA	1	1
NA	INV	TRA	TRA	TRA	503	178
NA	NOT_INV	TRA	TRA	TRA	99	73
DUP	INV	DUP	DUP	DUP	26	11
DUP	NOT_INV	DUP	DUP	DUP	6	3
DUP	INV	DUP	DUP	TRA	66	20
DUP	NOT_INV	DUP	DUP	TRA	23	12
DUP	INV	DUP	TRA	TRA	101	50
DUP	NOT_INV	DUP	TRA	TRA	50	39
DUP	INV	TRA	TRA	TRA	173	88
INV	INV	DUP	DUP	DUP	97	12
INV	INV	DUP	DUP	TRA	160	38
INV	INV	DUP	TRA	TRA	257	87
NA	INV	DUP	DUP	DUP	48	9
NA	NOT_INV	DUP	DUP	DUP	5	2
NA	INV	DUP	DUP	TRA	178	44
NA	NOT_INV	DUP	DUP	TRA	38	10
NA	INV	DUP	TRA	TRA	295	121
NA	NOT_INV	DUP	TRA	TRA	85	49
DEL	INV	DEL	DEL	DUP	1	1
DEL	NOT_INV	DEL	DUP	DUP	1	1
DEL	INV	DEL	DUP	DUP	1	1
DEL	INV	DEL	DUP	TRA	2	2

DEL	INV	DUP	TRA	TRA	96	52
DEL	NOT_INV	DUP	TRA	TRA	46	34
DUP	INV	DEL	DUP	DUP	2	1
DUP	NOT_INV	DEL	DUP	TRA	6	6
DUP	INV	DEL	DUP	TRA	5	3
DUP	INV	DEL	TRA	TRA	9	4
DUP	NOT_INV	DEL	TRA	TRA	1	1
INV	INV	DEL	DEL	DUP	2	1
INV	INV	DEL	DUP	DUP	5	2
INV	INV	DEL	DUP	TRA	16	12
NA	INV	DEL	DEL	DUP	1	1
NA	INV	DEL	DUP	TRA	11	4
DEL	INV	DUP	DUP	DUP	41	10
DEL	NOT_INV	DUP	DUP	DUP	8	6
DEL	INV	DUP	DUP	TRA	73	20
DEL	NOT_INV	DUP	DUP	TRA	25	11
DUP	NOT_INV	TRA	TRA	TRA	128	83
INV	INV	TRA	TRA	TRA	505	168

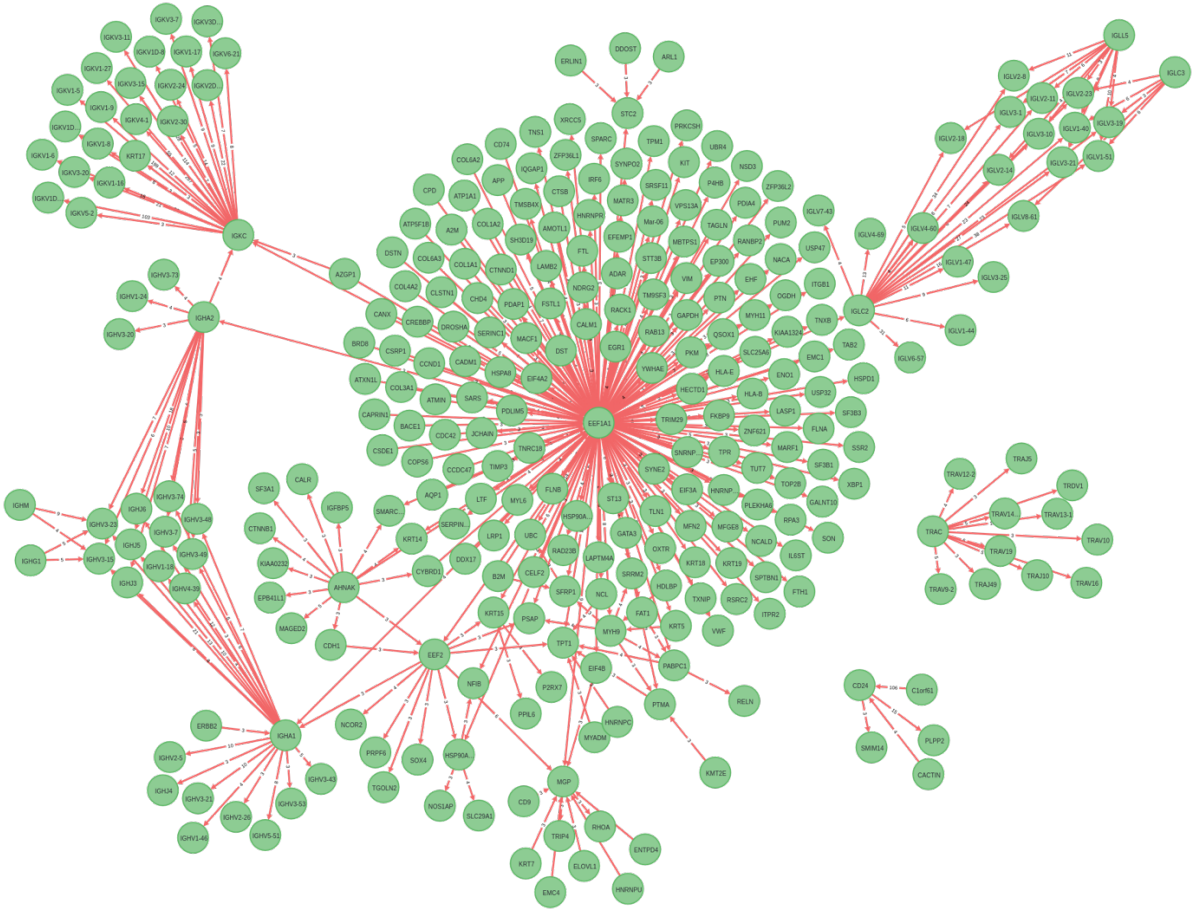
*** Type of the intrachromosomal SV. "NA" refers to cases where there was not any intrachromosomal SV.**
**** When the trisomy pattern involved one intrachromosomal SV, we determined whether it involved inversion (INV) or not (NOT_INV).**
N: number of; DEL: deletion; DUP: duplication; TRA: translocation.

Supplementary Table 3. Number of paired-reads from RNA-seq libraries from metastatic (in purple) and PCAWG normal samples (in yellow). We performed a downsampling of reads in each of the 18 metastases, setting the number of reads to 70M, as the mean number of reads in the PCAWG dataset.

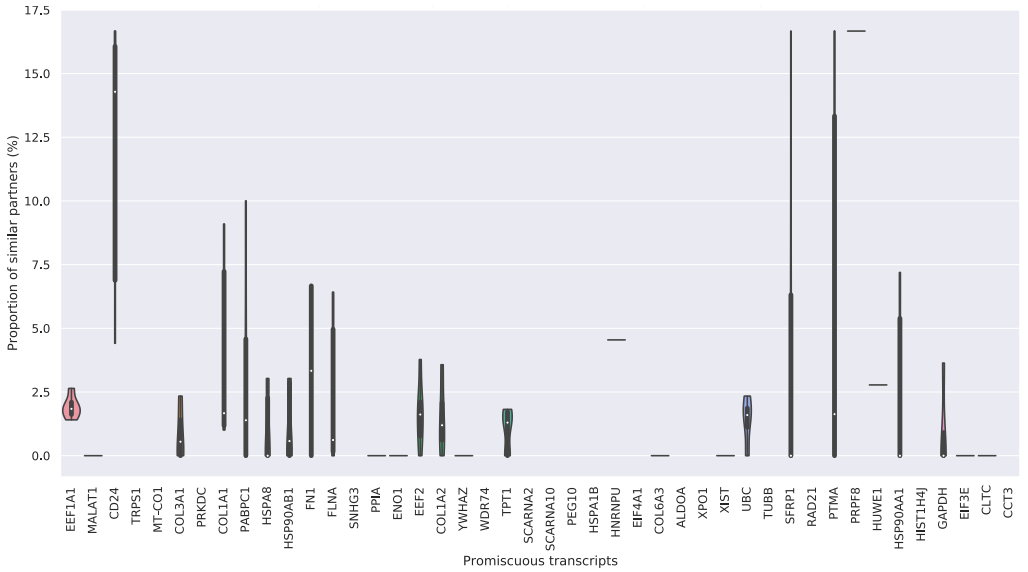
Sample ID	N pair reads
302_005_A	143165646
302_005_B	136137952
302_005_C	147603729
302_006_A	144599341
302_006_B	152355827
302_007	146847013
302_008_A	148825670
302_008_B	158985682
302_009	129105375
302_010_C	151597672
302_010_D	119526768
302_010_E	102549401
302_011_A	128461104
302_011_B	116814872
302_012_A	146771436
302_015	125129469
302_016	124503868
302_018	125910122
PCAWG 1	44777761
PCAWG 2	80845232
PCAWG 3	83800332
PCAWG 4	69895149
PCAWG 5	54053049
PCAWG 6	86151312



Supplementary Figure 1. Schematic representation of the in-silico reconstruction of the fusion *CD24-HSP90AB1* and the sequences obtained by Sanger sequencing. We sequenced the PCR product by using the primers HSP90AB1_III_F and HSP90AB1_III_R (displayed in the white boxes), obtaining the corresponding predicted sequences by the RNA-seq data, with and extended region of the transcript from *HSP90AB1* gene (in light blue) and a polyA-like tail near the fusion point. We further sequenced the PCR product using the primers Nested_HSP90AB1_III_F and Nested_HSP90AB1_III_R to capture the sequence of the fusion point. The *HSP90AB1* transcript was sequenced using the Nested_HSP90AB1_III_F primer, designed on the CD24 region, evidencing the fusion of these two molecules.



Supplementary Figure 2. Transcript fusion networks involving promiscuous transcripts in healthy breast sample. Network graph from the PCAGW normal sample 2 (using Neo4j®), showing the fusions (edges) between the genes displayed as nodes.



Supplementary Figure 3. Proportion of similar (E-value <= 10^{-3}) partners per promiscuous transcript in the normal breast samples. The violin plot displays the 44 promiscuous transcripts found in the metastatic samples from patient 302, of which only 28 were found in the normal samples. These shared promiscuous genes had partners not related.

8 REFERENCES

1. Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, B. F. No Title. *Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. (https://gco.iarc.fr/today, accessed March 2021)* https://gco.iarc.fr/today (2020).

2. Stratton, M., Campbell, P. & Futreal, P. The cancer genome. *Nature* **458**, 719–724 (2009).

3. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature 2000 406:6797* **406**, 747–752 (2000).

4. J, G. *et al.* The consensus molecular subtypes of colorectal cancer. *Nature medicine* **21**, 1350–1356 (2015).

5. AA, A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).

6. Williams, S. C. P. News feature: Capturing cancer's complexity: Some researchers believe that a tumor's heterogeneity provides crucial clues about how cancers respond to treatment. *Proceedings of the National Academy of Sciences of the United States of America* vol. 112 4509–4511 (2015).

7. FRANKLIN, R. E. & GOSLING, R. G. Molecular Configuration in Sodium Thymonucleate. *Nature 1953 171:4356* **171**, 740–741 (1953).

8. WATSON, J. D. & CRICK, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature 1953 171:4356* **171**, 737–738 (1953).

9. JC, V. *et al.* The sequence of the human genome. *Science (New York, N.Y.)* **291**, 1304–1351 (2001).

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

11. Abdellah, Z. *et al.* Finishing the euchromatic sequence of the human genome. *Nature 2004 431:7011* **431**, 931–945 (2004).

