TEL AVIV אוניברסיטת
UNIVERSITY תל אביב

Tel-Aviv University

Raymond and Beverly Sackler Faculty of Exact Sciences

The Blavatnik School of Computer Science and AI

Integrating mutation, copy number, and gene expression data to identify driver genes of recurrent chromosome-arm losses

Thesis submitted in partial fulfillment of graduate requirements for
the degree "Master of Sciences" in Tel-Aviv University
School of Computer Science and AI

By

Ron Saad

Under the supervision of

Prof. Ron Shamir

Prof. Uri Ben-David

December 25

Acknowledgments

I would like to express my sincere gratitude to the people who supported me during the course of this work and helped me bring it to completion.

First and foremost, I thank my supervisors, Prof. Ron Shamir and Prof. Uri Ben-David, for their guidance, encouragement, and invaluable insights throughout this research. Their mentorship has taught me invaluable lessons in conducting rigorous and high-quality scientific research. I feel privileged to have been mentored by them. Their patient, thoughtful, and excellence-driven guidance has made this journey deeply meaningful and has profoundly shaped my development as a researcher.

I am also sincerely grateful to all the members of the Shamir and Ben-David labs. Your collaborative spirit, thoughtful discussions, and constructive feedback created a supportive environment that made this work both enjoyable and enriching.

I would like to extend my thanks to my family for their constant encouragement, patience, and confidence in me. Their support has been a constant source of strength throughout this work.

Finally, I would like to gratefully acknowledge the Edmond J. Safra Center for Bioinformatics at Tel Aviv University for supporting me with a fellowship during my studies. Additional thanks to the funding agencies that supported this study: the European Research Council Starting Grant (grant #945674 to U.B.-D.), the Israel Science Foundation (grants #1805/21 to U.B.-D. and #2206/22 to R. Shamir), the Israel Cancer Research Fund (project grant to U.B.-D.), and the U.S.-Israel Binational Science Foundation (grant #2023236 to U.B.-D.).

Abstract

Cancer arises through the progressive accumulation of genetic and epigenetic alterations. Despite being one of the most common genomic abnormalities in cancer, aneuploidy – gain or loss of entire chromosomes or chromosome arms – remains poorly understood with respect to its tumor-promoting roles. In this work, we present a systematic framework integrating mutation, copy number, and gene expression data to identify candidate driver genes of cancer type-specific recurrent chromosome-arm losses across 20 cancer types, using ~7,500 tumors from The Cancer Genome Atlas.

By analyzing focal deletions and point mutations that co-occur, or are mutually exclusive, with chromosome-arm losses, we pinpoint 322 candidate drivers associated with 159 recurring events. Our approach identifies known aneuploidy drivers such as *TP53* and *PTEN*, while revealing multiple additional candidates, including tumor suppressors not previously linked to aneuploidy.

We leverage expression changes associated with chromosome-arm losses to propose cancer-promoting pathway-level alterations. Integrating these findings highlights key candidate drivers that underlie the observed expression alterations, reinforcing their biological relevance. We also provide a comprehensive catalog of candidate driver genes for recurrently lost chromosome-arms in human cancer.

Table of Contents

Chapter 1: Biological Background

In this chapter, we present the basic biological concepts that are relevant to this work.

1. Genetic Basis of Cancer Development

1.1 DNA Structure and Function

Deoxyribonucleic acid (DNA) is the fundamental hereditary material in all living organisms, encoding the instructions necessary for development, cellular function, and reproduction. DNA is composed of two antiparallel strands forming a double helix, consisting of nucleotide subunits: adenine (A), thymine (T), cytosine (C), and guanine (G). These bases pair specifically (A-T and C-G), allowing for the accurate replication and transcription of genetic information. The genetic code carried by DNA is transcribed into messenger RNA (mRNA), which then guides the synthesis of proteins, the molecular machines that perform most cellular functions.

1.2 Chromosomes

In all nucleated cells, most of the genomic material is organized into several very long DNA molecules called chromosomes. In human cells, DNA is arranged into 23 pairs of chromosomes, each containing thousands of genetic elements that control cellular processes. Each pair consists of one chromosome inherited from one of the parents, together forming two haploid sets of chromosomes. A single haploid set, containing 23 chromosomes, represents the genetic contribution from one parent. Each chromosome has a specialized region known as the centromere, which is essential for the accurate segregation of genetic material during cell division (described below). The centromere divides the chromosome into two distinct regions called arms: the short arm, referred to as the p arm, and the long arm, referred to as the q arm. The integrity and proper regulation of the DNA sequence are vital for maintaining cellular homeostasis[1].

1.3 Genes

A gene is a specific sequence of DNA that contains the information required to produce a functional product, typically a protein or functional RNA molecule. Genes are composed of coding regions (exons) interspersed with non-coding regions (introns), and are accompanied by regulatory DNA elements such as promoters, enhancers, and silencers that control when a gene is expressed. Proteins are composed of sequences built from a 20-letter alphabet of amino acids. The exons of protein-coding genes consist of nucleotide triplets, or codons, each specifying a particular amino acid. During gene expression of protein-coding genes, the DNA sequence of a gene is transcribed into mRNA by RNA polymerase. The mRNA is then translated into a protein by ribosomes, according to the genetic code. Not all genes encode proteins, some produce non-coding RNAs, such as microRNAs and long non-coding RNAs, which play critical regulatory roles in gene expression. The spatial and temporal regulation of gene expression ensures that different cell types maintain distinct identities and functions. Aberrations in gene regulation, such as mutations or epigenetic alterations, can disrupt normal cellular processes and contribute to diseases, including cancer[1].

1.4 Cell Cycle and DNA Replication

The cell cycle is the tightly regulated process by which a cell grows, duplicates its DNA, and divides into two daughter cells[2]. It consists of four main phases: G_1 (cell growth), S (DNA synthesis), G_2 (preparation for mitosis), and M (mitosis) (**Figure 1**). Accurate duplication of the genome occurs during the S phase, where each chromosome is replicated to ensure equal genetic material is passed on to daughter cells.

Mitosis, the division of the nucleus during the M phase, occurs in several steps. In prophase, the duplicated chromosomes condense and become clearly visible under a microscope. At this stage, a structure called the mitotic spindle begins to form - this is a network of protein fibers that helps separate the chromosomes later in mitosis. Each chromosome develops a kinetochore, a protein complex located at the centromere, which serves as the attachment point for the spindle fibers. During metaphase, the chromosomes align in the center of the cell, ensuring that each kinetochore is attached to fibers pulling toward opposite sides. In anaphase, the two identical copies of each chromosome, called sister chromatids, are pulled apart to opposite poles of the cell. Finally, in telophase, new nuclei form around the separated chromosome sets, followed by cytokinesis, where the cell physically splits into two new daughter cells.

The cell cycle has critical checkpoints that ensure proper division, particularly at G_1/S, G_2/M transitions, during metaphase-monitor DNA integrity, replication completeness, and chromosome alignment. Failures in these checkpoints can lead to errors such as DNA damage or chromosome missegregation, setting the stage for genomic instability and aneuploidy (detailed below), which are hallmarks of many cancers.

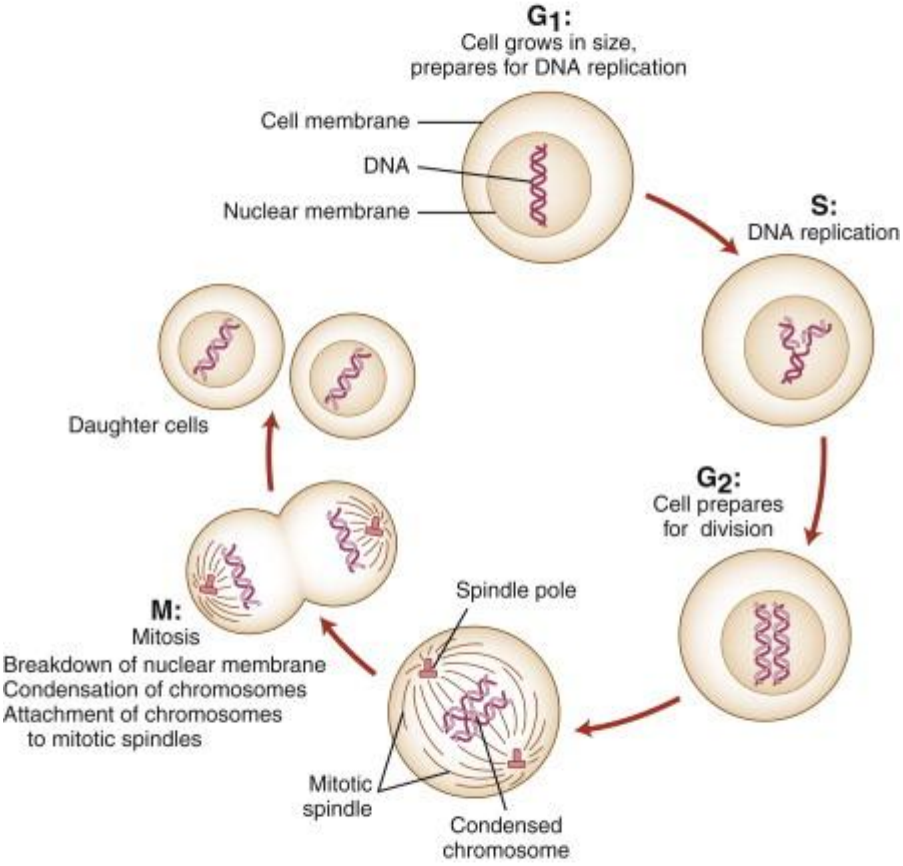

G1:
Cell grows in size, prepares for DNA replication
Cell membrane
DNA
Nuclear membrane
S:
DNA replication
G2:
Cell prepares for division
Spindle pole
Mitotic spindle
Condensed chromosome
M:
Mitosis
Breakdown of nuclear membrane
Condensation of chromosomes
Attachment of chromosomes to mitotic spindles
Daughter cells

Figure 1: Schematic overview of the cell-cycle phases (G_1, S, G_2, M). (Source: Human Genes and Genomes, Chapter 16 *The Genetics of Cancer*[2])

1.5 Cancer Evolution through Genetic Alterations

Cancer arises through an evolutionary process driven by genomic and epigenetic alterations[3]. These genomic alterations take several forms. Point mutations (PMs) involve the substitution of a single nucleotide. Insertions and deletions (indels) are the addition or removal of short DNA fragments. Structural variants (SVs) represent larger rearrangements of the genome, including deletions, duplications, inversions, and translocations. Copy number alterations (CNVs) are a class of SVs that involve gains or losses of large DNA regions, which can affect a single gene or, in some cases, entire chromosome arms or chromosomes. These genomic events enable cells to acquire hallmark cancer capabilities such as sustained proliferation, evasion of apoptosis, and immune escape. Accumulation and selection of these genomic alterations underpin tumor initiation, progression, and treatment resistance (**Figure 2)**.

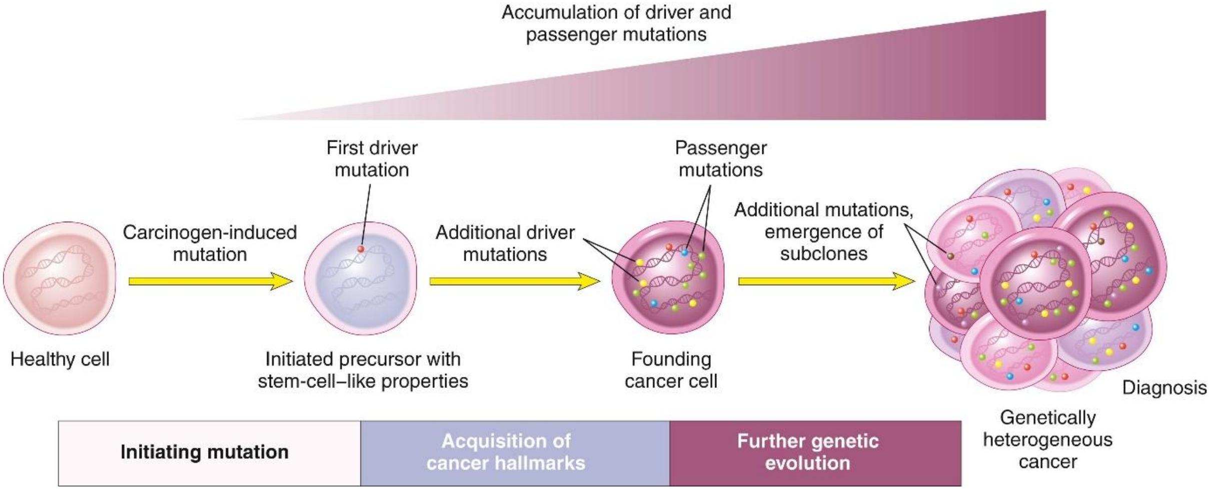

Accumulation of driver and passenger mutations
First driver mutation
Passenger mutations
Additional mutations, emergence of subclones
Carcinogen-induced mutation
Additional driver mutations
Healthy cell
Initiated precursor with stem-cell–like properties
Founding cancer cell
Diagnosis
Genetically heterogeneous cancer
Initiating mutation
Acquisition of cancer hallmarks
Further genetic evolution

Figure 2: Cancer evolution through accumulation of genomic alterations. (Source: https://clinicalpub.com/neoplasia/)

1.6 Aneuploidy

Aneuploidy is classically defined as a deviation from the normal chromosome number, resulting in a count that is not an exact multiple of the haploid set, due to whole-chromosome gains or losses. However, in cancer genomics, the term is often expanded to include copy number alterations of chromosome arms. It is important to note that the mechanisms leading to whole-chromosome mis-segregation differ significantly from those causing arm-level copy number changes. Yet, when studying the consequences of aneuploidy in cancer, these alterations are often analyzed together, since they both reflect large-scale event involving the gain or loss of hundreds or thousands of genetic elements with shared downstream effects. Aneuploidy is generally unfavorable for cellular fitness and viability in most biological contexts. However, it is well tolerated in cancer, prevalent in approximately 90% of solid tumors[4].

1.7 Mechanisms Leading to Aneuploidy

Aneuploidy results from chromosomal instability (CIN), a condition where cells frequently missegregate chromosomes during cell division[5]. Several key mechanisms contribute to CIN (**Figure 3**):

Spindle Assembly Checkpoint (SAC) Deficiencies: The SAC is a critical regulatory system that ensures chromosomes are properly attached to the spindle apparatus before anaphase onset. Defects in SAC components, such as mutations or misregulation of proteins like MAD2 or BUBR1, can lead to premature progression through mitosis despite erroneous chromosome attachments, resulting in missegregation and aneuploidy.

Chromosome Cohesion Defects: Proper cohesion between sister chromatids is essential for accurate chromosome segregation. Proteins like separase and securin regulate the separation of sister chromatids during anaphase. Malfunctions in these proteins can cause premature or delayed separation, leading to missegregation events and contributing to CIN.

Kinetochore-Microtubule Attachment Errors: Accurate chromosome segregation depends on proper attachments between kinetochores and spindle microtubules. Errors such as merotelic attachments, where a single kinetochore attaches to microtubules from both spindle poles, can occur. While these errors often evade detection by the SAC, they can lead to lagging chromosomes during anaphase, resulting in aneuploidy.

Centrosome Amplification and Multipolar Spindle Formation: Centrosomes serve as the primary microtubule-organizing centers in cells. Amplification of centrosomes can lead to the formation of multipolar spindles, increasing the likelihood of unequal chromosome segregation. Cells often attempt to cluster extra centrosomes to form a pseudo-bipolar spindle, but this process is error-prone and can contribute to CIN.

While CIN has been studied extensively, the precise roles of aneuploidy in cancer initiation and progression are still not fully understood.

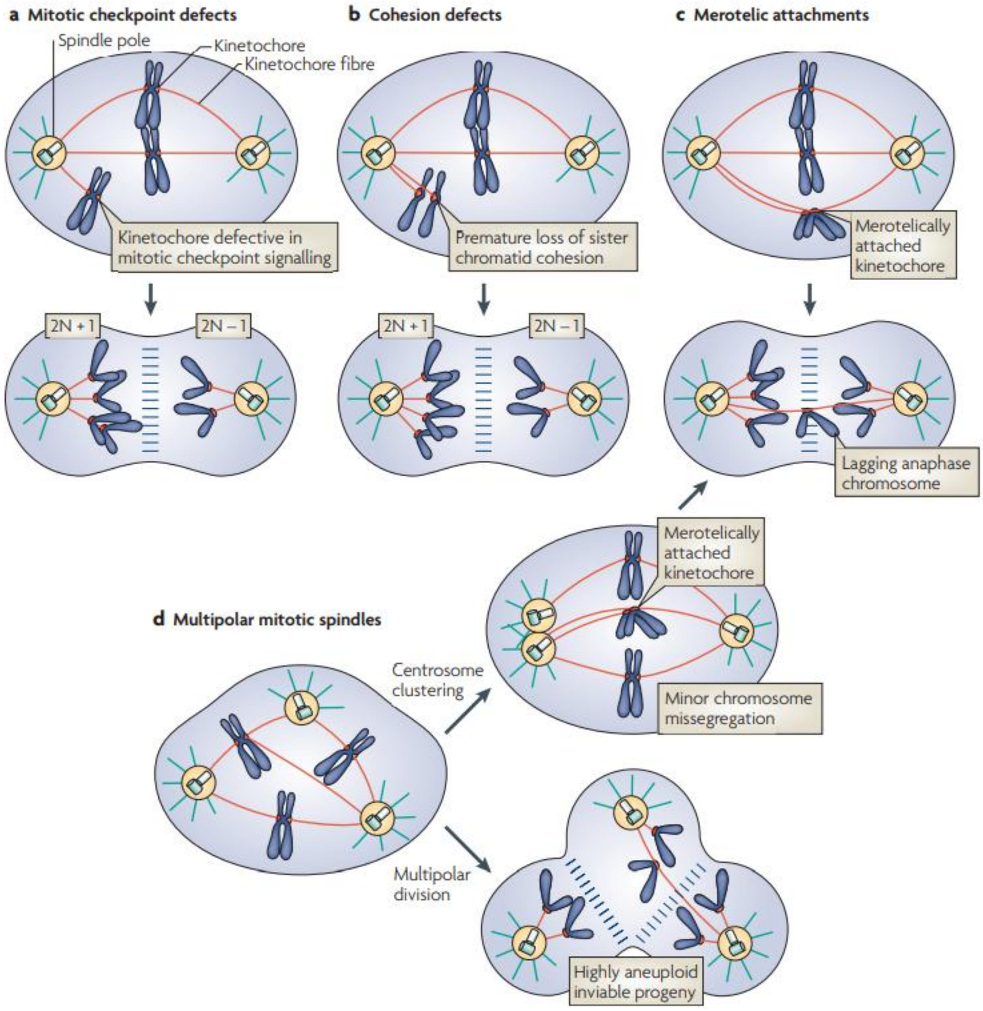

a Mitotic checkpoint defects
Spindle pole
Kinetochore
Kinetochore fibre
Kinetochore defective in mitotic checkpoint signalling
2N + 1
2N – 1
b Cohesion defects
Premature loss of sister chromatid cohesion
2N + 1
2N – 1
c Merotelic attachments
Merotelically attached kinetochore
Lagging anaphase chromosome
d Multipolar mitotic spindles
Centrosome clustering
Merotelically attached kinetochore
Minor chromosome missegregation
Multipolar division
Highly aneuploid inviable progeny

Figure 3: Mechanisms underling aneuploidy formation. (Source: Holland & Cleveland (2009)[5])

1.8 Aneuploidy Landscape in Cancer

An important observation made in recent years is that the role of specific alterations is highly context-dependent, influenced by factors such as tumor stage, cell type, and interactions with the immune system[6]. These factors contribute to the selection pressures that shape the complex landscape of aneuploidy in cancer. The recurrence of particular alterations within a given context suggests that these recurrent changes are positively selected, indicating their potential role in driving cancer initiation and progression.

2. Cancer Driver Genes

2.1 Definition and Identification of Cancer Driver Genes

Cancer driver genes harbor alterations conferring selective advantages essential for tumor initiation and progression. These genes can be oncogenes (OGs), whose overactivation in cancer, typically through amplification or gain-of-function mutations, promotes proliferation, survival, or other tumorigenic traits[7]. Examples include *MYC*, *KRAS*, and *EGFR*. Alternatively, they can be tumor suppressor genes (TSGs), which normally function to restrain cell growth, repair DNA damage, or induce apoptosis. Their inactivation, often via deletion, mutations, or epigenetic silencing, removes critical brakes on cell proliferation. Prominent TSGs include *TP53*, *RB1*, and *PTEN*.

In contrast, passenger genes are those affected by alterations that are not directly involved in cancer development. The identification of driver genes relies on multiple criteria, recurrence across independent tumors, functional validation using experimental models, and integrative computational analyses that incorporate genomic, transcriptomic, and epigenetic data.

2.2 Aneuploidy Driver Genes

The genes that drive common aneuploidies (hereafter referred to as ‘drivers’) remain poorly characterized, with only a handful of cases in which candidate driver genes have been demonstrated to underlie the recurrence of a specific aneuploidy. Even when specific strong drivers were identified, whether additional genes on the same arm also contribute to aneuploidy recurrence remains largely unknown. Identifying the elements that underlie the positive selection for specific recurring aneuploidies is highly challenging for several reasons[8]. First, each such event impacts hundreds or thousands of genetic elements, making it difficult to pinpoint the actual drivers. Every arm harbors multiple TSGs and oncogenes, but not every cancer-related gene would necessarily be the driver of an arm-level copy number alteration. Second, the effects of these alterations are highly context-dependent, making them difficult to study and to model accurately. Third, generating cancer models with specific aneuploidies remains a significant technical challenge, limiting our ability to investigate the effects of these events comprehensively.

Despite those challenges, there are some known aneuploidy drivers such as *TP53* and *PTEN,* which drive the deletion of chromosome arms 17p and 10q respectively and *MYC,* which drives the gain of chromosome arm 8q[9].

3. Genomic Alteration Profiling in Cancer

3.1 DNA Sequencing

DNA sequencing is a high-throughput assay used to read the content of DNA in a given sample. The most widely used approach today is Next-Generation Sequencing (NGS), named for being the second major generation of sequencing technologies[10] (**Figure 4).** NGS begins with extracting the DNA and breaking it into many short fragments, which then go through a process named library preparation, where short sequences named adaptors are attached to them allowing them to bind to the sequencing chip and to undergo the subsequent amplification and sequencing steps. The attached sequences are then amplified, creating clusters of identical sequences. Finaly, sequencing is preformed by creating a complementary strand for each sequence. At each step of the sequencing process, a fluorescently labelled nucleotide is added to the complementary strand, a laser is used to read which nucleotide was added to every cluster. This way content of each cluster is read position by position.

After sequencing, the resulting short reads are computationally aligned to a reference genome, enabling the detection of nucleotide changes relative to the reference. Additionally, the number of reads mapped to each regions in the DNA can be used to estimate the number of copies of that regions in the sequenced DNA.
In cancer research, sequencing is often performed on both a tumor sample and a matched normal sample from the same patient, allowing the identification of somatic mutations that arise during tumor development and are absent in the normal tissue.

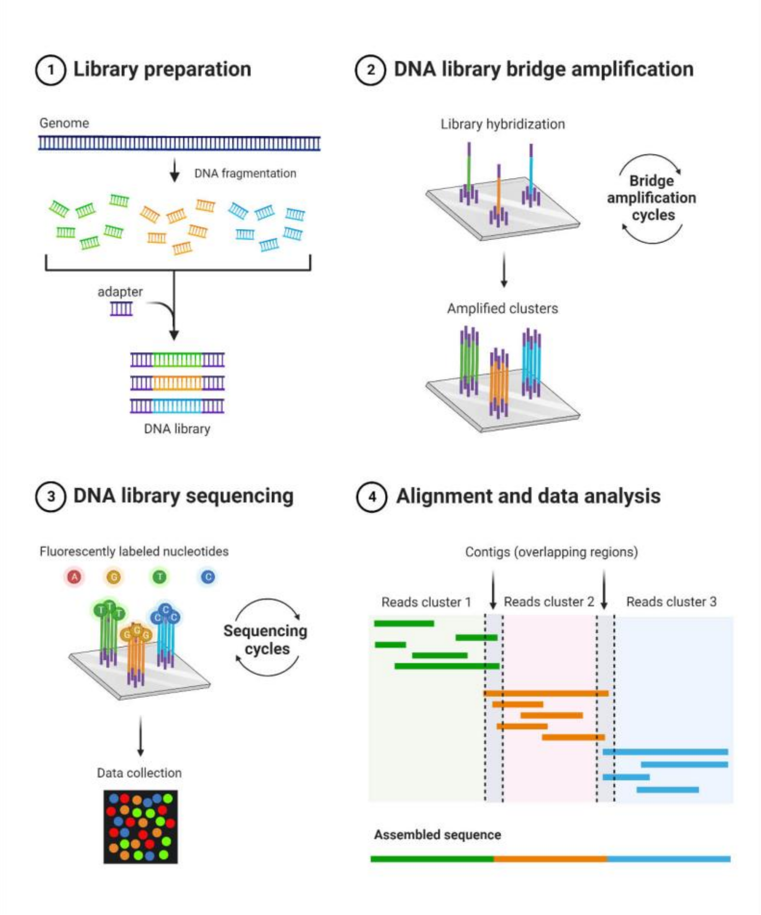


Figure 4: The steps of Next-Generation Sequencing. (Source: https://microbenotes.com/next-generation-sequencing-ngs/)

3.2 SNPs and SNP Arrays

Single-nucleotide polymorphisms (SNPs) are heritable variations at single base-pair positions in the genome that differ among individuals within a population. Each SNP represents a stable base substitution that occurs at a specific genomic locus and is present in a large fraction of the population (typically at least 1%). The vast majority of SNPs in humans present two versions in the population, known as alleles. An individual can have two copies of one allele, making him homozygous to that allele, or one copy of each, making him heterozygous.

A SNP array is an assay used to measure which alleles are present and how many copies of each are found across hundreds of thousands to millions of SNP sites in an individual's genome[11] (**Figure 5**). During the assay, the sample DNA is broken to small fragments which are then fluorescently labelled and matched against short DNA sequences known as probes. Each probe matches one allele of a single SNP, and each spot on the SNP array contains many identical copies of one probe. The DNA matching relies on the concept of DNA hybridization, a process in which single-stranded DNA fragments bind to complementary sequences. The closer the match between the sample DNA and the probe sequence, the stronger and more stable the binding. After hybridization and washing (to remove unbound or weakly bound DNA), the array is scanned using a laser. Each probe spot emits a fluorescent signal proportional to the amount of labelled sample DNA bound to that probe. Analyzing the fluorescence intensities of the two spots representing the two alleles of a given SNP can be used to calculate two important factors:

- Log_2 ratio - compares total intensity at each locus (across both spots) to the intensity measured for a normal reference, allowing to detect copy number gains and losses.
- B-allele frequency (BAF) - measures the relative contribution of one allele versus the other, allowing to infer how many copies of each allele are present

In cancer research, the Log_2 ratio of SNPs is a widely used tool for the detection of copy number alterations.

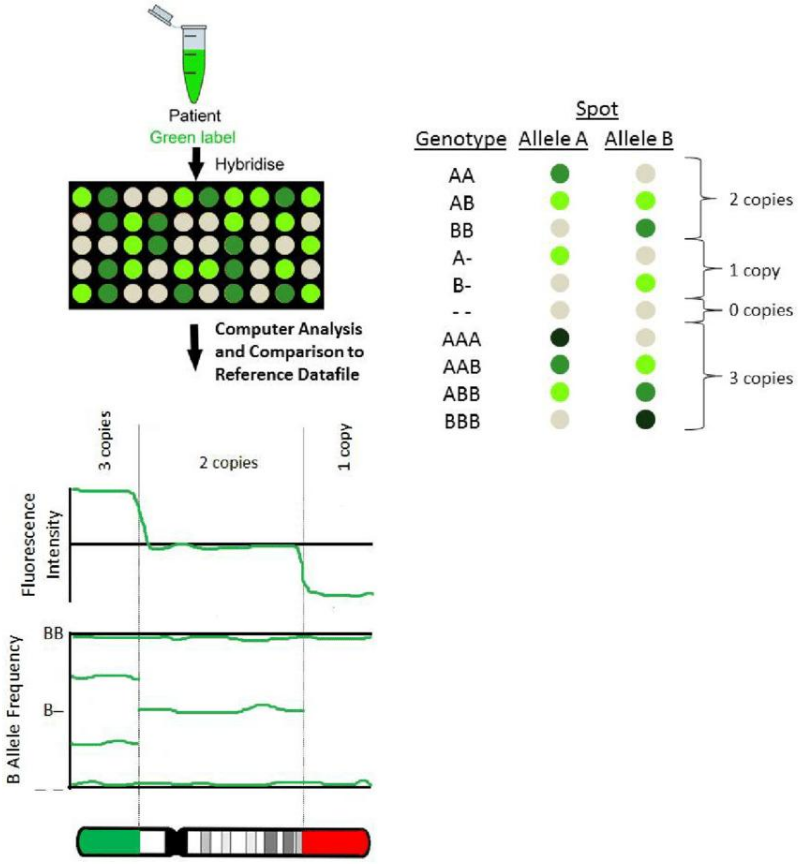


Figure 5: Single Nucleotide Polymorphism (SNP) arrays. (Source: Karampetsou et al. (2014)[12])

4. Gene Expression Profiling and Functional Interpretation

4.1 RNA Sequencing

RNA sequencing (RNA-seq) is a high-throughput technique used to quantify gene expression by directly measuring the abundance of RNA transcripts within a sample. In this approach, mRNA molecules are isolated, converted to complementary DNA (cDNA), and sequenced to produce short reads that are then aligned to a reference genome or transcriptome (as done in DNA sequencing). The number of reads mapped to each gene, commonly referred to as raw counts, serves as a proxy for that gene's expression level (although it is also affected by the gene's length, see below). Because the total number of reads obtained for each sample, referred to as its library size, can vary between sequencing runs or experimental batches, raw read counts must be normalized to ensure that expression levels are comparable across samples. Normalization adjusts for these differences in sequencing depth so that observed variation reflects true biological changes rather than technical artifacts.
In addition, when comparing expression between genes within the same sample, correction for gene length is important, since longer transcripts naturally yield more reads[13].