12. Metzker, M. L. Sequencing technologies — the next generation. *Nature Reviews Genetics 2010 11:1* **11**, 31–46 (2009).

13. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics 2016 17:6* **17**, 333–351 (2016).

14. Ulahannan, D., Kovac, M. B., Mulholland, P. J., Cazier, J.-B. & Tomlinson, I. Technical and implementation issues in using next-generation sequencing of cancers in clinical practice. *British Journal of Cancer* **109**, 827 (2013).

15. Dijk, E. L. van, Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends in Genetics* **30**, 418–426 (2014).

16. Guo, J. *et al.* Four-color DNA sequencing with 3′-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proceedings of the National Academy of Sciences* **105**, 9145–9150 (2008).

17. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics 2016 17:6* **17**, 333–351 (2016).

18. Paired-End vs. Single-Read Sequencing Technology. https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html.

19. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature 2012 490:7418* **490**, 61–70 (2012).

20. Muzny, D. M. *et al.* Comprehensive molecular characterization of human colon and rectal cancer. *Nature 2012 487:7407* **487**, 330–337 (2012).

21. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature 2008 455:7216* **455**, 1069–1075 (2008).

22. McLendon, R. *et al.* Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature 2008 455:7216* **455**, 1061–1068 (2008).

23. DJ, M. *et al.* Use of cancer-specific genomic rearrangements to quantify disease burden in plasma from patients with solid tumors. *Genes, chromosomes & cancer* **49**, 1062–1069 (2010).

24. J, L., J, B. & BC, B. Dose-dependent, complete response to imatinib of a metastatic mucosal melanoma with a K642E KIT mutation. *Pigment cell & melanoma research* **21**, 492–493 (2008).

25. J, R. *et al.* Multicenter phase II study of the oral MEK inhibitor, CI-1040, in patients with advanced non-small-cell lung, breast, colon, and pancreatic cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **22**, 4456–4462 (2004).

26. TS, M. *et al.* Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *The New England journal of medicine* **361**, 947–957 (2009).

27. AD, R. *et al.* Prognostic role of KRAS and BRAF in stage II and III resected colon cancer: results of the translational study on the PETACC-3, EORTC 40993, SAKK 60-00 trial. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **28**, 466–474 (2010).

28. Pajusalu, S. *et al.* Large gene panel sequencing in clinical diagnostics—results from 501 consecutive cases. *Clinical Genetics* **93**, 78–83 (2018).

29. B, R., M, T. & N, M. The promise of whole-exome sequencing in medical genetics. *Journal of human genetics* **59**, 5–15 (2014).

30. Katsanis, S. H. & Katsanis, N. Molecular genetic testing and the future of clinical genomics. *Nature Reviews Genetics 2013 14:6* **14**, 415–426 (2013).

31. A, P., MJ, E. & CM, P. Practical implications of gene-expression-based assays for breast oncologists. *Nature reviews. Clinical oncology* **9**, 48–57 (2011).

32. W, L., R, W., Z, Y., L, B. & Z, S. High accordance in prognosis prediction of colorectal cancer across independent datasets by multi-gene module expression profiles. *PloS one* **7**, (2012).

33. CW, D. *et al.* Expression signature of IFN/STAT1 signaling genes predicts poor survival outcome in glioblastoma multiforme in a subtype-specific manner. *PloS one* **7**, (2012).

34. J, B. *et al.* Biomarker discovery in non-small cell lung cancer: integrating gene expression profiling, meta-analysis, and tissue microarray validation. *Clinical cancer research : an official journal of the American Association for Cancer Research* **19**, 194–204 (2013).

35. Q, S. *et al.* Dissecting intratumoral myeloid cell plasticity by single cell RNA-seq. *Cancer medicine* **8**, 3072–3085 (2019).

36. ML, S. & I, T. Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges. *Molecular cell* **75**, 7–12 (2019).

37. Boveri, T. Concerning the Origin of Malignant Tumours by Theodor Boveri. Translated and annotated by Henry Harris. *Journal of Cell Science* **121**, 1–84 (2008).

38. P, R. A SARCOMA OF THE FOWL TRANSMISSIBLE BY AN AGENT SEPARABLE FROM THE TUMOR CELLS. *The Journal of experimental medicine* **13**, 397–411 (1911).

39. PH, D. & PK, V. Differences between the ribonucleic acids of transforming and nontransforming avian tumor viruses. *Proceedings of the National Academy of Sciences of the United States of America* **67**, 1673–1680 (1970).

40. D, S., HE, V., JM, B. & PK, V. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature* **260**, 170–173 (1976).

41. Loeb, L. A. & Harris, C. C. Advances in Chemical Carcinogenesis: A Historical Review and Prospective. *Cancer Research* **68**, 6863–6872 (2008).

42. Tabin, C. J. *et al.* Mechanism of activation of a human oncogene. *Nature 1982 300:5888* **300**, 143–149 (1982).

43. Reddy, E. P., Reynolds, R. K., Santos, E. & Barbacid, M. A point mutation is responsible for the acquisition of transforming properties by the T24

human bladder carcinoma oncogene. *Nature 1982 300:5888* **300**, 149–152 (1982).

44. Knudson, A. G. Mutation and Cancer: Statistical Study of Retinoblastoma. *Proceedings of the National Academy of Sciences* **68**, 820–823 (1971).

45. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483–1489 (2015).

46. I, M. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).

47. L, G. *et al.* Assessing matched normal and tumor pairs in next-generation sequencing studies. *PloS one* **6**, (2011).

48. Y, S. *et al.* High frequency of mutations of the PIK3CA gene in human cancers. *Science (New York, N.Y.)* **304**, 554 (2004).

49. Samuels, Y. *et al.* High Frequency of Mutations of the PIK3CA Gene in Human Cancers. *Science* **304**, 554–554 (2004).

50. JG, P. *et al.* EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science (New York, N.Y.)* **304**, 1497–1500 (2004).

51. TJ, L. *et al.* Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *The New England journal of medicine* **350**, 2129–2139 (2004).

52. W, P. *et al.* EGF receptor gene mutations are common in lung cancers from “never smokers” and are associated with sensitivity of tumors to

gefitinib and erlotinib. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 13306–13311 (2004).

53. Tate, J. G. *et al.* COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research* **47**, D941–D947 (2019).

54. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics 2013 45:10* **45**, 1113–1120 (2013).

55. Hudson, T. J. *et al.* International network of cancer genome projects. *Nature 2010 464:7291* **464**, 993–998 (2010).

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

57. Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature 2011 470:7332* **470**, 59–65 (2011).

58. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nature Reviews Genetics* **7**, 85–97 (2006).

59. Korbel, J. O. *et al.* Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *Science (New York, N.Y.)* **318**, 420 (2007).

60. PJ, C. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature genetics* **40**, 722–729 (2008).

61. ARTICLE An integrated map of structural variation in 2,504 human genomes. doi:10.1038/nature15394.

62. Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214–220 (2011).

63. Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109–1113 (2010).

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

65. Alaei-Mahabadi, B., Bhadury, J., Karlsson, J. W., Nilsson, J. A. & Larsson, E. Global analysis of somatic structural genomic alterations and their impact on gene expression in diverse human cancers. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 13768–13773 (2016).

66. Drier, Y. *et al.* Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Research* **23**, 228–235 (2013).

67. Yang, L. *et al.* Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**, 919–929 (2013).

68. Inaki, K. & Liu, E. T. Structural mutations in cancer: Mechanistic and functional insights. *Trends in Genetics* **28**, 550–559 (2012).

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

70. Zhang, Y. *et al.* A Pan-Cancer Compendium of Genes Deregulated by Somatic Genomic Rearrangement across More Than 1,400 Cases. *Cell Reports* **24**, 515–527 (2018).

71. Nangalia, J. & Campbell, P. J. Genome Sequencing during a Patient's Journey through Cancer. *New England Journal of Medicine* **381**, 2145–2156 (2019).

72. Currall, B. B., Chiangmai, C., Talkowski, M. E. & Morton, C. C. Mechanisms for Structural Variation in the Human Genome. *Current Genetic Medicine Reports* **1**, 81–90 (2013).

73. Neelsen, K. J. & Lopes, M. Replication fork reversal in eukaryotes: From dead end to dynamic response. *Nature Reviews Molecular Cell Biology* vol. 16 207–220 (2015).

74. Ciccia, A. & Elledge, S. J. The DNA Damage Response: Making It Safe to Play with Knives. *Molecular Cell* vol. 40 179–204 (2010).

75. Keeney, S., Lange, J. & Mohibullah, N. Self-organization of meiotic recombination initiation: General principles and molecular pathways. *Annual Review of Genetics* **48**, 187–214 (2014).

76. Alt, F. W., Zhang, Y., Meng, F. L., Guo, C. & Schwer, B. Mechanisms of programmed DNA lesions and genomic instability in the immune system. *Cell* vol. 152 417–429 (2013).

77. Doksani, Y. & de Lange, T. The role of double-strand break repair pathways at functional and dysfunctional telomeres. *Cold Spring Harbor Perspectives in Biology* **6**, (2014).

78. Parks, M. M., Lawrence, C. E. & Raphael, B. J. Detecting non-allelic homologous recombination from high-throughput sequencing data. *Genome Biology* **16**, (2015).

79. Ou, Z. *et al.* Observation and prediction of recurrent human translocations mediated by NAHR between nonhomologous chromosomes. *Genome Research* **21**, 33–46 (2011).

80. Roychowdhury, T. & Abyzov, A. Chromatin organization modulates the origin of heritable structural variations in human genome. *Nucleic Acids Research* **47**, 2766–2777 (2019).

81. Hoeijmakers, J. H. J. Genome maintenance mechanisms for preventing cancer. *Nature* vol. 411 366–374 (2001).

82. Prakash, R., Zhang, Y., Feng, W. & Jasin, M. Homologous recombination and human health: The roles of BRCA1, BRCA2, and associated proteins. *Cold Spring Harbor Perspectives in Biology* **7**, (2015).

83. Kim, H. & D'Andrea, A. D. Regulation of DNA cross-link repair by the Fanconi anemia/BRCA pathway. *Genes and Development* vol. 26 1393–1408 (2012).

84. Bennardo, N., Cheng, A., Huang, N. & Stark, J. M. Alternative-NHEJ is a mechanistically distinct pathway of mammalian chromosome break repair. *PLoS Genetics* **4**, 1000110 (2008).

85. McVey, M. & Lee, S. E. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends in Genetics* vol. 24 529–538 (2008).

86. Bhargava, R., Onyango, D. O. & Stark, J. M. Regulation of Single-Strand Annealing and its Role in Genome Maintenance. *Trends in Genetics* vol. 32 566–575 (2016).

87. Zhang, F., Carvalho, C. M. B. & Lupski, J. R. Complex human chromosomal and genomic rearrangements. *Trends in Genetics* vol. 25 298–307 (2009).

88. Zhang, F. *et al.* The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nature Genetics* **41**, 849–853 (2009).

89. Zhang, F. *et al.* The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nature Genetics* **41**, 849–853 (2009).

90. Liu, P. *et al.* Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell* **146**, 889–903 (2011).

91. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).

92. Baca, S., Prandi, D. & Lawrence, M. Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–77 (2013).

93. Rausch, T. *et al.* Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* **148**, 59–71 (2012).

94. Malhotra, A. *et al.* Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. *Genome Research* **23**, 762–776 (2013).

95. Kloosterman, W. P. *et al.* Chromothripsis is a common mechanism driving genomic rearrangements in primary and metastatic colorectal cancer. *Genome biology* **12**, (2011).

96. Voronina, N. *et al.* The landscape of chromothripsis across adult cancer types. *Nature Communications* **11**, (2020).

97. Stephens, P. J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005–1010 (2009).

98. Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109–1113 (2010).

99. Korbel, J. O. & Campbell, P. J. Criteria for inference of chromothripsis in cancer genomes. *Cell* vol. 152 1226–1236 (2013).

100. Gisselsson, D. *et al.* Chromosomal breakage-fusion-bridge events cause genetic intratumor heterogeneity. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 5357–5362 (2000).

101. Debatisse, M., le Tallec, B., Letessier, A., Dutrillaux, B. & Brison, O. Common fragile sites: Mechanisms of instability revisited. *Trends in Genetics* vol. 28 22–32 (2012).

102. Cortés-Ciriano, I. *et al.* Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nature Genetics* **52**, 331–341 (2020).

103. Zhang, W. *et al.* Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biology 2015 16:1* **16**, 1–12 (2015).

104. Kuksin, M. *et al.* Applications of single-cell and bulk RNA sequencing in onco-immunology. *European Journal of Cancer* **149**, 193–210 (2021).

105. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature 2012 489:7414* **489**, 57–74 (2012).

106. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics 2013 45:6* **45**, 580–585 (2013).

107. Wei, F.-Z. *et al.* Differential Expression Analysis Revealing CLCA1 to Be a Prognostic and Diagnostic Biomarker for Colorectal Cancer. *Frontiers in Oncology* **0**, 2342 (2020).

108. Lin, S. *et al.* Comprehensive analysis on the expression levels and prognostic values of LOX family genes in kidney renal clear cell carcinoma. *Cancer Medicine* **9**, 8624–8638 (2020).

109. Herold, T. *et al.* An eight-gene expression signature for the prediction of survival and time to treatment in chronic lymphocytic leukemia. *Leukemia 2011 25:10* **25**, 1639–1645 (2011).

110. Zhang, Y., Qian, J., Gu, C. & Yang, Y. Alternative splicing and cancer: a systematic review. *Signal Transduction and Targeted Therapy 2021 6:1* **6**, 1–14 (2021).

111. Singh, B. & Eyras, E. The role of alternative splicing in cancer. *https://doi.org/10.1080/21541264.2016.1268245* **8**, 91–98 (2017).

112. Sana, J., Faltejskova, P., Svoboda, M. & Slaby, O. Novel classes of non-coding RNAs and cancer. *Journal of Translational Medicine* **10**, 103 (2012).

113. Anastasiadou, E., Jacob, L. S. & Slack, F. J. Non-coding RNA networks in cancer. *Nature reviews. Cancer* **18**, 5 (2018).

114. GA, C. *et al.* Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 15524–15529 (2002).

115. Hu, X. *et al.* A Functional Genomic Approach Identifies FAL1 as an Oncogenic Long Noncoding RNA that Associates with BMI1 and Represses p21 Expression in Cancer. *Cancer Cell* **26**, 344–357 (2014).

116. YY, T. *et al.* PVT1 dependence in cancer with MYC copy-number increase. *Nature* **512**, 82–86 (2014).

117. Anastasiadou, E., Faggioni, A., Trivedi, P. & Slack, F. J. The Nefarious Nexus of Noncoding RNAs in Cancer. *International Journal of Molecular Sciences 2018, Vol. 19, Page 2072* **19**, 2072 (2018).

118. Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews Molecular Cell Biology 2020 22:2* **22**, 96–118 (2020).

119. Open Access Citation; Faraji, F. *et al.* Post-transcriptional Control of Tumor Cell Autonomous Metastatic Potential by CCR4-NOT Deadenylase CNOT7. *PLoS Genet* **12**, 1005820 (2016).

120. Patel, S. A. & Vanharanta, S. Epigenetic determinants of metastasis. *Molecular Oncology* **11**, 79–96 (2017).

121. Linde, N. *et al.* Macrophages orchestrate breast cancer early dissemination and metastasis. *Nature Communications* **9**, 1–14 (2018).

122. SA, M. *et al.* Mesenchyme Forkhead 1 (FOXC2) plays a key role in metastasis and is associated with aggressive basal-like breast cancers. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 10069–10074 (2007).

123. J, Y. *et al.* Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. *Cell* **117**, 927–939 (2004).

124. L, M. *et al.* Therapeutic silencing of miR-10b inhibits metastasis in a mouse mammary tumor model. *Nature biotechnology* **28**, 341–347 (2010).

125. SF, T. *et al.* Endogenous human microRNAs that suppress breast cancer metastasis. *Nature* **451**, 147–152 (2008).

126. Bray, F. *et al.* 394 CA: A Cancer Journal for Clinicians Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA CANCER J CLIN* **68**, 394–424 (2018).

127. Harbeck, N. *et al.* Breast cancer. *Nature Reviews Disease Primers* **5**, 1–31 (2019).

128. EV, B., G, M., I, R. & EM, J. Racial and ethnic disparities in the impact of obesity on breast cancer risk and survival: a global perspective. *Advances in nutrition (Bethesda, Md.)* **6**, 803–819 (2015).

129. M, P.-R., C, M.-T., JJ, V.-G., ER, F. & JM, S. Obesity and adverse breast cancer risk and outcome: Mechanistic insights and strategies for intervention. *CA: a cancer journal for clinicians* **67**, 378–397 (2017).

130. Mørch, L. S. *et al.* Contemporary Hormonal Contraception and the Risk of Breast Cancer. *http://dx.doi.org/10.1056/NEJMoa1700732* **377**, 2228–2239 (2017).

131. L, D. P., G, C.-P. & F, P. Breast cancer risk of hormonal contraception: Counselling considering new evidence. *Critical reviews in oncology/hematology* **137**, 123–130 (2019).

132. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature 2000 406:6797* **406**, 747–752 (2000).

133. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences* **98**, 10869–10874 (2001).

134. JX, R., Y, G., H, L., X, H. & ZM, S. Racial/ethnic differences in the outcomes of patients with metastatic breast cancer: contributions of demographic, socioeconomic, tumor and metastatic characteristics. *Breast cancer research and treatment* **173**, 225–237 (2019).

135. AG, W. & EP, W. Breast Cancer Treatment: A Review. *JAMA* **321**, 288–300 (2019).

136. C, K. *et al.* Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell* **173**, 879-893.e13 (2018).

137. Shackleton, M., Quintana, E., Fearon, E. R. & Morrison, S. J. Heterogeneity in Cancer: Cancer Stem Cells versus Clonal Evolution. *Cell* **138**, 822–829 (2009).

138. Shiovitz, S. & Korde, L. A. Genetics of breast cancer: a topic in evolution. *Annals of Oncology* **26**, 1291 (2015).

139. Kuchenbaecker, K. B. *et al.* Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. *JAMA* **317**, 2402–2416 (2017).

140. EF, C., KJ, M. & SD, M. Updates on breast cancer genetics: Clinical implications of detecting syndromes of inherited increased susceptibility to breast cancer. *Seminars in oncology* **43**, 528–535 (2016).

141. B, C. *et al.* Multi-gene panel testing for hereditary cancer predisposition in unsolved high-risk breast and ovarian cancer patients. *Breast cancer research and treatment* **163**, 383–390 (2017).

142. A, T. *et al.* Consensus for genes to be included on cancer panel tests offered by UK genetics services: guidelines of the UK Cancer Genetics Group. *Journal of medical genetics* **55**, 372–377 (2018).

143. Allred, D. C. Ductal Carcinoma In Situ: Terminology, Classification, and Natural History. *Journal of the National Cancer Institute. Monographs* **2010**, 134 (2010).

144. Cowell, C. F. *et al.* Progression from ductal carcinoma in situ to invasive breast cancer: Revisited. *Molecular Oncology* **7**, 859–869 (2013).

145. Polyak, K. Is Breast Tumor Progression Really Linear? *Clinical Cancer Research* **14**, 339–341 (2008).

146. T, O. Extracellular matrix components in breast cancer progression and metastasis. *Breast (Edinburgh, Scotland)* **22 Suppl 2**, (2013).

147. A, B., LJ, van't V. & MJ, B. An "elite hacker": breast tumors exploit the normal microenvironment program to instruct their progression and biological diversity. *Cell adhesion & migration* **6**, 236–435 (2012).

148. C, S., L, C., P, D. S., C, C. & M, L. Tumor-infiltrating lymphocytes in breast cancer according to tumor subtype: Current state of the art. *Breast (Edinburgh, Scotland)* **35**, 142–150 (2017).

149. D, N. & SEB, M. Immune Landscape of Breast Cancers. *Biomedicines* **6**, (2018).

150. M, R. *et al.* Remodeling of the methylation landscape in breast cancer metastasis. *PloS one* **9**, (2014).

151. KJ, P. *et al.* MicroRNA-335 inhibits tumor reinitiation and is silenced through genetic and epigenetic mechanisms in human breast cancer. *Genes & development* **25**, 226–231 (2011).

152. J, C. *et al.* Histone demethylase RBP2 is critical for breast cancer progression and metastasis. *Cell reports* **6**, 868–877 (2014).

153. DR, H. Metastasis suppression by BRMS1 associated with SIN3 chromatin remodeling complexes. *Cancer metastasis reviews* **31**, 641–651 (2012).

154. L, W. *et al.* CARM1 methylates chromatin remodeling factor BAF155 to enhance tumor progression and metastasis. *Cancer cell* **25**, 21–36 (2014).

155. Nacu, S. *et al.* Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *BMC Medical Genomics* **4**, 11 (2011).

156. KE, V. *et al.* Recurrent read-through fusion transcripts in breast cancer. *Breast cancer research and treatment* **146**, 287–297 (2014).

157. H, L., J, W., X, M. & J, S. Gene fusions and RNA trans-splicing in normal and neoplastic human cells. *Cell cycle (Georgetown, Tex.)* **8**, 218–222 (2009).

158. K, J. & H, L. Chimeric RNAs generated by intergenic splicing in normal and cancer cells. *Genes, chromosomes & cancer* **53**, 963–971 (2014).

159. Latysheva, N. S. & Babu, M. M. Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic Acids Research* (2016) doi:10.1093/nar/gkw282.

160. Tomlins, S. A. *et al.* Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer. *Science* **310**, 644–648 (2005).

161. Hochhaus, A., Rosée, P. Ia, Müller, M. C., Ernst, T. & Cross, N. C. P. Impact of BCR-ABL mutations on patients with chronic myeloid leukemia. *http://dx.doi.org/10.4161/cc.10.2.14537* **10**, 250–260 (2011).

162. Yu, Y.-P. *et al.* Identification of recurrent fusion genes across multiple cancer types. *Scientific Reports* **9**, (2019).

163. ROWLEY, J. D. A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia identified by Quinacrine Fluorescence and Giemsa Staining. *Nature 1973 243:5405* **243**, 290–293 (1973).

164. JD, R. Identificaton of a translocation with quinacrine fluorescence in a patient with acute leukemia. *Annales de genetique* **16**, 109–112 (1973).

165. L, Z., U, H., K, N. & G, K. Characteristic chromosomal abnormalities in biopsies and lymphoid-cell lines from patients with Burkitt and non-Burkitt lymphomas. *International journal of cancer* **17**, 47–56 (1976).

166. R, B. *et al.* A new translocation in Burkitt's tumor cells. *Human genetics* **53**, 111–112 (1979).

167. Fukuhara, S., Rowley, J. D., Variakojis, D. & Golomb, H. M. Chromosome Abnormalities in Poorly Differentiated Lymphocytic Lymphoma. *Cancer Research* **39**, (1979).