4.2 Transcriptomic Analyses of Functional Changes

Changes in cellular function can be inferred from alterations in gene expression.
By quantifying RNA abundance for each gene using RNA-seq and comparing gene expression patterns between two or more groups, it is possible to evaluate how cellular activity is affected under different biological or experimental conditions (for example, tumor versus normal tissue, treated versus control samples, or distinct genomic subtypes).These per-gene changes in expression provide a detailed view of the transcriptomic changes associated with genomic alterations, including aneuploidy, and serve as the foundation for uncovering broader functional changes at the pathway and cellular levels[14].

4.3 Gene Sets and Pathways

Individual gene-level changes often provide limited biological insight, as cellular functions are typically carried out by coordinated groups of genes acting within pathways or functional modules. To interpret transcriptomic alterations in a broader biological context, changes in gene expression are analyzed in terms of predefined gene sets - collections of genes that share common roles, regulatory mechanisms, or participation in specific pathways. These gene sets are curated from databases such as Gene Ontology (GO)[15], KEGG[16], Reactome[17].
By evaluating whether such sets show coordinated expression changes, it is possible to infer which biological processes are activated, repressed, or otherwise perturbed under specific conditions. This pathway-level perspective provides a higher level understanding of how genomic alterations, such as aneuploidy, influence cellular behavior.

Chapter 2: Computational Background

In this chapter, we present the main computational methods that are later used in our work.

1. Transcriptomic Analyses of Functional Changes

1.1 Differential Expression Analysis Using DESeq2

DESeq2[14] is a widely used statistical framework for analyzing differential gene expression between sample groups defined by one or more experimental conditions using RNA-seq count data. It models the raw read counts for each gene using a Negative Binomial (NB) distribution.

The core concept behind this modeling is as follows:

For a given sample j, the true expression level of gene i can be represented as the fraction of RNA molecules in the sample transcribed from that gene, denoted q_{ij}. The sequencing process can be viewed as repeatedly drawing RNA molecules at random from the transcript pool, where each molecule has a probability q_{ij} of being from gene i. Given a library size s_j the sample, the expected read count for gene i follows a Poisson process with mean $\mu_{ij} = s_j q_{ij}$. However, due to biological and technical variability between samples with the same experimental conditions, the true proportions q_{ij} vary. Modeling this variability by assuming q_{ij} follows a Gamma distribution gives a Gamma-Poisson mixture, which is equivalent to a Negative Binomial (NB) distribution. Thus, the observed counts are modeled as $K_{ij} \sim NB(\mu_{ij}, \alpha_i)$ where α_i is a gene-specific dispersion parameter describing how much extra variability (overdispersion) exists beyond the Poisson expectation.

The effect of experimental conditions is then modeled using a generalized linear model (GLM). GLMs extend linear regression models, allowing response data to follow other distributions from the exponential family beyond the normal distribution, including the Poison and NB distributions. The GLM has two more components beside the distribution. The linear predictor is a linear function of the dependent variables $\eta = \mathrm{X}\beta$. The link function connects the linear predictor to the mean of the distribution. In the case of the Poisson and NB distributions this function is the log function, i.e. $\log(\mu) = \eta = \mathrm{X}\beta$. In DESeq2 the expected read count μ_{ij} depends on one or more experimental factors x_j (e.g., control vs treatment). Using a log link function, the model is expressed as $log(\mu_{ij}) = log(s_j) + {x_j}^{\top}\beta_i$. The parameters of this model are estimated by maximum likelihood, assuming a NB distribution with the gene-specific dispersion parameter α_i. Since the number of samples is often limited, direct dispersion estimates can be noisy. To address this, DESeq2 employs an empirical Bayes shrinkage approach, which borrows information across all genes to stabilize the dispersion estimates while preserving genuine gene-to-gene differences in variability.

After estimating both the dispersions and the model coefficients, DESeq2 tests for differential expression by evaluating whether the coefficients associated with specific experimental conditions β_i differ significantly from zero using the Wald test. P-values obtained from these tests are adjusted for multiple comparisons using the Benjamini-Hochberg procedure

1.2 Gene Set Enrichment Analysis

Gene Set Enrichment Analysis[18] (GSEA) is a statistical method used to determine whether members of a predefined gene set S have a significant tendency to appear in the top or bottom of a ranked list of genes L.

In practice, genes are usually ranked by their association to one of two phenotypes (e.g., tumor vs. normal or treated vs. control), where association is measured by the $\log_2$ fold change of the gene's expression when comparing samples from the two phenotypes. GSEA then tests whether the genes belonging to a particular set are non-randomly distributed along this ranked list, specifically whether they show a statistically significant tendency to occur toward the top (upregulated) or bottom (downregulated) of the ranking (**Figure 6)**.

Formally, given a gene set S and an assignment of some phenotype correlation statistic r_g to each gene g, the algorithm walks down the list of all genes L ranked by r_g and calculates a running-sum statistic, increasing it when a gene in S is encountered (a "hit") and decreasing it otherwise (a "miss"). The increment when encountering gene $g_j \in S$ is weighted by the absolute value of its correlation statistic $|r_j|^p$ where p is a weighting parameter (usually set to 1), normalized by the sum $N_R = \sum_{g_k \in S} |r_{g_k}|^p$. The decrement for genes not in S is $\frac{1}{|L|-|S|}$.

The running-sum statistic at step i is the sum of "hit" and "miss" steps up to the that step

$$P_{hit}(S,i) = \sum_{g_{j \in S, j \leq i}} \frac{|r_j|^p}{N_R}$$

$$P_{miss}(S,i) = \sum_{g_{j \notin S, j \leq i}} \frac{1}{|L| - |S|}$$

The enrichment score (ES) is defined as the maximum deviation from zero of the running-sum statistic:

$$ES(S) = \max_i [P_{hit}(S,i) - P_{miss}(S,i)]$$

A positive ES indicates that genes in the set tend to occur near the top of the ranked list (upregulated), while a negative ES suggests enrichment at the bottom (downregulated).

The subset of genes that contributes most strongly to the enrichment signal (those appearing before the running-sum statistic reaches its maximum deviation from zero) is termed the leading-edge subset. This subset represents the core genes driving the enrichment, often highlighting the most biologically relevant members of the pathway associated with the observed phenotype.

Statistical significance is assessed by comparing the observed ES to a null distribution generated by calculating the enrichment score for random gene sets of the same size $|S|$. To enable comparison between gene sets of different sizes, each observed ES is divided by the mean of the absolute ES values obtained from its corresponding permutations that have the same sign, yielding a normalized enrichment score (NES). Finally, statistical significance across all tested gene sets is adjusted for multiple comparisons using false discovery rate (FDR), providing a robust estimate of pathway-level significance.

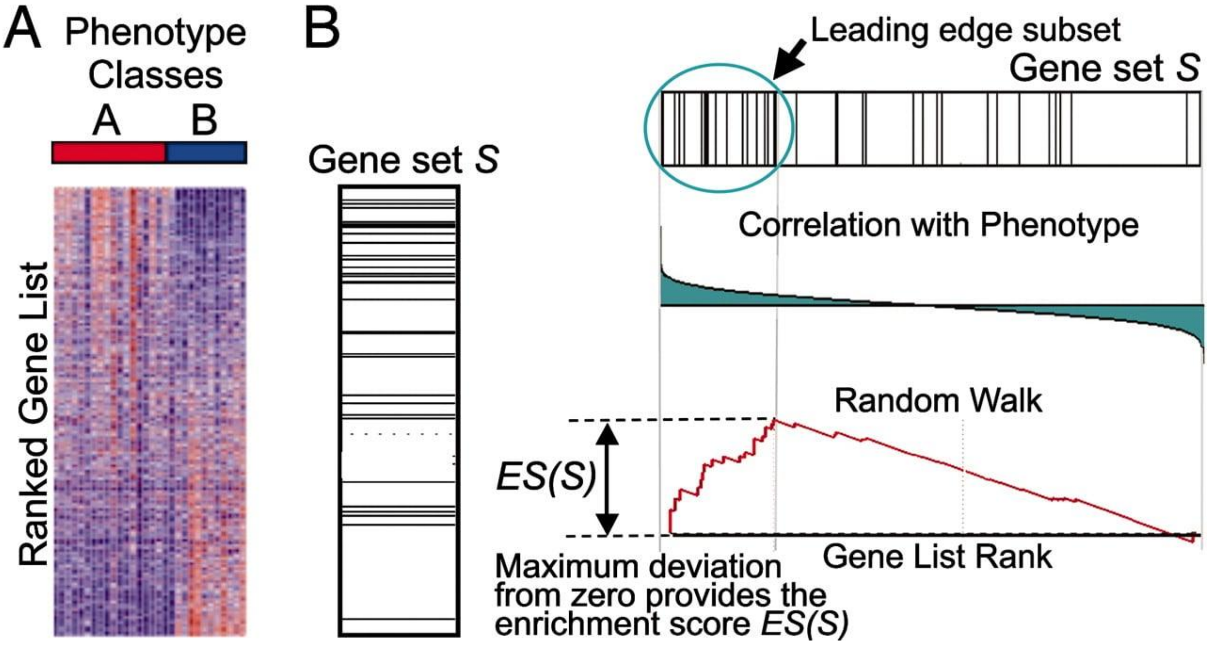


Figure 6: Overview of GSEA. (Source: Subramanian et al. (2005)[18])

2. Recurrence-Based Methods for Cancer Driver Detection

2.1 GISTIC2.0: Detection of Recurrent Copy Number Alterations

GISTIC2.0[19] (*Genomic Identification of Significant Targets in Cancer*) is a probabilistic framework for identifying genomic regions that show statistically significant recurrent somatic copy number alterations (SCNAs) across a cohort of tumor samples. It integrates information about both the amplitude (magnitude of copy number change) and frequency (number of affected samples) of each alteration, while correcting for background chromosomal instability through empirical permutation testing.

The algorithm begins with segmented copy number data derived from SNP arrays or sequencing, where each segment represents a contiguous genomic region with an approximately uniform $\log_2$ copy number ratio between the tumor sample and a reference, typically matched normal.

Each segment is represented as, $S_{i,j} = (chromosome, start, end, value)$,where $value$ is the $\log_2$ ratio for segment j in sample i.

The following steps are then performed

1. Identification of underlying SCNAs using Ziggurat Deconstruction:
   Ziggurat Deconstruction (ZD) models the observed segmented copy number profile of each a chromosome c as the outcome of an underlying sequence of independent SCNA events, collectively denoted h_c, and seeks the most likely history that produced the observed data σ_c (**Figure 7**), which is formulated as $h_c^* = \underset{h_c}{\operatorname{argmax}} Pr(\sigma_c|h_c) +$

$penalty(h_c)$ where the penalty enforces parsimony (implemented via the Bayesian Information Criterion, BIC). ZD models h_c by assuming the chromosome starts from a number of “basal” copy number levels k on which SCNAs act. In practice k is limited to 2, which intuitively can be interpreted as the basal levels of the chromosome arms, though not actually limited by ZD to the arm regions.

σ_c, which is formulated as $h_c^* = \underset{h_c}{\operatorname{argmax}} Pr(\sigma_c|h_c) + penalty(h_c)$ where the penalty enforces parsimony (implemented via the Bayesian Information Criterion, BIC).

For a given SCNA history h_c with k “basal” levels and n breakpoints the BIC is defined as $BIC(h_c) = -2ln\big(Pr(\sigma_c|h_c)\big) + (2k-1)ln(n)$.

Each event is characterized by its length L and amplitude A (its $\log_2$ ratio), and the probability of observing a particular configuration is expressed as the product of independent event probabilities, $Pr(\sigma_c|h_c) = \prod_i f(L_i, A_i)$ where $f(L,A)$ describes the background rate of events with a given length and amplitude. Because $f(L,A)$ is unknown initially, the algorithm first performs a constrained deconstruction in which each breakpoint is attributed to a single event assuming that amplifications and deletions do not alternate (i.e., a gain of a segment is never followed by a loss affecting the same segment and vice versa). In this initial pass, the chromosome profile is peeled back by iteratively merging the most extreme segments toward the baseline copy number, producing an empirical distribution of SCNA events whose smoothed frequencies (with 1% pseudocounts) define the initial $f(L,A)$.

ZD then alternates between two iterative steps: (1) deconstructing each chromosome under the current background model to infer the most likely sequence of events using penalized likelihood, and (2) re-estimating $f(L,A)$ from the pooled set of inferred events across all samples.

2. Probabilistic framework for scoring copy number events:

The probabilistic scoring framework in GISTIC2.0 converts observed focal copy-number alterations covering each genomic position into a likelihood-based score by asking how unlikely those focal events would be under a fitted background model of SCNA formation. To do so, GISTIC 2.0 uses representative loci in the genome known as markers, instead of each individual position. These can be SNPs for SNP array input or other predefined representative loci for sequencing-based input.

First, the SCNA events are classified as focal or broad using a predefined cutoff. Then, each marker is scored by how unlikely it to observe the set of focal events covering it $F_i = \{f_1, f_2, \ldots\}$ given the observed broad events covering it $B_i = \{b_1, b_2, \ldots\}$.

Assuming focal SCNAs are independent, the focal GISTIC score at marker i is defined as $FG_i = -\ln\big(Pr(F_i|B_i)\big) = -\ln\big(\prod_{f\in F_i} Pr(f|B_i)\big) = -\sum_{f\in F_i} ln\big(Pr(f|B_i)\big)$

Each focal event contributes the negative log-probability of observing an event of its length L and amplitude A under the background model. Across large pan-cancer datasets, the probability of observing SCNA of length L and amplitude A decreases approximately inversely with event length (for lengths shorter than a chromosome arm) and decays exponentially with amplitude. Consequently, the probability that any given genomic marker is covered by a focal SCNA of sub-arm length is roughly constant across lengths and can be modelled as $\Pr(f) = \alpha e^{-\alpha A}$ where α is a positive scaling parameter that is fit across all samples (and separately for amplifications and deletions).

Additionally, focal amplifications were largely empirically found to be independent of their underlying arm-level changes, but focal deletions showed a strong dependence on underlying arm-level loss. Specifically, the probability of observing a focal deletion decreases as the magnitude of the arm-level deletion increases. Incorporating this relationship, the conditional probability of observing a focal event with amplitude A affecting a marker, given an arm-level change of amplitude B is:

$$\Pr(f|B) = \begin{cases} \alpha_{amp} e^{-\alpha_{amp} A} & A > 0 \\ (1+B)\alpha_{del} e^{-\alpha_{del} A} & A < 0 \; and \; B > -1 \\ \epsilon \alpha_{del} e^{-\alpha_{del} A} & A < 0 \; and \; B \leq -1 \end{cases}$$

where ϵ is a small constant representing the exceedingly rare cases where the arm-level deletion exceeds one copy in magnitude.

Given this background probability model, the GISTIC score of each marker FG_i in each sample can be calculated. The per-sample scores are then summed across all samples to obtain a cohort-level score for each genomic marker. To evaluate the statistical significance of the score, the null distribution of scores expected by chance is evaluated through random permutation of marker positions across the genome. One-sided p-values are subsequently computed for each marker by comparing the observed scores to this null distribution, and multiple-hypothesis correction is applied using the FDR method.

3. Arbitrated Peel-off:

   The Arbitrated Peel-off algorithm is designed to identify independent focal peaks of recurrent copy number alteration while allowing overlapping SCNAs to contribute proportionally to multiple nearby peaks. In this way, each SCNA can contribute to more than one peak of cross-sample recuring SCNAs, reviling other (less recurrent) such peaks, which may be otherwise missed. Each SCNA i and peak region j are associated with a weight variable w_{ij}, representing the fraction of the SCNA's score attributed to that peak. Initially, all weights are set to zero. Although the exact formulation of the SCNA score s_i is not explicitly detailed, it is likely derived from the same probabilistic background model described earlier for random events, decreasing approximately linearly with event length and exponentially with amplitude. For each chromosome, all markers with q-value below 0.25 are considered for peak identification. If no marker meets this criterion, the markers with the minimal q-value is selected instead. The algorithm proceeds iteratively by selecting, at each step, the most significant unprocessed marker k and identifying the set S of SCNAs that overlap this region, which are then assigned to peaks through two complementary steps:

   1. Uncontested assignment - the scores of all SCNAs not previously assigned to a peak are fully assigned to k, i.e $w_{ij} = s_i$.
   2. Arbitrated assignment - the score of each SCNA that overlaps multiple significant peaks is divided between them, each getting a fraction of the score proportional to the sum of its uncontested score assignments. For each peak c in the group of peaks overlapping the SCNA C_i, a disjoint score D_c is defined as $D_c = \sum_{\substack{l: w_{lr}=0 \\ \forall r \neq c}} w_{lc}$, representing the cumulative weight of SCNAs uniquely assigned to that peak. The score of SCNA i is divided between all peaks in C_i according to $w_{ij} = \frac{D_j}{\sum_{c \in C_i} D_c} s_i$.

Finally, the score of each peak is defined as the sum of all weights assigned to it

$w_j = \sum_i wij.$

4. Determination of boundaries of significantly altered regions using RegBounder:
The final step in the GISTIC2.0 pipeline is the determination of the boundaries of each peak region, which are expected to contain the gene or genes targeted by recurrent copy number alteration. To establish statistical confidence for these boundaries, GISTIC2.0 first constructs a background distribution for the difference between the maximum and minimum marker scores within genomic windows of size n. This distribution is obtained by permuting SCNAs across the genome in each sample. For each identified peak, the algorithm then iteratively expands the region surrounding the peak marker as long as the observed difference between its minimal and maximal marker scores remains below the γth percentile of the background distribution for windows of the same size, for some predefined confidence level γ.

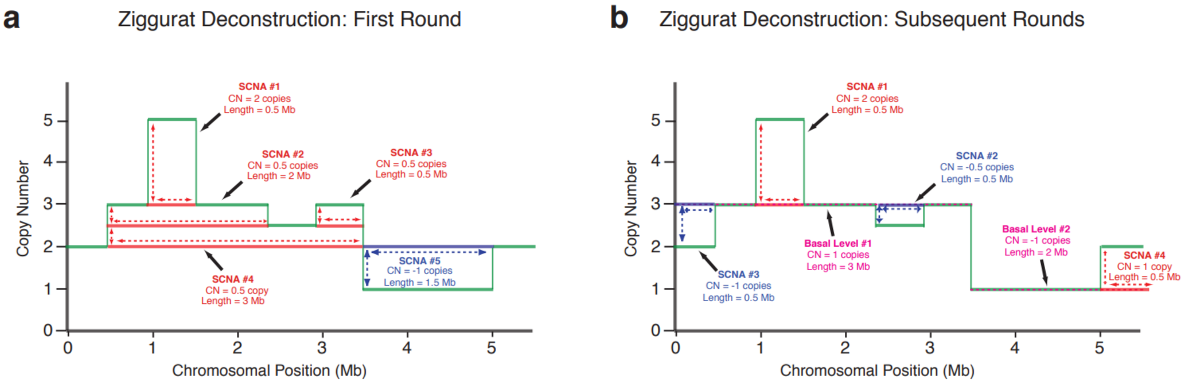


Figure 7: Ziggurat Deconstruction. (Source: Hermel et al. (2011)[19])

2.2 MutSig2CV

MutSig2CV[20] is a statistical framework for identifying genes whose observed somatic point mutation frequency exceeds the expectation based on background mutational processes. The method decomposes mutational significance into three complementary statistical components that capture distinct signals of positive selection.

1. Abundance (MutSigCV):
This component assesses whether a gene exhibits a higher mutation frequency than expected under a context-dependent background mutation rate (BMR) model. The BMR integrates three key sources of variability: (i) mutation category (e.g., transition mutations at CpG dinucleotides), (ii) patient-specific mutation rate differences, and (iii) gene-specific covariates that influence mutability. To account for these effects, genes are first embedded in a covariate space that captures genomic features known to modulate mutation rates. The default covariates are gene expression level, replication timing, and chromatin compartment organization. Each covariate is Z-normalized, and Euclidean distance within this multidimensional space is used to identify the local neighborhood of each gene, consisting of genes with similar mutational covariate profiles.

For each gene g, patient p, and mutation category c, the observed number of non-silent mutations $n_{g,c,p}^{nonsilent}$ is compared against a Beta-Binomial background model:

$$H(n_{g,c,p}^{nonsilent}, N_{g,c,p}^{nonsilent}, x_{g,c,p}, X_{g,c,p})$$

where N is the number of bases covered in the gene, x is the number of background (silent and non-coding) mutations observed in the gene and its neighboring genes, and X is the number of covered bases associated with x.

The mutation categories for gene g in patient p are then ranked by their event likelihood (higher rank means lower likelihood), and the top two categories with at least one observed mutation are used for scoring gene g in patient p.

Scores are aggregated across all patients to compute a gene-level statistic, and statistical significance is determined by comparing the observed score to a null distribution of scores. The null model is first computed for each patient separately, using its category-specific background mutation rate and calculating the probability of observing a score equal of higher to the one achieved for this patient. Then, the gene level p-value is calculated through convolution of all these probabilities, i.e. by calculating the probability of observing a score equal or higher at each patient.

2. Clustering (MutSigCL):
   This component evaluates whether somatic mutations within a gene exhibit significant spatial clustering, indicative of recurrent mutational hotspots that may reflect functional targeting. The analysis focuses exclusively on observed non-silent coding mutations within each gene. To estimate the expected background pattern of mutation placement, the positions of these mutations are randomly permuted, while preserving each mutation's context category, defined by local sequence context. During permutation, indels are allowed to move freely, whereas point mutations are restricted to positions of the same mutational context.
   For each permutation, a clustering score S_{CL} is computed, defined as the fraction of mutations that occur within hotspot regions, which are 3 base-pair windows containing at least two mutations and comprising at least 2% of the total mutations in the gene.
3. Functional conservation (MutSigFN):
   This component assesses whether mutations preferentially occur at evolutionarily conserved sites. For each gene, the same set of non-silent coding mutations and corresponding permutation framework are used as in MutSigCL. For each observed and permuted configuration, a functional score S_{FN} is computed as the mean conservation value of all mutated positions, derived from alignment of 45 vertebrate genomes to the human genome (the UCSC 'phyloP46way' track).

The results of the three tests are integrated to produce a unified gene-level p-value. First, the clustering and conservation components are combined into a joint p-value derived empirically from the joint probability distribution of their permutation-based null scores. Next, this joint p-value is combined with the MutSigCV p-value using the Fisher's method of combining p-values. Finally, multiple hypothesis testing correction is applied using the FDR.

2.3 PRODIGY

PRODIGY[21] is a tool designed for patient-specific ranking of cancer driving mutations. It ranks candidate genes by estimating their cumulative effect on dysregulated pathways, as identified by differential expression analysis.

As a first step, PRODIGY calculates the differential gene expression of each cancer sample by comparing it to a set of normal samples using DESeq2[14]. Genes with an absolute $\log_2$-fold change greater than a threshold β and statistically significant at an FDR below a threshold γ (default: β = 2, γ = 0.05) are then used to identify dysregulated pathways via the hypergeometric test.

Then, the mutated genes in each sample are scored using the following procedure:

Given a protein-protein interaction network $G = (V, E, W)$, where W are edge weights representing interaction confidence, a set of differentially expressed genes DEG and a list of dysregulated pathways for a pathway p represented by the graph $G_p = \left(V_p, E_p\right)$ the effect of each candidate gene g is estimated as follows:

1. Constructing a new graph $G_{p,g} = \left(V_{p,g}, E_{p,g}, W_{p,g}, P_{p,g}\right)$, where:
$$V_{p,g} = V_p \cup \{g\} \cup N\left(V_p\right) \cup N(\{g\})$$
$$E_{p,g} = E_p \cup \left\{(u, v) | u, v \in V_{p,g}\ and\ (u, v) \in E\right\}$$
$$W_{p,g}(u, v) = \begin{cases} 0.1, & u, v \in V_{p,g} \\ 1 - W(u, v), & otherwise \end{cases}$$
$$P_{p,g}(v) = \begin{cases} log(|FoldChange(v)|), & v \in DEG \cap V_p \\ -degree(v)^{\alpha}, & otherwise \end{cases}$$
Here $N(S)$ is the list of direct neighbors of set of vertices S in G and α is a parameter that controls the penalty assigned to nodes with high degree. By default, $\alpha = 0.05$.

2. Finding an approximate solution to the rooted prize collecting Steiner tree problem [22], i.e. a sub-tree $G` = (V`, E`)$ rooted at g maximizing the score:
$$S = \sum_{v \in V\prime} P(v) - \sum_{e \in E\prime} W(e)$$

3. Assigning a normalized effect score by dividing the score S of the solution by the maximal possible score $S_{max} = \sum_{v \in DEG \cap V_p\prime} log(|FoldChange(v)|)$

The final score of each gene, which is used to rank the genes, is defined as the sum of all it's normalized effect scores across all altered pathways.

Chapter 3: Methods and results

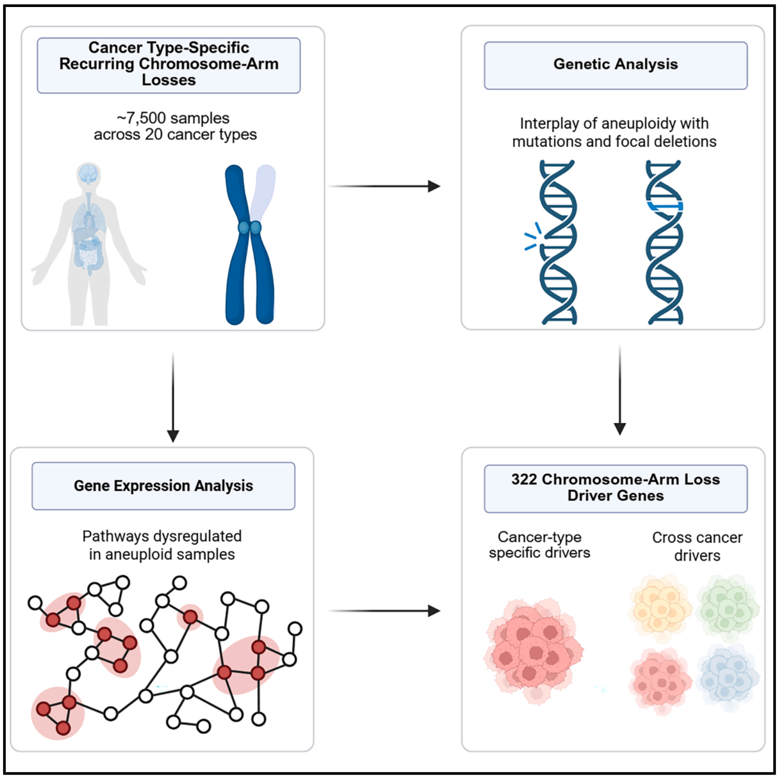
The methods and results of the thesis were published in Cell Reports (2025). The paper is included in the following pages. Supplemental information can be found online at https://doi.org/10.1016/j. celrep.2025.116455.

Cell Reports

Resource

Integrating mutation, copy number, and gene expression data to identify driver genes of recurrent chromosome-arm losses

Graphical abstract


Cancer Type-Specific Recurring Chromosome-Arm Losses
~7,500 samples across 20 cancer types
Genetic Analysis
Interplay of aneuploidy with mutations and focal deletions
Gene Expression Analysis
Pathways dysregulated in aneuploid samples
322 Chromosome-Arm Loss Driver Genes
Cancer-type specific drivers
Cross cancer drivers

Highlights

- A multi-omic approach identifies drivers of recurrent chromosome-arm losses in cancer
- 322 genes nominated as chromosome-arm loss drivers using mutation, copy number, and expression data
- Most drivers are cancer-type specific, but some act as universal aneuploidy drivers
- Integrated pathway analysis links candidate drivers to tumor transcriptional shifts

Authors

Ron Saad, Ron Shamir, Uri Ben-David

Correspondence

rshamir@tauex.tau.ac.il (R.S.),
ubendavid@tauex.tau.ac.il (U.B.-D.)

In brief

Saad et al. present a comprehensive framework that combines mutation, copy number, and expression data to identify 322 candidate genes driving recurrent chromosome-arm losses across 20 cancer types. Additionally, the study links these drivers to altered cellular pathways, providing a resource for understanding how aneuploidy promotes cancer.

Saad et al., 2025, Cell Reports *44*, 116455
November 25, 2025 © 2025 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.celrep.2025.116455

Cell Reports

CellPress
OPEN ACCESS

Resource

Integrating mutation, copy number, and gene expression data to identify driver genes of recurrent chromosome-arm losses

Ron Saad,[1,2,3] Ron Shamir,[1,3,4,*] and Uri Ben-David[2,3,4,5,*]
[1]The Blavatnik School of Computer Science and Artificial Intelligence, Tel Aviv University, Tel Aviv, Israel
[2]School of Medicine, Faculty of Medical & Health Sciences, Tel Aviv University, Tel Aviv, Israel
[3]Edmond J. Safra Center for Bioinformatics, Tel Aviv University, Tel Aviv, Israel
[4]These authors contributed equally
[5]Lead contact
*Correspondence: rshamir@tauex.tau.ac.il (R.S.), ubendavid@tauex.tau.ac.il (U.B.-D.)

https://doi.org/10.1016/j.celrep.2025.116455

SUMMARY

Aneuploidy is a hallmark of cancer, yet the genes driving recurrent chromosome-arm losses remain largely unknown. We present a systematic framework integrating mutation, copy number, and gene expression data to identify candidate driver genes of cancer type-specific recurrent chromosome-arm losses across 20 cancer types, using ~7,500 tumors from The Cancer Genome Atlas. By analyzing focal deletions and point mutations that co-occur, or are mutually exclusive, with chromosome-arm losses, we pinpoint 322 candidate drivers associated with 159 recurring events. Our approach identifies known aneuploidy drivers such as *TP53* and *PTEN*, while revealing multiple additional candidates, including tumor suppressors not previously linked to aneuploidy. We leverage expression changes associated with chromosome-arm losses to propose cancer-promoting pathway-level alterations. Integrating these findings highlights key candidate drivers that underlie the observed expression alterations, reinforcing their biological relevance. We provide a comprehensive catalog of candidate driver genes for recurrently lost chromosome-arms in human cancer.