168. Shtivelman, E., Lifshitz, B., Gale, R. P. & Canaani, E. Fused transcript of abl and bcr genes in chronic myelogenous leukaemia. *Nature 1985 315:6020* **315**, 550–554 (1985).

169. K, S. *et al.* Evidence of a new chimeric bcr/c-abl mRNA in patients with chronic myelocytic leukemia and the Philadelphia chromosome. *The New England journal of medicine* **313**, 1429–1433 (1985).

170. MJ, T. *et al.* Rearrangement of the MLL gene in acute lymphoblastic and acute myeloid leukemias with 11q23 chromosomal translocations. *The New England journal of medicine* **329**, 909–914 (1993).

171. Morrissey, J. *et al.* A Serine/Proline-Rich Protein Is Fused To HRX in t(4;11) Acute Leukemias. *Blood* **81**, 1124–1131 (1993).

172. Y, G. *et al.* The t(4;11) chromosome translocation of human acute leukemias fuses the ALL-1 gene, related to Drosophila trithorax, to the AF-4 gene. *Cell* **71**, 701–708 (1992).

173. SA, T. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science (New York, N.Y.)* **310**, 644–648 (2005).

174. Maher, C. A. *et al.* Transcriptome sequencing to detect gene fusions in cancer. *Nature 2009 458:7234* **458**, 97–101 (2009).

175. Maher, C. A. *et al.* Chimeric transcript discovery by paired-end transcriptome sequencing. *Proceedings of the National Academy of Sciences* **106**, 12353–12358 (2009).

176. Muzny, D. M. *et al.* Comprehensive molecular characterization of human colon and rectal cancer. *Nature 2012 487:7407* **487**, 330–337 (2012).

177. Stephens, P. J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature 2009 462:7276* **462**, 1005–1010 (2009).

178. C, S. *et al.* MHC class II transactivator CIITA is a recurrent gene fusion partner in lymphoid cancers. *Nature* **471**, 377–383 (2011).

179. K, Y. *et al.* The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene* **34**, 4845–4854 (2015).

180. Mitelman Database Chromosome Aberrations and Gene Fusions in Cancer. https://mitelmandatabase.isb-cgc.org/.

181. Nacu, S. *et al.* Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *BMC Medical Genomics 2011 4:1* **4**, 1–22 (2011).

182. DS, R. *et al.* SLC45A3-ELK4 is a novel and frequent erythroblast transformation-specific fusion transcript in prostate cancer. *Cancer research* **69**, 2737–2738 (2009).

183. Kumar, S., Razzaq, S. K., Vo, A. D., Gautam, M. & Li, H. Identifying fusion transcripts using next generation sequencing. *Wiley Interdisciplinary Reviews: RNA* **7**, 811–823 (2016).

184. Johansson, B. *et al.* Most gene fusions in cancer are stochastic events. *Genes, Chromosomes and Cancer* **58**, 607–611 (2019).

185. CA, M. *et al.* Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458**, 97–101 (2009).

186. Wang, Q., Xia, J., Jia, P., Pao, W. & Zhao, Z. Application of next generation sequencing to human gene fusion detection: computational tools, features and perspectives. *Briefings in Bioinformatics* **14**, 506 (2013).

187. M, C. *et al.* State-of-the-art fusion-finder algorithms sensitivity and specificity. *BioMed research international* **2013**, (2013).

188. M, C. *et al.* State of art fusion-finder algorithms are suitable to detect transcription-induced chimeras in normal tissues? *BMC bioinformatics* **14 Suppl 7**, (2013).

189. A, M. *et al.* Comrad: detection of expressed rearrangements by integrated analysis of RNA-Seq and low coverage genome sequence data. *Bioinformatics (Oxford, England)* **27**, 1481–1488 (2011).

190. N, P., M, S., T, C. & E, R. CRAC: an integrated approach to the analysis of RNA-seq reads. *Genome biology* **14**, (2013).

191. H, G. *et al.* FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics (Oxford, England)* **27**, 1922–1928 (2011).

192. Y, H. *et al.* FuMa: reporting overlap in RNA-seq detected fusion genes. *Bioinformatics (Oxford, England)* **32**, 1226–1228 (2016).

193. A, M. *et al.* nFuse: discovery of complex genomic rearrangements in cancer using high-throughput sequencing. *Genome research* **22**, 2250–2261 (2012).

194. Gingeras, T. R. Implications of chimaeric non-co-linear transcripts. *Nature 2009 461:7261* **461**, 206–211 (2009).

195. Kumar, S., Razzaq, S. K., Vo, A. D., Gautam, M. & Li, H. Identifying Fusion Transcripts Using Next Generation Sequencing Advanced Review. *Wiley interdisciplinary reviews. RNA* **7**, 811 (2016).

196. Neckles, C., Rajan, S. S. & Caplen, N. J. Fusion transcripts: Unexploited vulnerabilities in cancer? *Wiley Interdisciplinary Reviews: RNA* **11**, e1562 (2020).

197. P, A. & MR, L. The oncogene ERG: a key factor in prostate cancer. *Oncogene* **35**, 403–414 (2016).

198. CA, F. Pathogenesis of NUT midline carcinoma. *Annual review of pathology* **7**, 247–265 (2012).

199. CA, F. NUT Carcinoma: Clinicopathologic features, pathogenesis, and treatment. *Pathology international* **68**, 583–595 (2018).

200. Yu, Y.-P. *et al.* Identification of recurrent fusion genes across multiple cancer types. *Scientific Reports 2019 9:1* **9**, 1–9 (2019).

201. YM, W. *et al.* Identification of targetable FGFR gene fusions in diverse cancers. *Cancer discovery* **3**, 636–647 (2013).

202. McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, Mastrogianakis GM, Olson JJ, Mikkelsen T, Lehman N, Aldape K, Yung WK, Bogler O, Weinstein JN, VandenBerg S, Berger M, Prados M, Muzny D, Morgan M, Scherer S, Sabo A, Nazareth L, Lewis L, Hall O, Zhu, T. E. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).

203. Futreal, P. A. *et al.* A census of human cancer genes. *Nature Reviews Cancer* **4**, 177–183 (2004).

204. Hudson (Chairperson), T. J. *et al.* International network of cancer genome projects. *Nature* **464**, 993–998 (2010).

205. Puente, X. S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).

206. Krzywinski, M. *et al.* Circos: An information aesthetic for comparative genomics. doi:10.1101/gr.092759.109.

207. Gene Ontology Consortium, T. *et al.* Gene Ontology: tool for the unification of biology. doi:10.1038/75556.

208. Gene Ontology Consortium, T. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research* **45**, (2017).

209. Mi, H. *et al.* PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research* **45**, (2017).

210. Calabrese, C. *et al.* Genomic basis for RNA alterations in cancer. *Nature* **578**, 129–136 (2020).

211. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**, (2015).

212. Anders, S. & Huber, W. Differential expression analysis for sequence count data. doi:10.1186/gb-2010-11-10-r106.

213. Wong, E., Baur, B., Quader, S. & Huang, C. H. Biological network motif detection: Principles and practice. *Briefings in Bioinformatics* **13**, 202–215 (2012).

214. Dobin, A. *et al.* Sequence analysis STAR: ultrafast universal RNA-seq aligner. **29**, 15–21 (2013).

215. Li, B. & Dewey, C. N. *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome*. http://www.biomedcentral.com/1471-2105/12/323 (2011) doi:10.1186/1471-2105-12-323.

216. Y, C. & PS, M. Gene expression analysis via multidimensional scaling. *Current protocols in bioinformatics* **Chapter 7**, (2005).

217. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410 (1990).

218. Haas, B. et al. STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *bioRxiv 120295* (2017) doi:doi:10.1101/120295.

219. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**, (2011).

220. Bignell, G. R. *et al.* Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. doi:10.1101/gr.6522707.

221. Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, (2010).

222. Menghi, F. *et al.* The tandem duplicator phenotype as a distinct genomic configuration in cancer. doi:10.1073/pnas.1520010113.

223. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the*

National Academy of Sciences of the United States of America **107**, 139–144 (2010).

224. Yilong Li, Nicola D Roberts, Joachim Weischenfeldt, Jeremiah A Wala, Ofer Shapira, Steven E. Schumacher, Ekta Khurana, Jan Korbel, Marcin Imielinski, Rameen Beroukhim, P. J. C. Patterns of structural variation in human cancer. *bioRxiv 181339* (2017) doi:10.1101/181339.

225. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. | *Nature* | **578**, (2020).

226. Papatheodorou, I. *et al.* Expression Atlas update: from tissues to single cells. *Nucleic Acids Research* **48**, D77–D83 (2020).

227. Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications* **10**, (2019).

228. Wiley Online Library, in *et al.* Tandem duplication of chromosomal segments is common in ovarian and breast cancer genomes. *Journal of Pathology J Pathol* **227**, 446–455 (2012).

229. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* (1969) doi:10.1038/s41586-020-1969-6.

230. Takano, T. *et al.* Glutamate release promotes growth of malignant gliomas. *Nature Medicine* **7**, 1010–1015 (2001).

231. Seidlitz, E. P., Sharma, M. K., Saikali, Z., Ghert, M. & Singh, G. Cancer cell lines release glutamate into the extracellular environment. *Clinical and Experimental Metastasis* **26**, 781–787 (2009).

232. Herner, A. *et al.* Glutamate increases pancreatic cancer cell invasion and migration via AMPA receptor activation and Kras-MAPK signaling. *International Journal of Cancer* **129**, 2349–2359 (2011).

233. Stepulak, A. *et al.* AMPA antagonists inhibit the extracellular signal regulated kinase pathway and suppress lung cancer growth. *Cancer biology & therapy* **6**, 1908–15 (2007).

234. Stepulak, A. *et al.* NMDA antagonist inhibits the extracellular signal-regulated kinase pathway and suppresses cancer growth. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15605–15610 (2005).

235. Zhou, Z., Si, G., Zhang, Y. & Zheng, K. Robust clustering by identifying the veins of clusters based on kernel density estimation. *Knowledge-Based Systems* **159**, 309–320 (2018).

236. Matioli, L. C., Santos, S. R., Kleina, M. & Leite, E. A. A new algorithm for clustering based on kernel density estimation. *Journal of Applied Statistics* **45**, 347–366 (2018).

237. Oden, A. & Wedel, H. Arguments for Fisher's Permutation Test. *The Annals of Statistics* **3**, 518–520 (2007).

238. Fi, M. O. & Garriga, G. C. *Permutation Tests for Studying Classifier Performance Markus Ojala*. *Journal of Machine Learning Research* vol. 11 (2010).

239. Hadi, K. *et al.* Distinct Classes of Complex Structural Variation Uncovered across Thousands of Cancer Genome Graphs. *Cell* **183**, 197-210.e32 (2020).

240. Zhang, C. Z. *et al.* Chromothripsis from DNA damage in micronuclei. *Nature* **522**, 179–184 (2015).

241. Siegel, M. B. *et al.* Integrated RNA and DNA sequencing reveals early drivers of metastatic breast cancer. *The Journal of Clinical Investigation* **128**, 1371–1383 (2018).

242. Sallinen, H. *et al.* Comparative transcriptome analysis of matched primary and distant metastatic ovarian carcinoma. *BMC Cancer 2019 19:1* **19**, 1–11 (2019).

243. Fares, J., Fares, M. Y., Khachfe, H. H., Salhab, H. A. & Fares, Y. Molecular principles of metastasis: a hallmark of cancer revisited. *Signal Transduction and Targeted Therapy 2020 5:1* **5**, 1–17 (2020).

244. Elia, I. *et al.* Breast cancer cells rely on environmental pyruvate to shape the metastatic niche. *Nature 2019 568:7750* **568**, 117–121 (2019).