INTRODUCTION

Aneuploidy, an abnormal number of chromosomes or chromosome arms (hereafter referred to as "arms") in a cell, has long been recognized as a hallmark of cancer.[1] Despite being very common in cancer, occurring in ~90% of solid tumors,[2] its role in cancer initiation and progression is still not fully understood.[3]

Advances in molecular biology research tools, such as high-throughput sequencing and CRISPR, along with improved data collection efforts through comprehensive databases like The Cancer Genome Atlas (TCGA)[4] and Cancer Cell Line Encyclopedia,[5] have significantly enhanced our ability to characterize aneuploidy in cancer. These developments provided deeper insights into the complex landscapes of chromosomal alterations in cancer[2] and the selection pressures that shape them.[1,3,6–12]

An important observation made in recent years is that the role of specific alterations is highly context dependent, influenced by factors such as tumor stage, cell type, and interactions with the immune system.[1,3] These factors contribute to the selection pressures that shape the complex landscape of aneuploidy in cancer. We and others have recently demonstrated the importance of negative selection in shaping the aneuploidy landscapes of human cancer.[6,13] Nonetheless, the recurrence of particular alterations within a given context suggests that these recurrent changes are positively selected, indicating their potential role in driving cancer initiation and progression.[3] Importantly, however, the genes that drive recurring aneuploidies (hereafter referred to as "drivers") remain poorly characterized, with only a handful of cases in which candidate driver genes have been demonstrated to underlie the recurrence of a specific aneuploidy.[10,11,14] Even when strong drivers were identified, whether additional genes on the same arm also contribute to aneuploidy recurrence remains largely unknown.[3]

Identifying the elements that underlie the positive selection for specific recurring aneuploidies is highly challenging for several reasons.[3] First, each event impacts hundreds or thousands of genetic elements, making it difficult to pinpoint the actual drivers. Every arm harbors multiple tumor-suppressor genes (TSGs) and oncogenes, but not every cancer-related gene would necessarily be the driver of an arm-level copy number alteration. Second, the effects of these alterations are highly context dependent, making them difficult to study and model accurately. Third, generating cancer models with specific aneuploidies remains a significant technical challenge, limiting our ability to investigate the effects of these events comprehensively.

Previous studies in this field have focused on identifying characteristics of arms that influence their likelihood of being deleted or amplified. For instance, Davoli et al.[15] demonstrated that the density and potency of TSGs are positively correlated with the frequency of arm loss and negatively correlated with the frequency of its gain, while oncogene density and potency exhibited the opposite trends. We recently corroborated these findings and further identified compensation by paralogs as being positively correlated with an arm's loss frequency.[6] While these studies have provided valuable insights into the relationship between arm features and aneuploidy prevalence, they do not directly address the identification of specific drivers of these events. A recently developed tool, BISCUT,[13] begins to address this challenge by analyzing telomere- and centromere-bound copy-number event distributions. Notably, the systematic approach described in[13] was strictly based on copy-number analyses. We therefore speculated that integrating other genomic modalities, namely point mutations and gene expression, could further improve the identification of candidate aneuploidy driver genes.

In this work, we propose a framework for identifying driver genes of cancer type-specific recurring arm losses. Our analysis leverages mutation, gene expression, and copy number data from ~7,500 tumor samples across 20 cancer types in the TCGA dataset. We examine the relationship between arm loss events and other genomic alterations that can inactivate a given gene, specifically point mutations and focal deletions. These analyses allowed us to identify genes frequently affected both by arm losses and by other mechanisms, suggesting that they function as TSGs whose bi-allelic inactivation involves an arm loss. We also identified genes predominantly affected by a single mechanism, either an arm loss or a focal event but not both. This may indicate that their mono-allelic inactivation provides a selective advantage to cancer, though their bi-allelic loss may not, or that the combined effect of the arm loss and the focal event on closely residing genes is detrimental. Lastly, we used gene expression data to identify changes in pathway activity associated with recurrent arm losses and link these alterations to our candidate driver genes to further refine our list of arm-loss driver genes.

We identified known drivers of specific recurring aneuploidies, such as *TP53*[16–18] and *PTEN*,[19–21] and also nominated additional candidates, overall proposing 322 drivers for 159 cancer type-specific recurring arm losses. Some of these candidates are established TSGs not previously associated with aneuploidy, whereas others do not have an established role in human tumorigenesis. We therefore provide a comprehensive resource of candidate driver genes for all recurrently lost arms in human cancer. Experimental validations will be required to demonstrate the driving role of the newly identified candidate genes.

RESULTS

Approach overview

A schematic representation of our approach for identifying cancer type-specific driver genes associated with recurring arm losses is shown in Figure 1. We first calculated the prevalence of loss of each arm across 20 cancer types, using ~7,500 samples from TCGA. We defined an arm loss as recurring in a specific cancer type if it was lost in >20% of the samples. This analysis identified 230 recurring arm losses across cancer types, hereinafter referred to as chromosome arm-cancer type (CA-CT) pairs (Figure 1A). For each arm and cancer type in the identified CA-CT pairs, we conducted genetic and expression analyses to identify potential driver genes and uncover the consequences of arm loss. The genetic analysis was done by using mutations and copy number data (Figure 1B). To identify drivers of arm loss, we defined four perturbation patterns: (1) Genes frequently affected by a focal deletion when the arm is lost (FD + AL), (2) genes frequently affected by a point mutation when the arm is lost (PM + AL), (3) genes frequently affected by focal deletion only when the arm is not lost (FD-AL), and (4) genes commonly affected by a point mutation only when the arm is not lost (PM-AL). The expression analysis was performed by comparing tumor samples with the arm loss to those without it or to normal samples (Figure 1C). Finally, we summarized the genetic and expression results for each arm across cancer types and perturbation patterns (Figure 1D).

Focal deletions co-occurring with chromosome-arm losses

We began by identifying focal deletions on a particular arm that frequently co-occur with the loss of that arm (pattern FD + AL; Figure 2A). For each CA-CT pair, we applied GISTIC2.0[22] to samples harboring the arm loss, identifying 133 focal deletions across 108 pairs. These deletions result in bi-allelic inactivation of specific chromosomal regions. We found that most genes were rarely affected by both focal and arm-level losses simultaneously (shown for BRCA in Figure 2B and for all other tumor types in Data S1A–S1S). The regions that are frequently co-deleted by both an arm loss and a focal deletion are rather small (median size of ~1.1 Mb, encompassing a median of 11 genes), consistent with a strong negative selection against the bi-allelic loss of large chromosomal regions.[23] In regions with three or fewer genes, we nominated the genes with the highest deletion prevalence as candidate drivers of the combined focal and arm losses. For larger regions, we nominated candidate drivers only if they appeared in a small region in other cancer types (see STAR Methods). It is important to note that we only considered protein-coding genes, although the recurring loss of the region may be driven by a non-coding genetic element.

The full list of genes identified in this pattern analysis is provided in Table S1A. A prominent example is observed in Chr10q in multiple cancer types, including GBM, LUSC, and SARC (Figures 2C–2E and Data S1T). In this case, the bi-allelic deletions affect small regions containing only a few genes, and our analysis identified the known TSG *PTEN* as the culprit of this arm loss. Indeed, *PTEN* was previously reported to drive the loss of Chr10q in GBM and other tumor types.[19–21]

Another interesting example is observed in arm Chr8p (Figures 2F–2H and Data S1U–S1AA). In some cancer types, such as HNSC, the co-deleted region is very small, and a single driver gene, *CSMD1*, was identified (Figure 2F). In other tumor types, such as READ and SARC (Figures 2G and 2H), larger deleted regions that encompass *CSMD1* were identified, with *CSMD1* being the most commonly deleted gene in all cases. This strongly suggests that *CSMD1*—a known TSG in various cancer types[24–26] not linked before to Chr8p loss—is an important (albeit not necessarily sole) driver of this aneuploidy.

Overall, this analysis identified 115 potential driver genes, within 33% of the CA-CT pairs. Of these, in about 65% of the

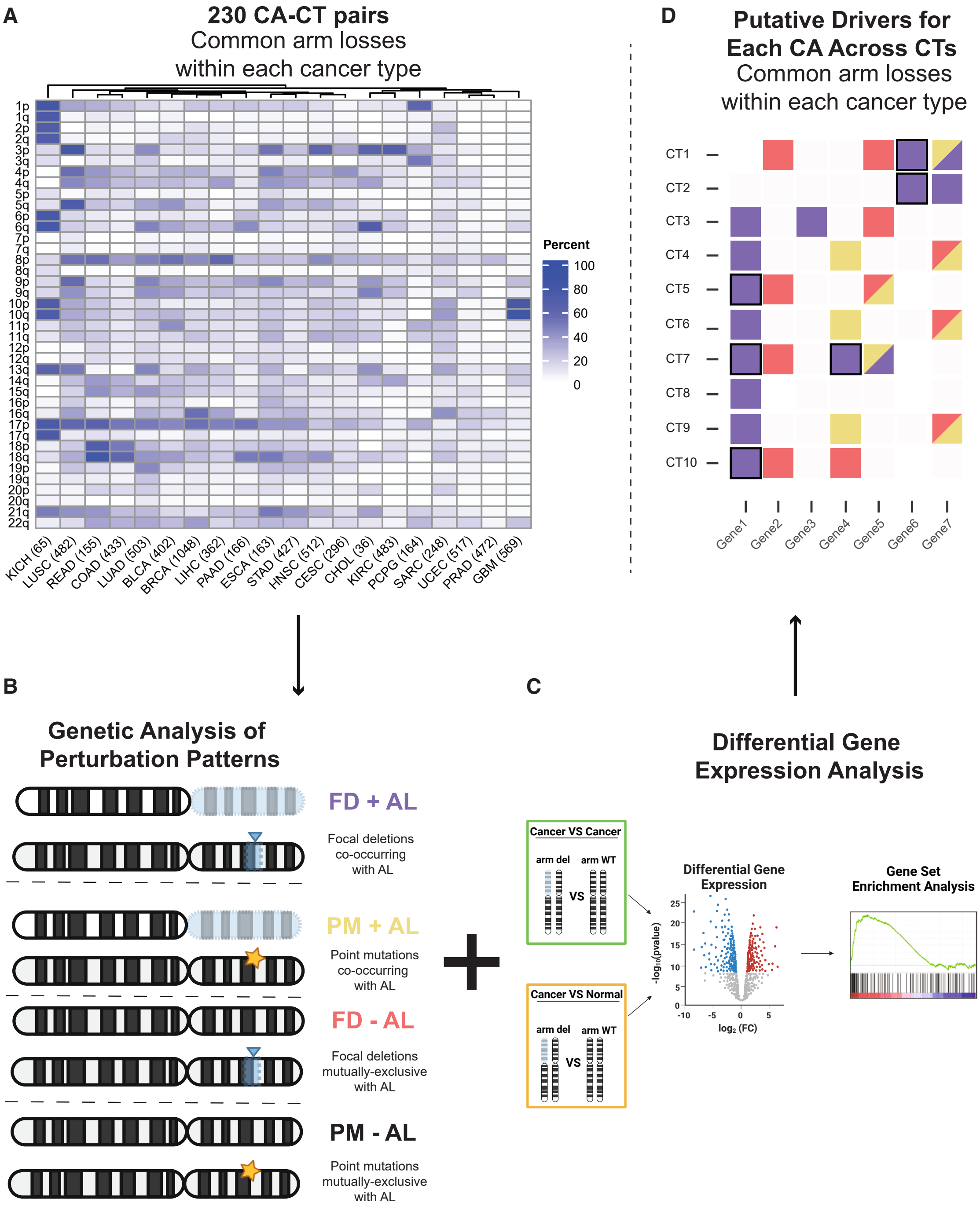


Figure 1. Approach overview

(A) The prevalence of loss of each chromosome arm (CA) across 20 cancer types (CTs) was calculated in ∼7,500 TCGA samples. 230 arms were identified as recurrently lost in specific cancer types.

(B) Genetic analysis was performed on each recurrently lost arm in the corresponding cancer type to nominate drivers of the arm loss.

(C) Expression changes associated with each such arm loss were studied in the corresponding cancer type to identify pathways associated with the arm loss.

(D) Results of these analyses were then integrated across cancer types and visualized. AL: arm loss.

cases, a single gene was nominated, whereas in the rest of the cases, two or three drivers were proposed (Figure 2I; Table S1A). The analysis revealed new potential drivers of recurrent arm losses, such as *CSMD1* in Chr8p, in addition to strong TSGs that have been previously assumed to be aneuploidy drivers, such as *PTEN* in Chr10q.

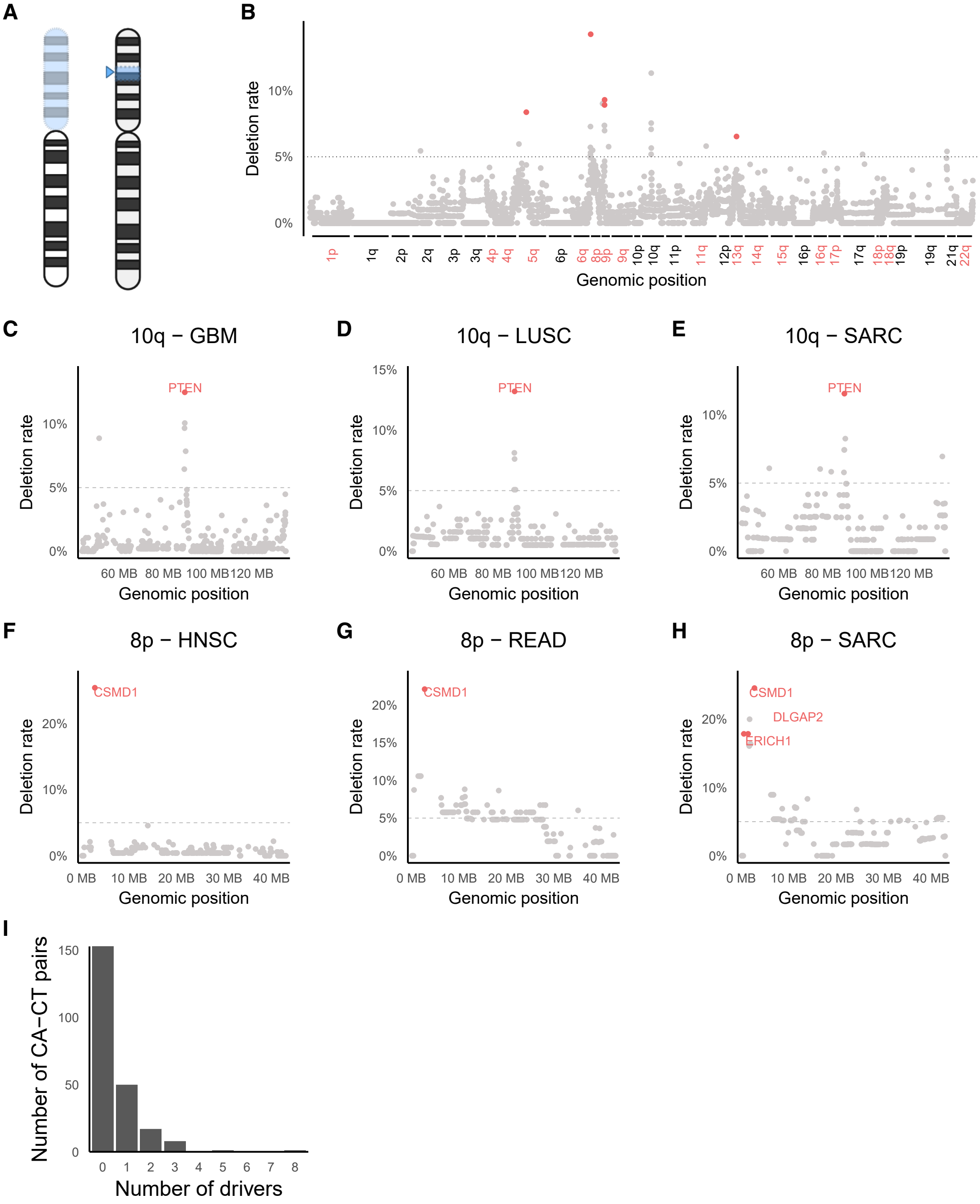


Figure 2. Focal deletions co-occurring with arm losses

(A) Schematic representation of the FD + AL inactivation pattern.

(B) The prevalence of focal deletions co-occurring with arm losses in BRCA. The focal deletion rate of each gene when the other copy of the arm on which it resides is lost in BRCA samples. Arms recurrently lost are colored in red. Genes identified as drivers by this pattern are colored in red.

(C–E) The prevalence of focal deletions co-occurring with Chr10q loss in GBM (C), LUSC (D), and SARC (E).

(F–H) The prevalence of focal deletions cooccurring with Chr8p loss in HNSC (F), READ (G), and SARC (H).

(I) Distribution of the number of drivers identified in CA-CT pairs for the FD + AL pattern.

Point mutations co-occurring with chromosome-arm losses

Next, we investigated genes that are frequently mutated when the arm on which they reside is lost (pattern PM + AL; Figure 3A). For each CA-CT pair, we applied MutSig2CV[27] to the group of samples harboring the arm loss, focusing on genes located on that arm. The results for BRCA are shown in Figure 3B and for all other tumor types, in Data S2A–S2S.

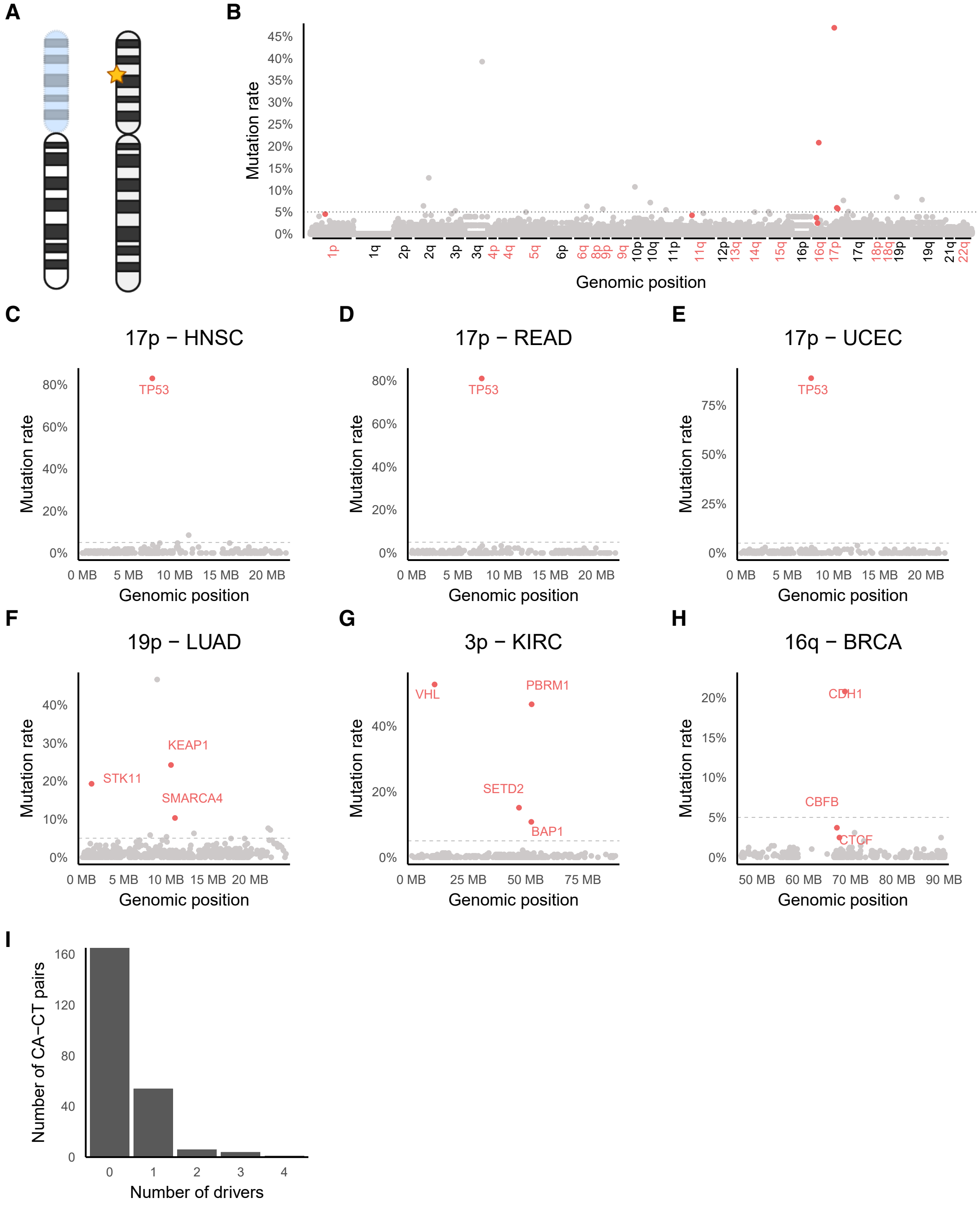


Figure 3. Point mutations co-occurring with arm losses

(A) Schematic representation of the PM + AL inactivation pattern.

(B) The prevalence of gene mutations co-occurring with arm losses in BRCA. Mutation rate of each gene when the other copy of the arm on which it resides is lost in BRCA samples. Recurrently lost arms are colored in red. Genes identified as drivers by this pattern are colored in red.

(C–E) The prevalence of gene mutations co-occurring with Chr17p loss in HNSC (C), READ (D), and UCEC (E).

(F) The prevalence of gene mutations co-occurring with Chr19p loss in LUAD.

(G) The prevalence of gene mutations co-occurring with Chr3p loss in KIRC.

(H) The prevalence of gene mutations co-occurring with Chr16q loss in BRCA.

(I) Distribution of the number of drivers identified in CA-CT pairs for the PM + AL pattern.

The full list of genes identified in this pattern analysis is provided in Table S1A.

A notable example of this pattern is *TP53*, which is frequently mutated in conjunction with a Chr17p loss. This association was observed in virtually all cancer types with frequent loss of Chr17p (Figures 3C–3E and Data S2T–S2AD). *TP53* is a well-established driver of Chr17p loss, and its bi-allelic inactivation through a combination of mutation and arm loss was observed in several cancer types before.[16–18]

Unlike *TP53*, most genes identified in this analysis are frequently mutated in only one or two cancer types. For example, while Chr19p loss is common across three cancer types, candidate driver genes fitting this pattern for this recurrent arm loss were identified only in LUAD (Figure 3F). Similarly, for losses of Chr16q and Chr3p, genes fitting this pattern were found only in three and six of the nine relevant cancer types, respectively. In both cases, the identified driver gene were cancer type specific: for example, *TGFBR2* was nominated as a driver of Chr3p loss only in HNSC, whereas *VHL* and *SETD2* were nominated as drivers of the same aneuploidy only in KIRC (Figure 3G and Data S2AE and S2AF). Likewise, *CDH1* was nominated as a driver of Chr16q loss only in BRCA (Figure 3H and Data S2AG and S2AH). *CDH1* was also identified as a driver in UCEC through the analysis of FD-AL pattern (see below). The identification of *TGFBR2* as a driver of Chr3p loss highlights the advantage of using MutSig2CV[27] over relying solely on mutation rates. *TGFBR2,* a known TSG, does not exhibit the highest mutation rate among genes on Chr3p. However, by incorporating additional gene-level features, MutSig2CV successfully prioritizes it over genes showing similar mutation rates. The relationship between mutation rates and MutSig2CV scores is illustrated in Data S2AK.

Overall, this analysis identified a total of 82 potential driver genes, within 28% of the CA-CT pairs. Of these, in the vast majority (83%) of cases, a single gene was identified, whereas in the other cases, the impacted region contained multiple (2–4) genes (Figure 3I; Table S1A). This analysis identified well-known TSGs, such as *TP53* in several tumor types, *VHL*[28] in kidney cancer, and *SMARCA4*[29] in lung cancer, but also genes that had not been previously implicated in driving recurrent arm losses, such as *NSD1* and *RASA1* in LUSC and HNSC, respectively (Data S2AI and S2AJ).

Focal deletions mutually exclusive with chromosome-arm losses

Some driver genes may be lost either through an arm loss or through a more focal loss but not by both losses simultaneously. This is either because the driver itself is an essential gene that cannot be bi-allelically inactivated or because of essential genes located near the driver. Therefore, we investigated focal deletions that are prevalent in a particular arm only when the other copy of the arm is not lost (pattern FD-AL; Figure 4A). These mutually exclusive events were relatively large, spanning large fractions of the respective arm (a median length of 6.9 Mb and of 159 genes; exemplified for BRCA in Figure 4B and for all other tumor types, in Data S3A–S3S). For each CA-CT pair, we applied GISTIC2.0[22] to samples lacking the arm loss to identify recurrent focal losses.

Since the identified chromosomal regions were large, harboring tens to hundreds of genes, most likely mere passengers, additional filtering steps were required for driver nomination. We therefore used PRODIGY,[30] a tool that we previously developed to identify driver genes based on their proximity to members of dysregulated pathways in a protein-protein interaction network (see STAR Methods). However, we noticed that many of the nominated candidates were in fact known oncogenes, likely due to the large number of genes per region and the lack of directionality requirements in our analysis. Therefore, to increase the confidence in the identified candidates, we further refined our approach to focus solely on 1,166 known TSGs, compiled from the Cancer Gene Census (CGC)[31] and the Tumor Suppressor Gene Database v2.0 (TSGene).[32] While this filtering step limited the scope of this particular analysis, it greatly increases the confidence in the validity of the identified drivers.

The full list of genes identified in this pattern analysis is provided in Table S1A. An example of this pattern is seen on Chr18q, which contains a sizable region frequently lost in the absence of the arm loss across multiple cancer types. Our analysis identified *SMAD4* as a candidate driver of Chr18q loss in eight cancer types, while also nominating other, cancer type-specific drivers (Figures 4C–4E and Data S3T–S3V).

Another interesting example is that of Chr4q, for which different regions were found to be frequently lost, in a tumor type-dependent manner (Figures 4F–4H and Data S3W–S3AA). One region, located at the end of that arm, was observed across several cancer types, and it includes two prominent candidate driver genes, *CASP3* and *FAT1*. *CASP3* is a known TSG via its role in apoptosis,[33] and *FAT1* is an atypical cadherin that was also shown to function as a TSG in some cancer types.[34] Another region, situated in the middle of Chr4q, appeared only in a subset of the cancer types, such as LIHC, suggesting that the drivers that reside within that region are tumor type specific (Figure 4F and Data S3W–S3AA).

Overall, this analysis identified 166 potential driver genes, within 42% of the CA-CT pairs (Figure 4I; Table S1A). Of these, in almost all cases (90%) one or two genes were identified. Some of these genes, like *CASP3* and *MAPK10*, have not been proposed as aneuploidy drivers before.

To evaluate the impact of the chosen TSG list on driver identification, we performed additional analyses using subsets of our comprehensive TSG list, with varying levels of stringency. TSGene provides a large catalog of experimentally supported and literature-derived TSGs,[32] whereas the CGC database provides a curated set of genes with substantial evidence for their role in cancer, and the CGC tier one category includes only genes with the highest level of evidence supporting their tumor suppressor function.[31] When restricting the lists to protein-coding genes located on autosomes, we obtained 989 genes from TSGene, 322 from CGC, and 251 from CGC tier one. Using these subsets for filtering PRODIGY results, we identified 140 cancer-specific candidate drivers with TSGene, 76 with CGC (56 of them overlapping with TSGene, p value < 2e-16), and 66 with CGC tier one (54 of them overlapping with TSGene, p value < 2e-16) (Figure S1). These strong overlaps support the robustness of the original analysis and also allow focusing on *bona fide*

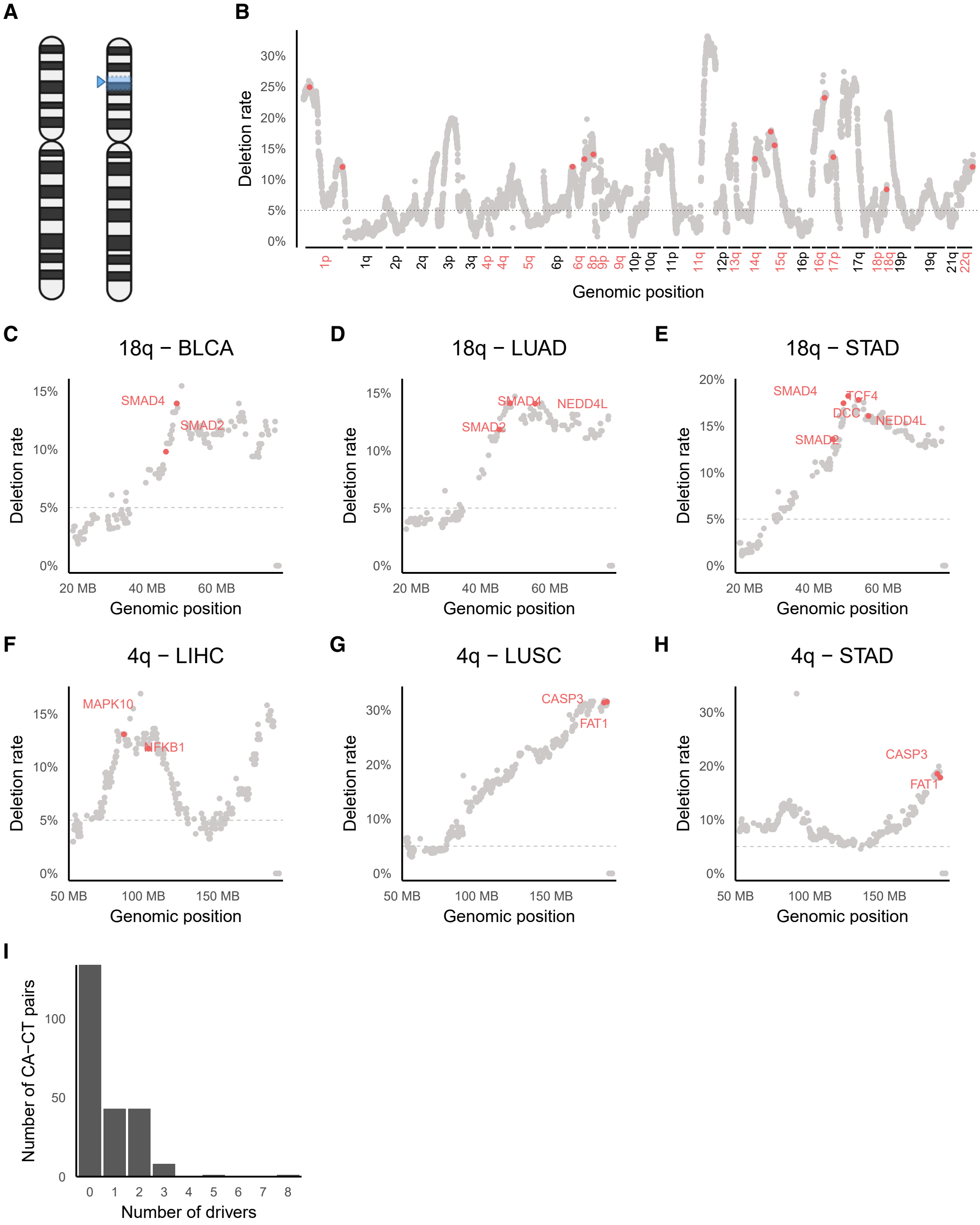


Figure 4. Focal deletions mutually exclusive with arm losses

(A) Schematic representation of the FD-AL inactivation pattern.