245. Engström, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods 2013 10:12* **10**, 1185–1191 (2013).

246. Zhao, S. Assessment of the Impact of Using a Reference Transcriptome in Mapping Short RNA-Seq Reads. *PLOS ONE* **9**, e101374 (2014).

247. Kentsis, A., Pikman, Y., Wang, L., Bernt, K. M. & Winters, A. C. MLL-Rearranged Leukemias-An Update on Science and Clinical Approaches STRUCTURe AND FUNCTiON OF wiLD-TYPe MLL1 Mixed-Lineage Leukemia 1 (MLL1) Protein Structure and Binding Partners. **5**, 1 (2017).

248. Meyer, C. *et al.* The MLL recombinome of acute leukemias in 2017. *Nature Publishing Group* **32**, 273–284 (2017).

249. Sankar, S. & Lessnick, S. L. Promiscuous Partnerships in Ewing's Sarcoma. doi:10.1016/j.cancergen.2011.07.008.

250. Tomlins, S. A. *et al.* Role of the TMPRSS2-ERG Gene Fusion in Prostate Cancer 1,2,3. *Neoplasia* **10**, 177–188 (2008).

251. Parris, T. Z. *et al.* Genome-wide multi-omics profiling of the 8p11-p12 amplicon in breast carcinoma. *Oncotarget* **9**, 24140–24154 (2018).

252. Engqvist, H. *et al.* Transcriptomic and genomic profiling of early-stage ovarian carcinomas associated with histotype and overall survival. *Oncotarget* **9**, 35162–35180 (2018).

253. Quail, M. A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nature Methods 2008 5:12* **5**, 1005–1010 (2008).

254. Houseley, J. & Tollervey, D. Apparent Non-Canonical Trans-Splicing Is Generated by Reverse Transcriptase In Vitro. *PLOS ONE* **5**, e12271 (2010).

255. J, C., A, C., G, Z. & RA, V. Reverse transcriptase template switching and false alternative transcripts. *Genomics* **88**, 127–131 (2006).

256. Zhu, S., Li, W. & Cao, Z. Does MMLV-RT lacking RNase H activity have the capability of switching templates during reverse transcription? *FEBS Letters* **520**, 185–185 (2002).

257. Djebali, S. *et al.* Evidence for transcript networks composed of chimeric rnas in human cells. *PLoS ONE* **7**, 28213 (2012).

258. Rodríguez-Martín, B. *et al.* ChimPipe: Accurate detection of fusion genes and transcription-induced chimeras from RNA-seq data. *BMC Genomics* **18**, 7 (2017).

9 APPENDIX

RESEARCH

Clustering and Graph Mining Techniques for Classification of Complex Structural Variations in Cancer Genomes

Gonzalo Gomez-Sanchez[1,2*†], Luisa Delgado-Serrano[3†], David Carrera[1], David Torrents[3,4] and Josep Ll. Berral[1,2*]

[*]Correspondence:
gonzalo.gomez@bsc.es;
josep.berral@bsc.es
[1]Barcelona Supercomputing Center (BSC), Department of Computer Science, Barcelona, Spain
[†]Gonzalo Gomez-Sanchez and Luisa Delgado-Serrano contributed equally to this work.

Abstract

Background: For many years, a major question in the field of cancer genomics has been the identification of those variations that can have a functional role in cancer, and distinguish from the majority of genomic changes that have no functional consequences. This is particularly challenging when considering complex chromosomal rearrangements, which are often composed of multiple DNA breaks, resulting in difficulties to classify and interpret them functionally. Despite recent efforts towards the classification of structural variants (SVs), more robust statistical frames are needed to better classify these variants, and to isolate those that derive from specific molecular mechanisms.

Results: We present a new statistical approach to analyze SVs patterns from 2392 real tumor samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium and to identify significant recurrence, which can inform of relevant mechanisms involved in the biology of tumors. The method is based on recursive KDE clustering of 152,926 SVs, graph mining techniques and statistical measures. The proposed methodology was able not only to identify complex patterns but also to prove them as not random occurrences. Furthermore, a new class of pattern that was not previously described has been identified.

Conclusions: We developed a new and unbiased methodology for clustering SVs to search further for complex patterns by using a cost-efficient graph mining method. Followed by deep statistical analysis and applying randomization techniques, our proposed framework allows for discerning between stochastic chromosomal rearrangements and complex patterns that might have specific underlying mechanisms present in different cancer types.

Keywords: complex rearrangements; structural variants; cancer genomics; clustering; graph mining; motif finding

Background

Cancer is a disease mainly driven by genetic alterations that take place in somatic cells. The somatic mutations in a cancer cell genome involve several distinct classes of DNA sequence change. A particularly important class of somatic mutations are structural variations that consist of genomic rearrangements, such as large deletions, large insertions, tandem duplications, inversions, and translocations [1]. In the last decade, whole-genome analysis has shown that several SVs are not independent events driven by genome instability, but they are acquired through a "single-hit" event involving several DNA breaks, usually resulting in complex genome rear-

rangements. However, there is not a standard methodology to identify recurrent and statistically significant patterns of SVs, being the key for understanding the underlying mechanisms of complex rearrangements, by which they might contribute to the development of tumors.

A few classes of complex chromosomal rearrangements have been described in tumor genomes. For example, in 2011, Stephens and co-workers described an SV pattern characterized by multiple (sometimes hundreds) rearrangements that occur within a restricted portion of the genome, involving normally one, but also rarely two chromosomes [2]. In another study, Baca and co-workers reported another specific pattern of chromosomal rearrangements in prostate tumors called Chromoplexy, which is characterized by a closed chain of translocations involving several chromosomes [3]. Recently, the PCAWG Consortium collected whole genome sequencing data from 2392 cancers across 36 tumor types, produced by the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) projects [4]. There, Li and co-workers described a replication-based mechanism of structural variation that results in varied chromosomal structures with low-level copy number gains and recurring inverted rearrangements [5]. Despite all these efforts to classify and characterize these complex events, a major fraction of the identified SVs in the PCAWG study remained "unclassified."

In order to fulfill this gap, we developed an innovative statistical approach to be able to discriminate between stochastic chromosomal rearrangements, probably due to general genome instability, from those patterns that might have specific and recurrent molecular mechanisms behind. The generation of such a workflow will allow the overall improvement of classification methods for the discrimination of mutations and to identify particular SV signatures as markers of tumor formation and progression.

Here, we applied this new statistical frame to 2392 tumor genomes from the PCAWG Consortium, including more than 152,926 SVs. This method takes into account the local distribution of SVs in every sample and is optimized using the global distribution across the dataset, using a Kernel Density Estimation function [6, 7]. The aim of the clustering is to join the rearrangements that are likely derived from the same molecular mechanism, as they share some topological properties. We assessed that the clustering approach joins rearrangements not randomly by performing a permutation test. Then, we provided a graph mining method to analyze the SV patterns, using advanced high performing technologies to reduce the computational cost [8, 9]. Finally, we adapted a methodology proposed by Wong [10] to obtain the level of significance of the different patterns based on the Abundance, a measure that indicates the overrepresentation or underrepresentation of a pattern against a random scenario.

By overcoming currently unsolved challenges of SVs classification in cancer, our results provide insights into different complex patterns of SVs emerging from possible chromatin conformations that allow interactions of different genomic regions that are occurring in a cancer cell.

Methods

Our main strategy for the identification of complex chromosomal rearrangements is summarized in Fig. 1. Preceded by a quality check and pre-processing of the

PCAWG data, the main workflow is composed of three major steps: KDE clustering, graph mining, and motif finding.

Defining clusters to identify the SVs involved in complex rearrangements

The clustering method was based on Kernel Density Estimation (KDE) [6, 7], a non-parametric statistical method to estimate the probability density function of a random variable. In this study, the random variable is the position of the SV, which is defined by the points where the rearrangements occur in the genome (breakpoints). Clustering those breakpoints that correspond to the same single rearrangement event is crucial to later classify complex patterns of SVs. We chose this clustering method because it uses a density estimation of the breakpoints as a starting point, which allowed us to rely both on the closeness of the breakpoints and their density. Using the Gaussian Kernel based on normal distribution, the only hyperparameter to be set was the bandwidth [11, 12]. This value defines how the density estimation is going to be: increasing the bandwidth leads to bigger clusters, whereas low values generate smaller and sparser clusters. The final size of each cluster will depend on both the selected bandwidth and the density of the breakpoints for each particular case. We had to set a bandwidth that provides the lowest intra-cluster distance, defined as the highest distance between two breakpoints within the same cluster, and the highest inter-cluster distance, defined as the lowest distance between two breakpoints of adjacent clusters, both illustrated in Fig. 2. Optimizing these two distances makes the clusters to be as separate from each other as possible and as small as possible. These distances were obtained for all samples at the same time, fixing the same bandwidth value for every sample. Therefore, taking into account the global breakpoints distribution across all the samples to set the bandwidth value, we were able to avoid potential biases derived from a particular sample distribution and to join together two clusters or not when needed.

Since the human genome is organized into 23 pairs of chromosomes, we performed the clustering locally at every chromosome. Figure 2 shows how the method works using different bandwidth values on the same region of a given chromosome. Once the clustering was done, the next step was to locate all the peaks of the function and assign the breakpoints to the closest peak. These peaks represent the cluster centers to use for all the breakpoints assigned to each cluster at the graph mining step (see below).

In order to improve the clustering resolution, a recursive 2-step clustering was carried out: after the first KDE clustering process, we performed a second clustering inside every cluster. To avoid already described complex patterns, such as Chromothripsis, the breakpoints looping over the same region were discarded after the first clustering round. This process made the mining of motifs computationally more efficient, avoiding noise into the second step of the clustering. In the second round of clustering, different bandwidth hyper-parameters were set to compare intra-cluster and inter-cluster distances. Since the clustering method was based on a density estimation function, we ensured a linear growth of the number of operations with the increase of data. Since both the density estimation and the final cluster selection only interact with data from a region of the chromosome at a time, the number of operations of the method will always be smaller than n^2, where n is the

number of breakpoints, avoiding high computational expenses. To provide a better understanding of the method, the pseudocode of the full clustering process can be found in Algorithm 1.

Algorithm 1: KDE Recursive Clustering

```
   Input: sample's breakpoints: sample
   Output: sample's clusters: cluster_dict
1  set bw_1, bw_2
2  for every sample do
3      init cluster_dict
4      read breakpoints
5      for every chromosome do
6          kde := compute kde (bw_1)
7          clusters := local maximus in kde
8          for every cluster do
9              remove Cromothripsis
10             kde_cl := compute kde (bw_2)
11             for every breakpoint do
12                 clusters := local maximus in kde_cl
13                 cluster_bk := find closest cluster
14                 cluster_dict[bkpoint] := cluster_bk
                   /* Assign breakpoints to the closest cluster                    */
15     save cluster_dict
```

We validated that our clustering approach was not joining random SVs by performing two tests. First, we generated simulated datasets 100 times by pooling together all the breakpoints of the samples, and creating new samples with random rearrangements. In the first test, we estimated the average dispersion of breakpoints in each simulated dataset. We used as a dispersion measure the standard deviation of the difference of base pair between adjacent breakpoints in a chromosome. Then, we compared the average dispersion distribution from the simulated datasets against the average dispersion of breakpoints in the original dataset performing a one sample Z-Test. In the second test, we applied the KDE clustering method to each simulated dataset as described for the original dataset. For each permutation we calculated the average cluster density defined as the average number of breakpoints per cluster and compared to the average number of breakpoints per cluster in the original dataset using a one sample Z-Test.

Graph Mining to search for complex rearrangements

The clustering process set out every sample as a graph where the breakpoint clusters are represented as vertices and the edges connecting these vertices correspond to the rearrangements. Since vertices could be composed of several breakpoints from different rearrangements, different graphs could be generated. To narrow down the survey of graphs, we focused only on Hamiltonian cycles (mentioned further only as *cycles*), where every vertex is connected to two other vertices (Fig. 3).

To find and count rearrangement patterns inside each graph, we used a search approach method based on the VSIGRAM method [13], following a vertical approach and finding the frequent subgraphs in a depth-first fashion. As the subgraph mining problem becomes computationally hard (NP-hard), we performed a pruned search with max size = 6. The graph-based data mining for SV pattern searching includes four steps: deduplicate edges, generate the graph, subgraph mining, and reduce similar patterns.

Deduplicate edges

Since every cluster can include more than one breakpoint, it is likely to find clusters with more than one edge going to one another cluster. These edges were therefore duplicated and had to be deduplicated, simply removing all of them except one.

Generate the graph

Next, we generated graphs for each sample, considering the cluster centers as the vertices, and the unique edges as the connecting edges of the vertices.

Subgraph mining

The used method for subgraph mining visited the graph through depth-first search, allowing parallelism, e.g. by splitting each starting vertex to be processed at the same time. At every vertex, we looked for all the possible connected paths of size 1. Then, these subgraphs were the candidates for looking for all the possible connected paths of size 2. The process was repeated for the paths of sizes 3, 4, 5, and 6. A graphic representation of this process can be found in Fig. 4 and the corresponding pseudocode in Algorithm 2.

Algorithm 2: Sub-graph mining

```
   Input: graph per sample: clusters, edges
   Output: subgraph's list per sample: subgraphs
1  set search_size
2  for every sample do
3      read clusters, edges
4      init size := 1
5      init subgraphs := clusters
6      while size ≤ search_size do
7          for every subgraph do
8              if size(subgraph) = size then
9                  get connected clusters (edges)
                   /* Find clusters connected to the actual subgraph      */
10                 for every conected cluster do
11                     subgraphs.add([subgraph,cluster])
                       /* Add the new subgraphs to keep the search with   */
12     save subgraphs
```

Reduce similar patterns

All of the subgraphs obtained from the vertices from a given sample were stored together and duplicated cases were eliminated by matching canonical labels and edge hashes.

Defining statistically significant patterns

In order to discern statistically significant patterns from random distributions, we compared frequencies between real observations and random observations from simulated datasets using a measure called Abundance (Δ), proposed by Wong [10].

Abundance measure

As defined in (1), we computed Δ for a given *cycle*, comparing f_{input}, which is defined as the frequency of a pattern in the original dataset with $\overline{f}_{random}$, the

mean of the frequencies of a pattern in N simulated random datasets. ε is a pseudo-count (Laplace smoothing) to prevent the ratio from exploding when frequencies are small. Δ can take values between -1, underrepresented and +1, overrepresented, being 0 the value for a pattern with the same representation in the original data than in the random datasets.

$$\Delta = \frac{f_{input} - \overline{f}_{random}}{f_{input} + \overline{f}_{random} + \varepsilon} \quad (1)$$

Dataset simulation test

In order to keep the same distribution of clusters as the original dataset, we randomized the edges between the clusters (the rearrangements). The randomization of the edges was performed using an adaptation of the switching method presented by Wong [10] to the graph abstraction previously described above. This method consists of repeatedly selecting two random edges A-B and C-D and exchanging the ends to form two new edges, e.g, A-D and B-C. The resulting graph keeps the same vertices and edges count. This method has a drawback: we cannot be certain when the graph is adequately randomized, but numerical studies have shown that enough random switching samples (100 x E) are adequate to achieve a randomized set, where E is the total number of edges across all samples [14]. Therefore, we generated 100 simulated datasets as follows: we removed the original edges of every sample and randomly assigned the same amount of edges to each sample every time.

Results

Clusters of SVs from complex patterns

The purpose of the clustering process is to join the rearrangements that belong to the same mutation event. Therefore, in order to select the optimal bandwidths and carry out the 2-step KDE clustering, we ran several experiments with different bandwidth values, observing that the resolution of a 1-step KDE clustering is limited by the size of the chromosomes; the density estimation was exactly the same using any bandwidth equal or smaller than 1000. A first inspection of the results showed low resolution, as breakpoints were being clustered despite being separated by hundreds of thousands base pairs, indicating the need to perform a second clustering to improve the resolution.

The final selected values for the method were bandwidth 1 = 1000 for the first step since it ensured the maximum resolution and bandwidth 2 = 400 for the second step since it showed high inter-cluster distances while still having small intra-cluster distances. As seen in Fig. 5, selecting a higher bandwidth the breakpoints were clustered with a considerable increase of the intra-cluster distance while almost not increasing the inter-cluster distance. Opposite, selecting a lower bandwidth the behavior was smaller intra-cluster distance but with a significant decrease in the inter-cluster distance.

To determine whether the obtained clusters were composed by random rearrangements, we first analyzed the distribution of the breakpoints in the original dataset. After comparing the dispersion of breakpoints in the simulated datasets with the dispersion from the original dataset, we got a p-value of smaller than 1^{-5}, indicating

that the breakpoint locations were not following a random distribution in the cancer genomes. Furthermore, we compared the cluster density in the simulated data and the original dataset finding that the cluster density of the original dataset was unlikely obtained in a random simulation (p-value $<1^{-5}$). Therefore, the clusters we obtained implementing the 2-step KDE clustering contain SVs that are likely mechanically linked and not just random occurrences.

Motif Finding

Using the graph mining technique allowed us to to convert our pattern search across all the genome of every sample in a simpler graph search. Within High Performing Computing environments that are based on Apache HBase [15], HDFS [16] and Spark [17] we are able to distribute the computational load across several machines. We used three machines with an Intel® Xeon(R) CPU E5-2630 v4 @2.20GHz processor, 128MB of RAM, and 20 cores each. Using these technologies, the search across 2392 samples was done in less than a day. The use of High Performing Computing methods becomes crucial for the analysis of simulated datasets, where we must repeat the methodology for 100 simulations.

Here, we only focused on *cycles* limited to a size of 6. The *cycle* with a size of 3, named *triangle*, was the pattern more recurrent across the different cancer samples. Its confidence was almost twice the confidence of the next simplest *cycle*, composed of only 1 edge more (Table 1).

Cycle size	Confidence	Average	Frequency
3	814	4.68	3817
4	417	6.75	2817
5	260	4.04	1051
6	188	44.43	8354

Table 1 Statistical values for the evaluated *cycles*. The values obtained are defined as follows. Confidence, which provides the number of samples that have at least one *cycle* occurrence. Average which refers to the average of the number of *cycles* happening in the samples. And finally, frequency, the sum of all the occurrences of the *cycle* across the whole dataset.

The challenge in the identification of complex patterns is to discern between the distributions of rearrangements that are the sum of random unrelated occurrences from those that are mechanically associated. We measured the significance of the patterns by calculating the Abundance (Δ). All the *cycles* evaluated in this study were overrepresented as shown in Fig. 6 (all *cycles* got positive values of Abundance.) However, as the number of rearrangements of the *cycle* increased, the Abundance decreased, being the *triangle*, the most overrepresented pattern.

Pattern significance across cancer types

Analyzing the behavior of the *cycles* in each cancer type, the abundances differed between tumor types (see Fig. S1). The *triangle* pattern again predominated over the majority of cancers, with the exceptions of Bone-Osteosarc, Kidney-ChRCC, Lymph-CLL, and Uterus-AdenoCA. Furthermore, there are tumor types that were more similar in terms of abundances of particular *cycles*. For example, Bladder-TCC, Bone-Osteosarc, Breast-AdenoCA, Breast-LobularCA, ColoRect-AdenoCA, Eso-AdenoCA, Head-SCC, Kidney-ChRCC, Lung-AdenoCA, Lung-SCC, Ovary-AdenoCA, Panc-AdenoCA, Prost-AdenoCA, SoftTissue-Leimyo,

Stomach-AdenoCA, Uterus-AdenoCA had high Abundance for most of the *cycles*. In contrast, Breast-DCIS, Cervix-AdenoCA, Myeloid-AML, Myeloid-MPN had Abundance = 0 for every *cycle* or almost every *cycle*. This group was clearly composed of cancer types without enough samples or complexity. The rest of the cancer types lied somewhere in the middle, having Abundance values not as high as the first group but not having all of them to 0 either: Biliary-AdenoCA, Bone-Benighm, Bone-Epith, CNS-GBM, CNS-Medullo, CNS-Oligo, CNS-PiloAstro, Cervix-SCC, Kidney-RCC, Liver-HCC, Lymph-BNHL, Lymph-CLL, Panc-Endocrine, Skin-Melanoma, SoftTissue-Liposarc, Thy-AdenoCA.

Characterization of Triangle types

We further characterized the *triangle* pattern since it was the most overrepresented and recurrent across all the samples. Known patterns of structural variants that could coincide with these *triangles* have been described based on the orientation of chromosomal segments at the breakpoints and their associated copy-number alterations. Using these criteria, we subclassified the *triangle* patterns into four different categories: i) Chromoplexy described by Baca et al. [3] where usually there is not DNA gain and even, there could be a minimal loss (balanced rearrangements); ii) Cycles of templated insertions, characterized by copy number gains and inverted rearrangements [5]; iii) Non-canonical chromothripsis, a pattern that was recently described [18], which can involve different chromosomes with frequently inverted rearrangements with oscillating copy-number alterations; iv) The fourth pattern, that we here have called Chromotrikona (from the Greek chromo for chromosome and from the Sanskrit trikona for triangle), do not correspond to any other pattern previously described and is characterized by the presence of frequent inverted rearrangements with no significant gains or losses of DNA.

Once we set the four classes of *triangles*, their abundances were estimated (see Fig. S2). Since we already knew that *triangles* were overrepresented, we expected to have a high abundance in all types. However, we noticed that Chromoplexy and Chromotrikona patterns were the most overrepresented types. These abundance similarities may be generated due to an overlapping of *triangles* of both types, having one or more clusters in common. Since we knew that clusters could have more than one breakpoint, they could be linked to different clusters, forming different *triangles* and therefore, different *triangle* types. We calculated the number of clusters that had in common every pair of *triangles* (see Fig. S3). As expected, Chromoplexy patterns had more common clusters with Chromotrikona patterns. Furthermore, this behavior was also maintained for Cycles of templated insertions and Non-canonical chromothripsis. These results suggest that these patterns could share some underlying properties as they are found in the same genomic regions.

We also performed an analysis of how these *triangle* types were distributed among the different cancer types. We excluded cancer types having less than 10 samples with *triangles* to avoid possible bias due to the low number of samples. The presence of the *triangle* types were heterogeneous across cancer types (Fig. 7). For instance, Chromoplexy was more common than the other *triangle* types in Lymph-BNHL, Uterus-AdenoCA, Panc-AdenoCA, Head-SCC, Ovary-AdenoCA, Prost-AdenoCA, and Breast-AdenoCA, while Cycles of Templated Insertions was predominant in

Bone-Osteosarc or Skin-Melanoma. Chromotrikona predominated only in Kidney-RCC and was the less represented pattern in Bone-Osteosarc, Liver-HCC, Head-SCC, Skin-Melanoma and SoftTissue-Liposarc.

Discussion

The identification and classification of complex patterns in cancer genomes are not well explored. The complexity of the data and the lack of certainty about the relevant cases claims new strategies that allow us to get insights into their underlying role in tumorigenesis.