(B) The prevalence of gene-level focal deletions when the other copy of the arm is not lost in BRCA. Focal deletion rate of each gene given that the other copy of the arm on which it resides is not lost in BRCA samples. Recurrently lost arms are colored in red. Genes identified as drivers by this pattern are colored in red.

(C–E) The prevalence of gene-level focal deletions when Chr18q is not lost in BLCA (C), LUAD (D), and STAD (E).

(F–H) The prevalence of gene-level focal deletions when Chr4q is not lost in LIHC (F), LUSC (G), and STAD (H).

(I) Distribution of the number of drivers identified in CA-CT pairs for the FD-AL pattern.

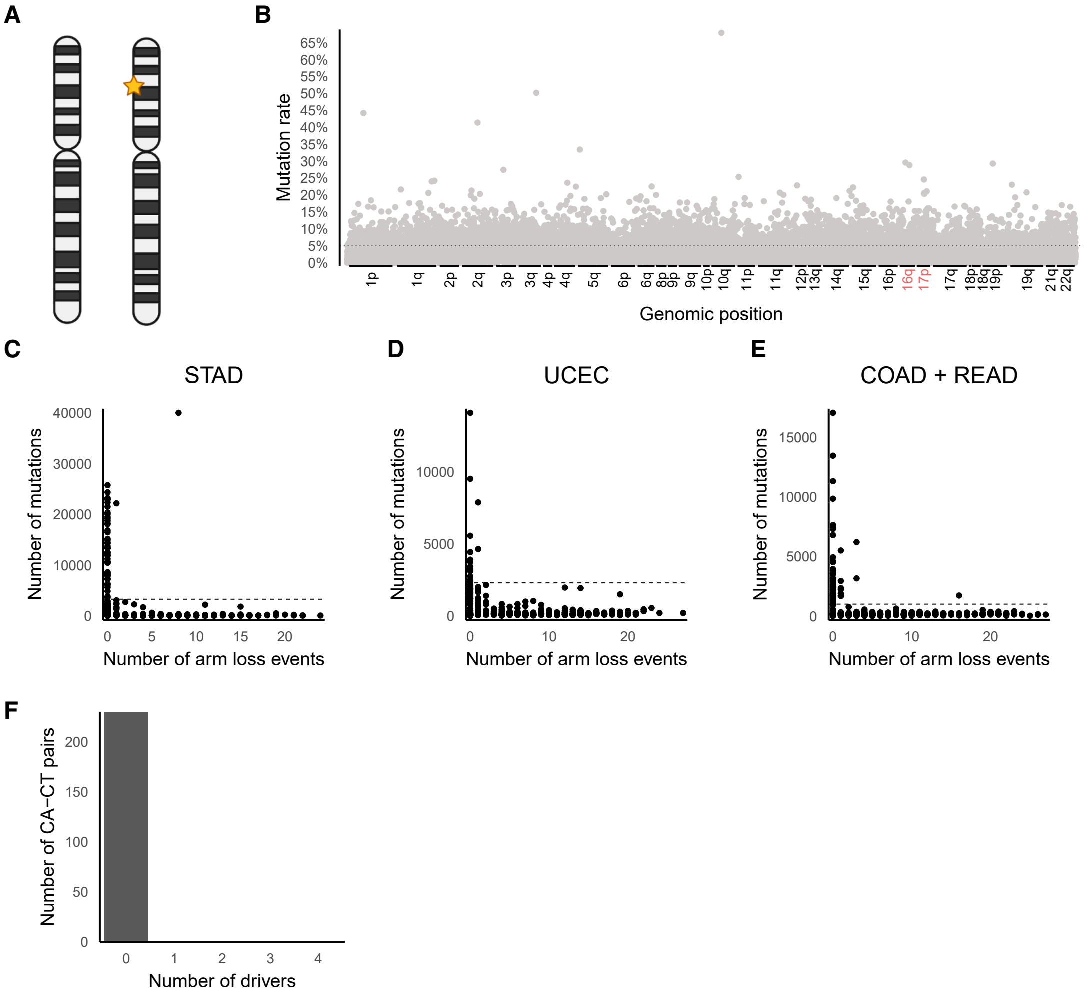


Figure 5. Point mutations mutually exclusive with arm losses

(A) Schematic representation of the PM-AL inactivation pattern.

(B) The prevalence of gene-level mutations in the absence of arm loss in UCEC. Mutation rate of each gene when the other copy of the arm on which it resides is not lost in UCEC samples. Recurrently lost arms are colored in red.

(C–E) Relationship between the number of mutations and number of arm deletions in STAD (C), UCEC (D), COAD, and READ (E) samples. Each dot shows the rates for one sample.

(F) Distribution of the number of drivers identified in CA-CT pairs for the PM-AL pattern.

TSGs for future validation studies. Results for the analyses using the different TSG lists are provided in Table S1B.

Point mutations mutually exclusive with chromosome-arm losses

Applying the same logic, we next searched for genes that are frequently mutated only in the absence of arm-level loss. For each CA-CT pair, we identified genes that were mutated significantly more frequently when the other copy of the arm on which they reside was not lost (pattern PM-AL; Figure 5A). We identified 374 candidate genes, all of which were found in the UCEC, STAD, and COAD-READ cancer types. However, in contrast to the patterns described above, the genes that were identified in this analysis were spread more or less evenly throughout the genome, with most genes presenting a surprisingly high mutation rate, as demonstrated in Figure 5B for UCEC (See Data S4 for the other tumor types). Furthermore, only ∼10% of these genes were known TSGs, compared to 60% and 83% in the FD + AL and PM + AL patterns, respectively (the FD-AL pattern analysis considered only TSGs to begin with).

These results suggest that the genes identified might not be true aneuploidy drivers. Importantly, the four cancer types in which the genes were identified are all known to have hypermutated, chromosomally stable subtypes.[35–37] Indeed, we found a subgroup of samples in these cancer types with a very high mutation count but very few, or even zero, arm losses (Figures 5C–5E), suggesting that the high mutation rate in these genes was mutually exclusive to arm losses simply because of this strong negative association, indicating that these were not true arm loss driver genes (Figure 5F).

Consolidation of the genetic analyses

Overall, we identified 322 CA-CT candidate drivers involving 140 unique genes. Approximately 71% of these are known TSGs. Importantly, however, our analysis nominated only ~8.6% of all known TSGs as putative drivers of arm losses, highlighting the value of harnessing genomic data to distinguish putative drivers from TSGs that just happen to reside on a recurrently lost arm.

Most of the candidate driver genes fit only one of the inactivation patterns discussed above (Figure 6A), highlighting the importance of employing complementary approaches for their nomination. Genes that fit multiple patterns are particularly noteworthy, though, as they provide stronger evidence supporting their roles as aneuploidy drivers. *FAT1*, *PTEN*, *RB1*, and *SMAD4* are the only genes that fit all three patterns. Indeed, all except *FAT1* were previously suggested to drive their respective arm loss.[19,38,39]

Interestingly, most putative drivers are specific to only one or a few cancer types (Figure 6B). However, a subset of genes appears in multiple cancer types, and 15 genes (11% of the proposed drivers) were identified in five or more cancer types, highlighting their potential significance as universal drivers (Figure 6B). For example, *TP53*, the most common TSG in human cancer, was identified as a driver of Chr17p loss in 14 different tumor types.

To further validate our list of candidate driver genes, we assessed their overlap with known essential genes and oncogenes, hypothesizing that arm loss drivers should be depleted of such genes. A list of 585 essential genes was obtained from the Online Gene Essentiality (OGEE) database,[40] and a list of 341 oncogenes was derived from OncoKB.[41] Out of our 140 unique candidate drivers, only four were known essential genes and six were known oncogenes (Figure 6C). Notably, eight out of these ten genes were identified exclusively by the FD-AL analysis, indicating that these genes appeared on our list of known TSGs despite also being classified as essential genes or oncogenes. This observation further supports the potential benefit of employing a more stringent TSG filtering approach, as discussed previously. Additionally, the essential gene CTCF and the oncogene KLF5 were identified by the PM + AL analysis; however, both of these genes have been previously reported to also have tumor-suppressive roles. Therefore, our 140 candidate genes did not contain any *bona fide* essential gene or oncogenes without any known tumor suppressive role.

Validation of putative breast cancer candidate driver genes using the METABRIC dataset

To validate our pipeline using an independent dataset, we applied it to breast cancer data from the METABRIC study.[42] We first computed arm-level aneuploidy scores using the available gene-level copy-number data (see STAR Methods), confirming that the arm loss patterns were highly similar between the TCGA breast cancer (BRCA) cohort and the METABRIC cohort (Figure S2A). Next, we identified candidate driver genes by applying our pipeline adjusted to the METABRIC data (STAR Methods). Overall, the driver genes identified in the METABRIC dataset significantly overlapped those identified in the TCGA BRCA dataset. For the FD + AL pattern, we identified 16 candidate drivers, including 3 of the 5 genes identified in this pattern in TCGA BRCA and 5 out of 44 identified across all TCGA cancer types (hypergeometric test p values = 6.75e-09 and 3.85e-10, respectively; Figure S2B). For the FD-AL pattern, we identified 18 candidate drivers, including 4 of the 12 genes identified in TCGA BRCA and 12 out of 74 identified across all TCGA cancer types (p values = 2.04e-05 and 3.34e-11, respectively; Figure S2C). Lastly, for the PM + AL pattern, we identified 12 candidate drivers, including 5 of the 8 genes identified in TCGA BRCA and 8 out of 42 identified across all TCGA cancer types (p values = 3.67e-15 and 3.25e-19, respectively; Figure S2D). We note that the higher number of candidate drivers identified in METABRIC compared to TCGA BRCA is likely due to the substantially larger sample size (~twice as many samples) of the METABRIC dataset. The full results for the METABRIC analysis are provided in Table S1C.

Comparison of putative driver genes between WGD+ and WGD− tumors

Whole-genome duplication (WGD) is associated with both genomic instability and aneuploidy. Tumors that have undergone WGD (hereafter referred to as WGD+) exhibit higher rates of aneuploidy and distinct recurring aneuploidy patterns.[2,7] Given this association, we aimed to determine whether different drivers might be identified in WGD+ vs. WGD− tumors. To address this, we ran our pipeline separately for each group. Overall, most driver genes identified were common to both groups (Figure S3). We observed strong agreement of GISTIC2.0 and MutSig2CV scores, used for driver identification in the FD + AL and PM + AL patterns, respectively, between these two analyses and between each of them to our original analysis (which included all tumor samples). The MutSig2CV scores of the WGD+ and WGD− analyses show Pearson's correlation of 0.70 with p value = 5.01e-12, and the GISTIC2.0 scores for these analyses have Pearson's correlation of 0.76 with p value = 7.77e-25. Specifically, out of the 46 and 57 cancer-specific drivers identified in the PM + AL pattern for WGD+ and WGD− tumors, respectively, 30 were shared (hypergeometric test p value < 2e-16), with 5 genes identified exclusively in one WGD group (and not in the other one or in the original analysis). Similarly, for the FD + AL pattern, out of 94 and 83 drivers identified for WGD+ and WGD− tumors, 52 were shared (p value < 2e-16), with 18 genes identified exclusively in one WGD group (Figure S3). The FD-AL pattern results showed a weaker, yet still very significant, overlap (85 drivers in WGD+, 158 drivers in WGD−, 37 shared; p value < 2e-16; 75 genes identified exclusively in one WGD group), likely due to the increased complexity of this analysis. Although some candidate drivers might not be identified in the separate WGD+ or WGD− analyses simply due to the smaller group sizes and the resulting reduced statistical power, others may be truly WGD status dependent. One such candidate is SMARCA4, a known TSG and chromatin remodeler previously associated with genomic stability,[29] which was identified in the WGD+ analysis but not in the WGD− analysis. Other genes that are identified exclusively in one WGD group are also of interest. The complete results from the WGD status-stratified analyses are provided in Tables S1D and S1E.