Here we have proposed a statistical framework to fulfill this gap. First, we used a KDE-based clustering method identifying adjacent SVs that are not independent events but belonged to the same single event. The KDE clustering has been proven to be fast and simple and very suitable for distribution based-clustering tasks without setting a priori number of clusters [19, 20]. Facing the lack of reference complex patterns of SVs to compare with, we presented a statistical approximation to prove that the clusters of SVs were not by chance, indicating that they must be related to each other [21, 22].

For the detection of motifs to identify the complex chromosomal rearrangements, we adapted a graph mining strategy with a measure of significance for each found pattern [23]. Similar motif finding algorithms based on randomizations have been already proved successfully such as FANMODE [24], MODA [25], and NetMode[26]. All these studies agree that the need to apply the methods to both the original and simulated datasets translates into a high computational burden. We used parallelization and HPC tools to decrease the computational cost of the method [27], as well as narrow down the search to patterns of size 6. The selected measure for analyzing the significance of the motifs, the Abundance, is directly related to the z-score of the pattern but normalized, allowing us to compare among different patterns [28].

Taken together, we here present the development and application of a new unbiased methodology for the classification of complex SV patterns in tumor genomes. Applying this method to more than 150 thousand SVs from the PCAWG cohort we could identify existing known patterns, as well as a new pattern (Chromotrikona) composed of three SVs that involves balanced inversions between distinct DNA regions in 2 or 3 chromosomes. This represents a significant step forward towards the understanding of the role of complex structural rearrangements in cancer.

Conclusions

In this study, we presented the development of a new statistical strategy for the classification of complex rearrangements in cancer, which is key to understanding the role and the impact of structural variation in the origin and evolution of tumors. Considering the current expansion of AI approaches for the analysis of complex biological data, this study highlights the necessity to establish robust, unbiased, and accurate statistical frames that are the foundation of more complex machine learning algorithms.

The new strategy proposed in this study fulfilled this end, being composed of an unbiased clustering solution based only on the data distribution, a robust motif finding algorithm that can be easily parallelizable to decrease the computational

cost of such an extensive search and a final statistical measure that accurately ranks the obtained patterns in terms of significance.

The results showed the identification of different known patterns in cancer samples as well as a new pattern not previously described. This recurrent pattern, called Chromotrikona, is defined by inverted rearrangements where there are no significant gains or losses of DNA. The development of methods for studying complex patterns of SVs allows us to have insights into new patterns but also understand the genesis of chromosomal rearrangements without limited resolutions. Such genomic rearrangements are the result of subverted biological processes by which they contribute to cancer development.

Acknowledgements
We acknowledge the access to data from PCAWG Consortium which provided SVs data. We thank the patients and their families for their participation in the ICGC and TCGA projects.

Funding
Not applicable

Abbreviations
- **SV**: Structural Variant.
- **PCAWG**: Pan-Cancer Analysis of Whole Genomes.
- **KDE**: Kernel Density Estimation.
- **ICGC**: International Cancer Genome Consortium.
- **TCGA**: The Cancer Genome Atlas.
- **HPC**: High Performance Computing.

Availability of data and materials
All the data analyzed during the current study are available in the data repositories from ICGC data portal.

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Authors' contributions
GG and JL designed the methodology. GG developed the methods and performed the main statistical analysis. LD characterized the triangle patterns. GG and LD conceived the work and wrote the paper. All authors read and approved the final manuscript.

Author details
[1]Barcelona Supercomputing Center (BSC), Department of Computer Science, Barcelona, Spain. [2]Universitat Politècnica de Catalunya (UPC), Barcelona, Spain. [3]Barcelona Supercomputing Center (BSC), Department of Life Sciences, Barcelona, Spain. [4]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.

References
1. Boveri, T.: Concerning the origin of malignant tumours by theodor boveri. translated and annotated by henry harris. Journal of cell science **121**(Supplement 1), 1–84 (2008)
2. Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D., Lau, K.W., Beare, D., Stebbings, L.A., *et al.*: Massive genomic rearrangement acquired in a single catastrophic event during cancer development. cell **144**(1), 27–40 (2011)
3. Baca, S.C., Prandi, D., Lawrence, M.S., Mosquera, J.M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T.Y., Ghandi, M., *et al.*: Punctuated evolution of prostate cancer genomes. Cell **153**(3), 666–677 (2013)
4. The, I., of Whole, T.P.-C.A., Consortium, G., *et al.*: Pan-cancer analysis of whole genomes. Nature **578**(7793), 82 (2020)
5. Li, Y., Roberts, N.D., Wala, J.A., Shapira, O., Schumacher, S.E., Kumar, K., Khurana, E., Waszak, S., Korbel, J.O., Haber, J.E., *et al.*: Patterns of somatic structural variation in human cancer genomes. Nature **578**(7793), 112–121 (2020)
6. Sheather, S.J.: Density estimation. Statistical science, 588–597 (2004)
7. Kim, J., Scott, C.D.: Robust kernel density estimation. The Journal of Machine Learning Research **13**(1), 2529–2565 (2012)
8. Dowd, K., Severance, C.: High performance computing (2010)
9. Hager, G., Wellein, G.: Introduction to high performance computing for scientists and engineers (2010)

10. Wong, E., Baur, B., Quader, S., Huang, C.-H.: Biological network motif detection: principles and practice. Briefings in bioinformatics **13**(2), 202–215 (2011)
11. Jones, M.C., Marron, J.S., Sheather, S.J.: A brief survey of bandwidth selection for density estimation. Journal of the American statistical association **91**(433), 401–407 (1996)
12. Chiu, S.-T.: Bandwidth selection for kernel density estimation. The Annals of Statistics, 1883–1905 (1991)
13. Kuramochi, M., Karypis, G.: Finding frequent patterns in a large sparse graph. Data mining and knowledge discovery **11**(3), 243–271 (2005)
14. Milo, R., Kashtan, N., Itzkovitz, S., E. J. Newman, M., Alon, U.: On the uniform generation of random graphs with prescribed degree sequences. Tech rep **21** (2004)
15. Team, A.H.: Apache hbase reference guide. Apache, version **2**(0) (2016)
16. Borthakur, D., *et al.*: Hdfs architecture guide. Hadoop Apache Project **53**(1-13), 2 (2008)
17. Spark, A.: Apache spark. Retrieved January **17**, 2018 (2018)
18. Cortés-Ciriano, I., Lee, J.J.-K., Xi, R., Jain, D., Jung, Y.L., Yang, L., Gordenin, D., Klimczak, L.J., Zhang, C.-Z., Pellman, D.S., *et al.*: Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. Nature genetics **52**(3), 331–341 (2020)
19. Matioli, L., Santos, S., Kleina, M., Leite, E.: A new algorithm for clustering based on kernel density estimation. Journal of Applied Statistics **45**(2), 347–366 (2018)
20. Zhou, Z., Si, G., Zhang, Y., Zheng, K.: Robust clustering by identifying the veins of clusters based on kernel density estimation. Knowledge-Based Systems **159**, 309–320 (2018)
21. Odén, A., Wedel, H., *et al.*: Arguments for fisher's permutation test. Annals of Statistics **3**(2), 518–520 (1975)
22. Ojala, M., Garriga, G.C.: Permutation tests for studying classifier performance. Journal of Machine Learning Research **11**(6) (2010)
23. Wong, E., Baur, B., Quader, S., Huang, C.-H.: Biological network motif detection: principles and practice. Briefings in bioinformatics **13**(2), 202–215 (2012)
24. Wernicke, S., Rasche, F.: Fanmod: a tool for fast network motif detection. Bioinformatics **22**(9), 1152–1153 (2006)
25. Omidi, S., Schreiber, F., Masoudi-Nejad, A.: Moda: an efficient algorithm for network motif discovery in biological networks. Genes & genetic systems **84**(5), 385–395 (2009)
26. Li, X., Stones, D.S., Wang, H., Deng, H., Liu, X., Wang, G.: Netmode: Network motif detection without nauty. PloS one **7**(12), 50093 (2012)
27. Kim, W., Diko, M., Rawson, K.: Network motif detection: Algorithms, parallel and cloud computing, and related tools. Tsinghua science and technology **18**(5), 469–489 (2013)
28. Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., Alon, U.: Superfamilies of evolved and designed networks. Science **303**(5663), 1538–1542 (2004)

Figures

Fig. 1 Workflow applied to identify complex rearrangements in PCAWG genomes. Simple data pre-processing was performed before implementing the recursive clustering. Then, the graph mining method was applied to find patterns. Finally, the motif finding strategy was applied to determine the statistically significant patterns

Fig. 2 Kernel Density Estimation of breakpoint clusters from chromosome 3 setting bandwidth values of (a) 3000 and (b) 8000. Blue dots represent the locations of the breakpoints, the blue line is the kernel density estimation and red lines the obtained cluster peaks. The inter and intra-cluster distances are shown in green and red, respectively

Fig. 3 Circular representation of human genome with *cycles* of different sizes

Fig. 4 Graphic representation of the subgraph mining process. We performed the search for every vertex of the sample until every possible connection of size 6 was found. Since we did not implement any control during the algorithm, every pattern was likely to be found more than one time and had to be reduced in the following step. This method allowed us to parallelize the search in several machines to reduce computational time

Fig. 5 Total inter and intra-cluster distances for the whole dataset using the 2-step KDE clustering with different bandwidth values

Fig. 6 Abundance values for the analyzed *cycles*. Its value can go from -1, underrepresented, to +1, overrepresented. The Abundance of a single rearrangement (1 SV) is also shown as a control value. Its value is 0 since we fix the rearrangements during the simulation of the random datasets, which means that its representation is the same in every dataset

Fig. 7 Confidence intervals of the mean of the frequency for each *triangle* type throughout cancer types. Only cancers with more than 10 samples with *triangles* were showed

Additional Files

Additional file 1 (pdf)

Figures S1 showing the abundance values of the evaluated cycles for the 36 cancer types. Figure S2 showing the abundance values for different triangle categories and Figure S3 showing common clusters between every pair of triangle types.

Figures

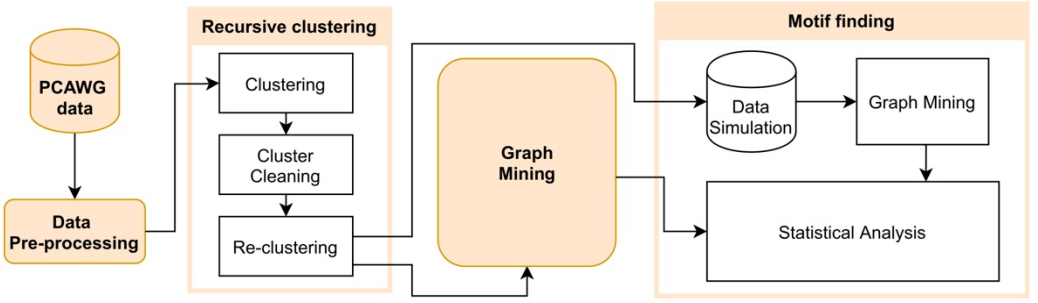


Figure 1

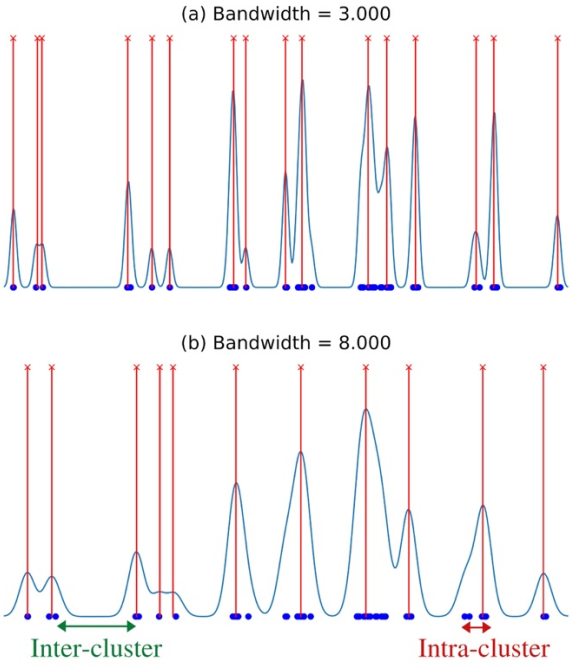


Figure 2

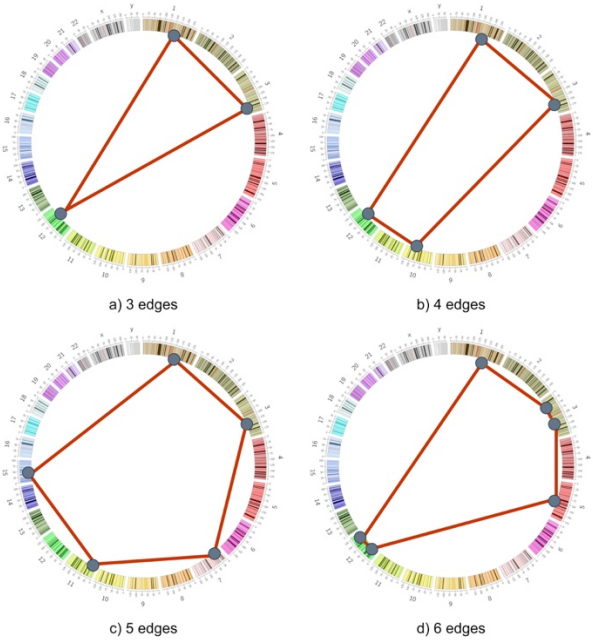


Figure 3

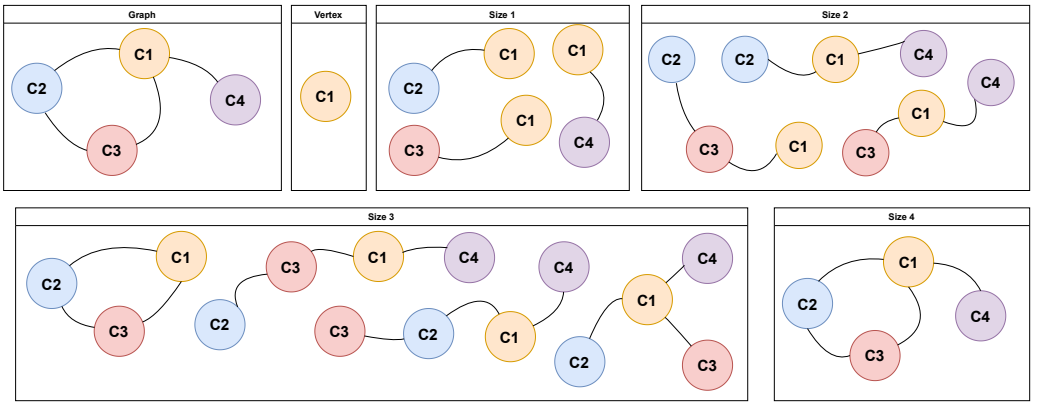


Figure 4

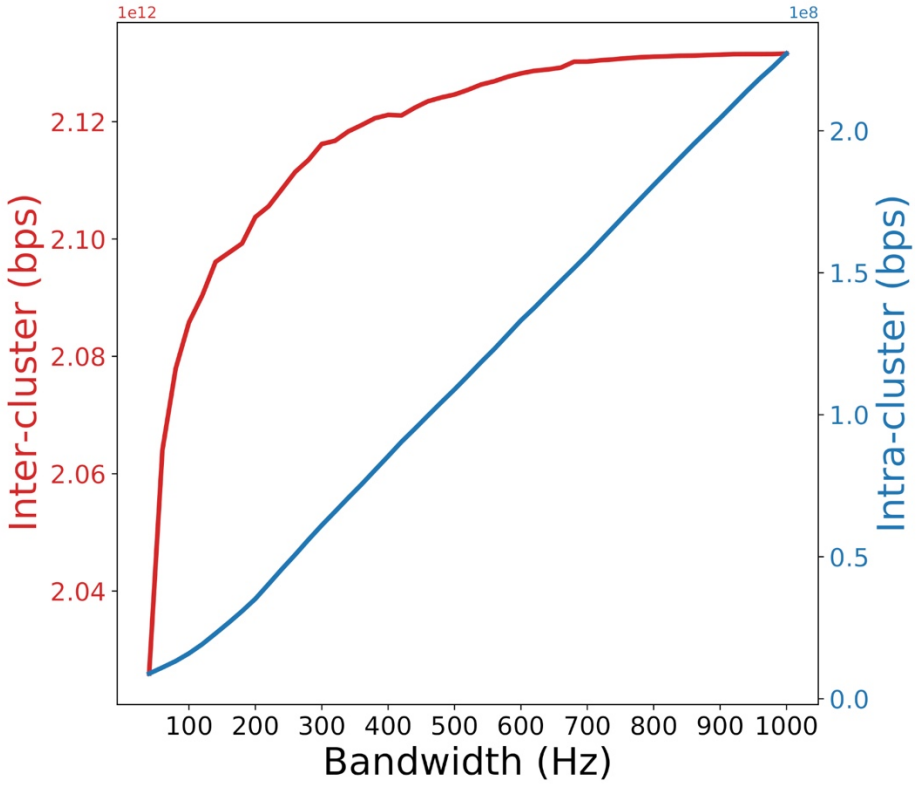


Figure 5

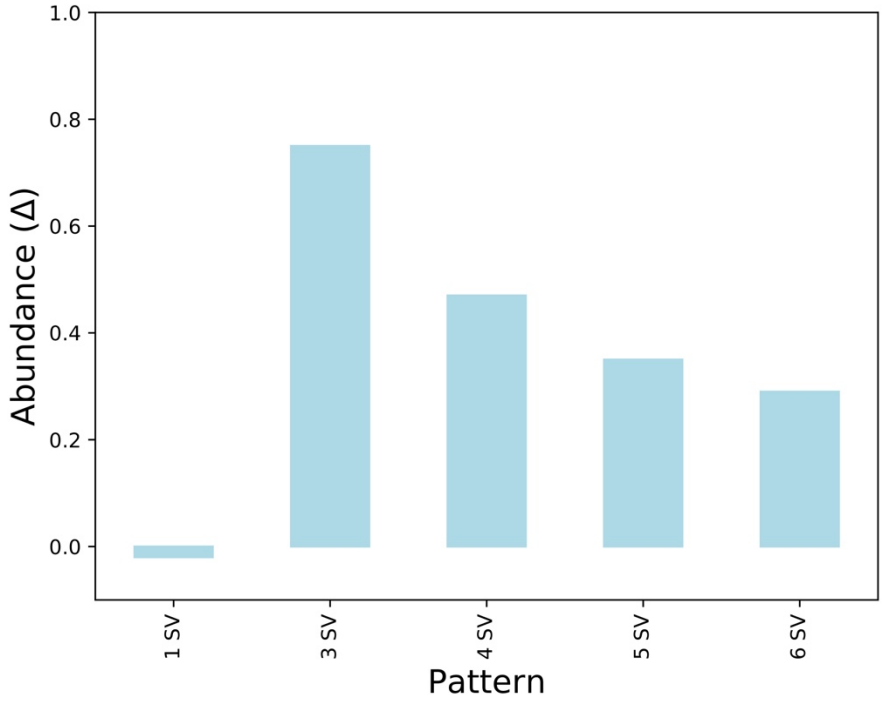


Figure 6

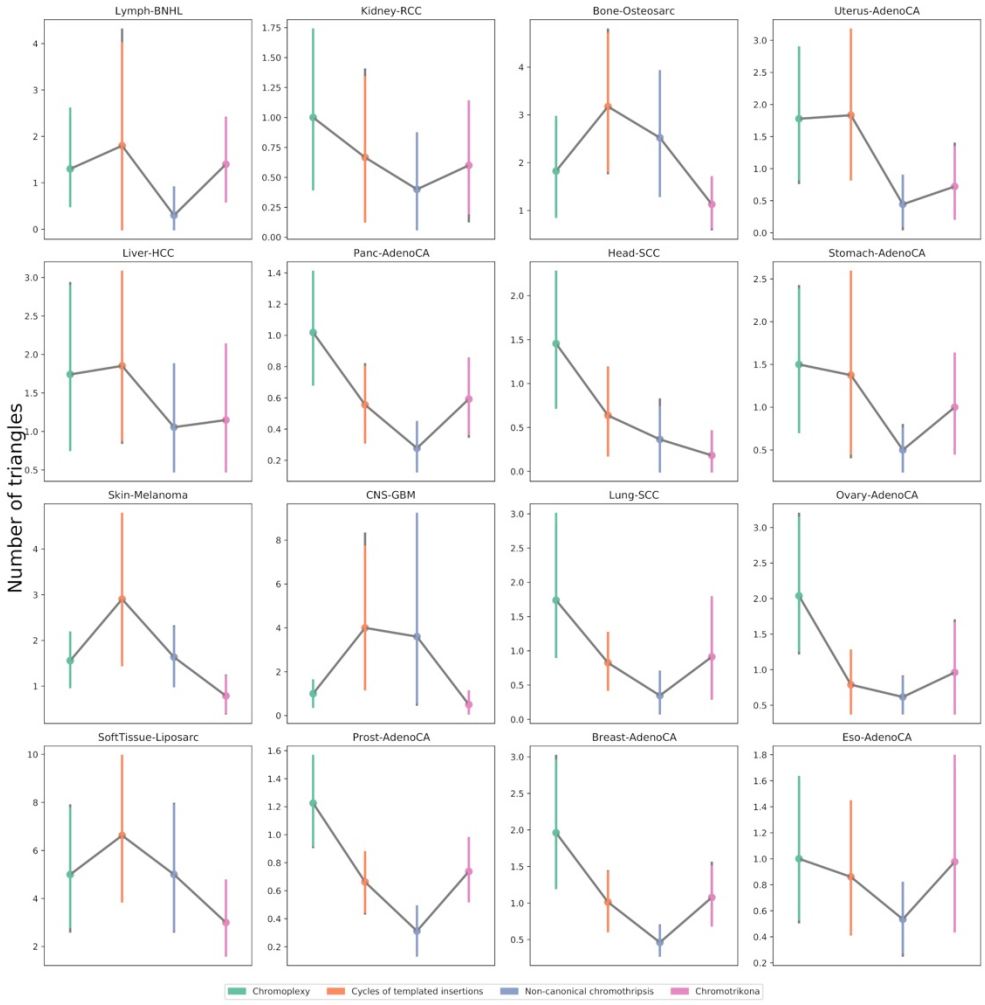

Lymph-BNHL
Kidney-RCC
Bone-Osteosarc
Uterus-AdenoCA
Liver-HCC
Panc-AdenoCA
Head-SCC
Stomach-AdenoCA
Skin-Melanoma
CNS-GBM
Lung-SCC
Ovary-AdenoCA
SoftTissue-Liposarc
Prost-AdenoCA
Breast-AdenoCA
Eso-AdenoCA
Number of triangles
Chromoplexy
Cycles of templated insertions
Non-canonical chromothripsis
Chromotrikona

Figure 7

UNIVERSITAT DE BARCELONA