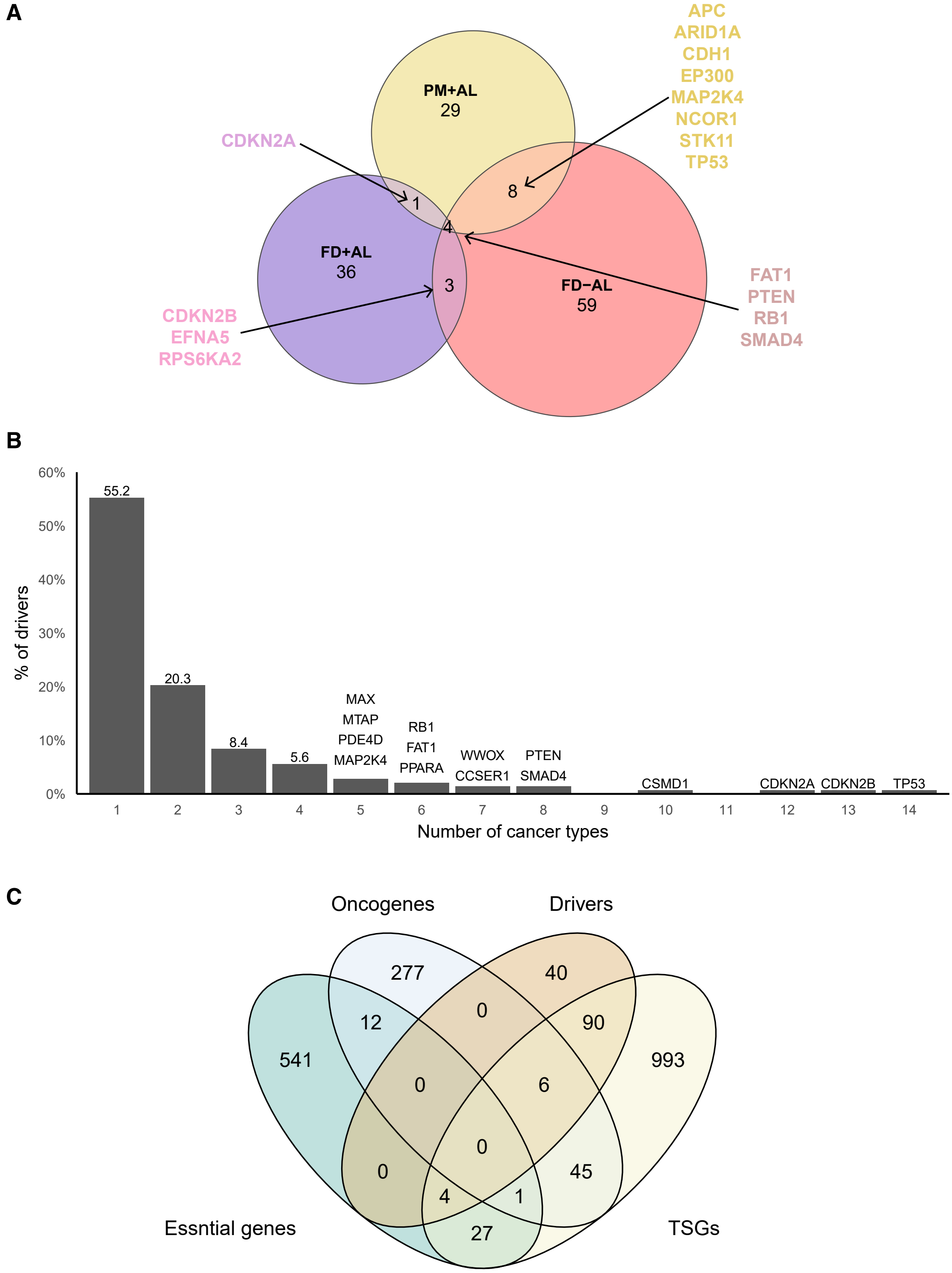
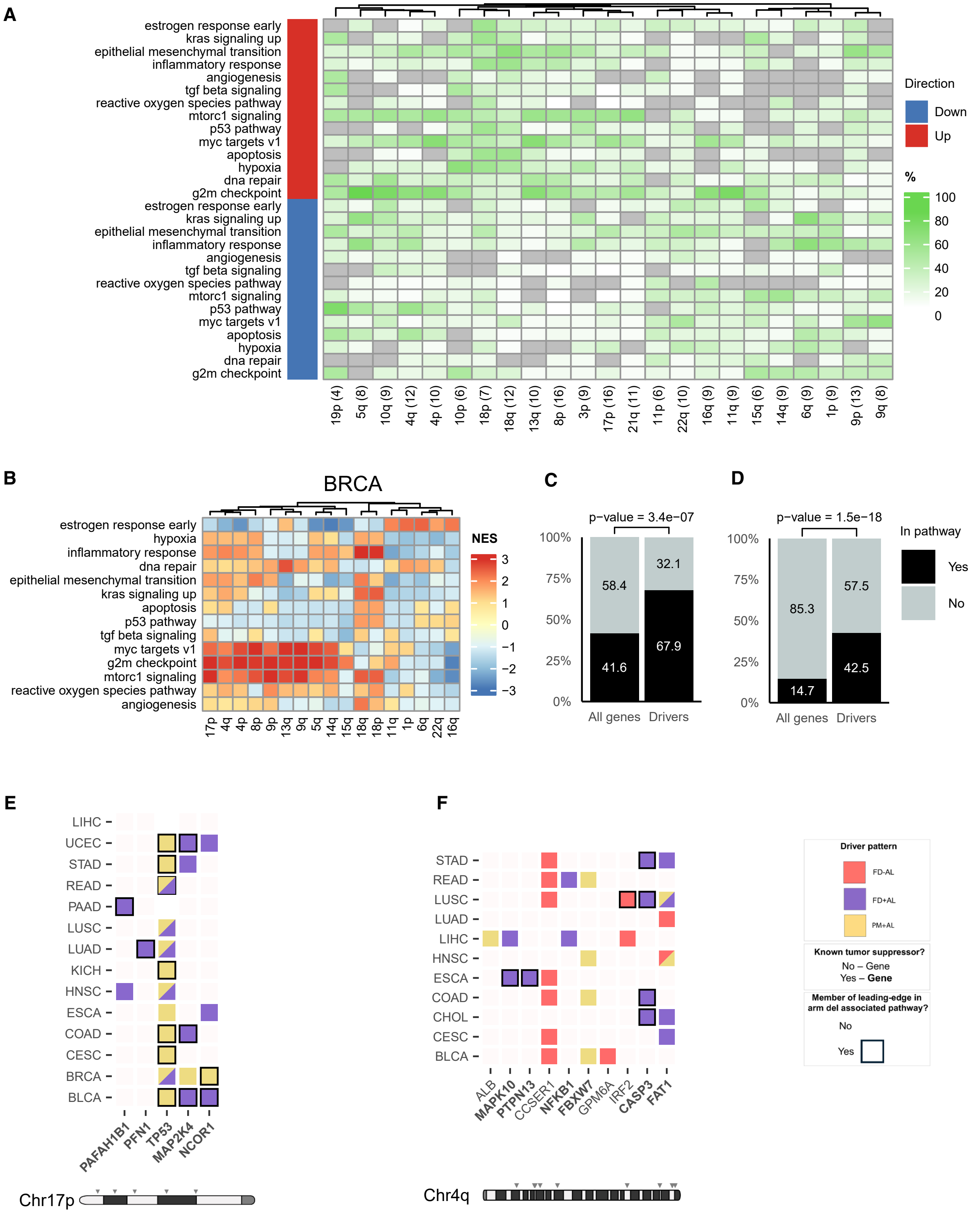


Figure 6. Consolidation of the genetic analyses

(A) Venn diagram of the candidate drivers identified by each inactivation pattern.

(B) Recurrence of candidate drivers across cancer types. The histogram shows the distribution of drivers according to the number of cancer types in which they were identified. Drivers that were identified in five or more cancer types are named above their bar.

(C) Venn diagram illustrating the overlap between the candidate drivers, essential genes, oncogenes, and tumor suppressor genes (TSGs).



(legend on next page)

Gene expression changes associated with recurrent chromosome-arm losses

To identify the gene expression consequences of specific arm losses in each cancer type, we performed the following comparisons for each CA-CT pair (Figure 1C): (1) within-cancer: cancer samples that harbor the arm loss were compared to cancer samples without the loss, and (2) between cancer and normal tissue: cancer samples that harbor the arm loss were compared to the normal-adjacent tissue samples.

For each comparison, we performed a differential gene expression (DGE) analysis followed by gene set enrichment analysis (GSEA)[43] using the MSigDB "Hallmark" gene sets collection[44] (Table S2). This collection includes 50 gene sets that represent well-defined biological states or processes, offering a high-level perspective on the differences between groups in each comparison. Our preliminary within-cancer type comparisons showed consistent patterns across different arms. We hypothesized that these findings reflect the general effects of high aneuploidy levels. To account for this, we repeated the analysis controlling for the confounder effect of aneuploidy levels in these comparisons using IPTW (see STAR Methods). On average, 19 and 17 "Hallmark" gene sets were differentially expressed between the aneuploid tumors and the non-aneuploid tumors or the normal adjacent tissues, respectively (Figures S4A and S4B). It is important to note that the sample sizes of normal tissues were considerably smaller than those of the tumor groups, limiting the statistical power for comparisons involving normal samples. While these results suggest that arm losses may indeed have strong transcriptional consequences, the observed gene expression changes were likely influenced by additional factors associated with the recurrent arm losses, complicating the identification of the true gene expression consequences of these aneuploidies.

Next, we sought to determine whether a specific arm loss is associated with consistent gene expression changes across cancer types. For each recurrent arm loss, we calculated the fraction of cancer types in which each gene set was significantly up- or downregulated (Figures 7A and S4C). When analyzing the results of the within-cancer type comparisons, we found that, in most cases, the gene expression changes associated with an arm loss were not ubiquitous across cancer types (Figure 7A; Table S2). In contrast, when analyzing the cancer vs. normal tissue comparisons, the effects were consistent across arms (Figure S4C; Table S2), indicating that these comparisons were likely dominated by general effects of cancer.

Finally, we evaluated the similarity of the gene expression changes that are associated with each arm losses within each cancer type (shown for BRCA in Figures 7B and S4D, and for all other tumor types, in Data S5 and S6). In the within-cancer type comparisons, multiple recurrent arm losses were associated with the same gene expression changes across arms (Figure 7B and Data S5), probably reflecting the co-occurrence of arm losses,[7] as well as general chromosome-independent effects of aneuploidy (or other aneuploidy-associated genomic alterations).[45–47] The cancer vs. normal tissue comparisons showed that the same pathways were associated with nearly all recurrent arm losses, further indicating that the observed effects in this comparison were largely influenced by general cancer-related gene expression changes (Figure S4D and Data S6).

Integrating the genetic analyses with the gene expression analyses

To further validate our findings and highlight the most interesting candidate drivers, we assessed whether the putative drivers were enriched for genes that belong to pathways that are significantly altered upon that arm loss. To do so, we applied GSEA to the gene expression results from the within-cancer type comparisons discussed above using the 1,293 pathways of size between 10 and 500 from the Reactome pathway database.[48] We identified 33,412 such dysregulated pathways across all 230 CA-CT pairs. We found that our putative drivers were indeed significantly more likely to participate in dysregulated pathways (Figure 7C). Additionally, we observed that the putative drivers were highly enriched in the leading-edge subsets of the dysregulated pathways, namely the subsets of genes with the highest contribution to a pathway's enrichment signal as defined in[43] (Figure 7D; see STAR Methods).

These results increase our confidence that pathways containing drivers in their leading-edge are more likely to be directly associated with the respective arm loss. Similarly, they boost our confidence that our approach indeed identifies biologically meaningful driver genes, which affect the gene expression

Figure 7. Integrating the genetic driver analysis with the gene expression analysis

(A) Pathway dysregulation associated with recurrent arm losses across cancer types. For each arm, the percentage of cancer types in which a given "Hallmark" gene set is significantly dysregulated (out of all cancer types showing a recurrent loss of the arm) was calculated, by comparing tumors with the arm loss to those without it. Selected cancer-related "Hallmark" gene sets are shown.

(B) Pathway dysregulation associated with recurrent arm losses in BRCA. For each BRCA recurrent arm loss, the upregulation or downregulation of selected cancer-related "Hallmark" gene sets IS shown, comparing tumors with vs. without each arm loss. NES: GSEA normalized enrichment score. BRCA, breast cancer.

(C) Enrichment of drivers in aneuploidy-associated pathways. For each CA-CT pair, the dysregulated Reactome pathways were identified by performing GSEA between the expression in tumor samples with the arm loss and those without it. The plot shows the fraction of drivers vs. the fraction of all genes belonging to such pathways. *p* value is calculated using Fisher's exact test.

(D) Enrichment of drivers in the leading-edge subsets of aneuploidy-associated pathways. The plots are as in (C) but showing the fraction of drivers vs. the fraction of all genes belonging to the leading-edge subsets of the dysregulated pathways.

(E) A summary visualization of the candidate drivers of Chr17p loss across cancer types. The candidate drivers of Chr17p loss identified in each cancer type are shown in a matrix, with rows corresponding to the cancer types and columns corresponding to the drivers. The colors reflect the perturbation pattern(s) in which each gene was identified. Drivers present in the leading-edge subsets of at least one dysregulated pathway are highlighted with a black border. The names of known TSGs appear in bold. The chromosome diagram shows the locations of the drivers, in the order in which they appear in the matrix.

(F) A summary visualization of the candidate drivers of Chr4q loss across cancer types. Same plot as in (E) but for Chr4q.

patterns of the tumors. We therefore integrated the genetic analyses with the gene expression analyses. For example, in tumor types in which Chr17p is recurrently lost, the cell cycle checkpoint and programmed cell death pathways are downregulated in Chr17p-loss samples compared to Chr17p-wildtype samples across several cancer types (Table S2). *TP53* appears in the leading edge of these gene sets and is proposed by our analysis to be a strong driver of this recurring aneuploidy (Table S1A). Similarly, Chr4q loss is associated with the downregulation of cell death pathways in STAD, COAD, and LUSC (Table S2), presumably driven by the inactivation of *CASP3* (Table S1A). Another noteworthy observation, seen in BLCA, GBM, and HNSC, is that apoptosis, cell cycle checkpoint, and programmed cell death pathways are upregulated in tumors with Chr9p loss compared to healthy tissues but are downregulated in comparison to Chr9p-wildtype samples (Table S2). This pattern suggests that these pathways may be generally upregulated in certain cancer types, with Chr9p loss mitigating this detrimental activation through the downregulation of *CDKN2A* and *CDKN2B*, which we identify as putative drivers of this aneuploidy (Table S1A).

Lastly, we provide a visual summary of the putative drivers of each CA-CT pair (Figures 7E and 7F and Data S7). For instance, Figure 7E provides an integrated view of the putative drivers of Chr17p loss. *TP53* appears across all cancer types, primarily through the PM + AL pattern. Four other genes are known TSGs that were proposed as drivers of this arm loss in different cancer types. These genes also frequently appear in the leading-edge subsets of dysregulated pathways identified in the GSEA analysis (Figure 7E). In contrast, the putative drivers of Chr4q loss reveal a more diverse, tumor-type-specific group of drivers, including both known TSGs and less well-characterized genes (Figure 7F). Most of these proposed drivers are specific to a handful of cancer types. Notable candidates include *CCSER1*, commonly focally deleted in the absence of Chr4q; *CASP3*, a known TSG residing in a small region frequently bi-allelically deleted across many cancers; and *FAT1*, another well-established TSG identified across many cancer types through all three patterns. Summary plots for all other recurrently lost arms are shown in Data S7. The full lists of putative drivers, together with the patterns and the pathways to which they belong, are provided in Table S1A.

Improved Charm scores based on our candidate drivers

Previous studies by Davoli et al.[15] and Jubran & Slutsky et al.[6] identified features of chromosome arms that influence their likelihood of being gained or lost in cancer. Specifically, Davoli et al.[15] demonstrated that the density and potency of TSGs are positively correlated with arm-loss prevalence, proposing *Charm*^{TSG} as a measure of the strength of the selection pressure to lose a copy of each chromosome arm.[15] To investigate how our candidate driver genes perform compared to these *Charm*^{TSG} scores, we developed a new scoring metric based exclusively on our candidate driver genes, *Charm*^{drivers} (rather than on all arm-residing TSGs in Davoli et al.; see STAR Methods). The new driver-based scores yielded a stronger correlation with the prevalence of arm loss, compared to the previous TSG-based scores (Pearson's r = 0.73, *p* value = 3.35e-05 for our model vs. r = 0.53, *p* value = 8.70e-03 for the Davoli et al. model; Figure S5). This increases the confidence in, and demonstrates the utility of, our candidate driver gene results.

DISCUSSION

In this study, we presented a framework for identifying driver genes associated with cancer type-specific recurrent arm losses. Our analysis revealed that focal deletions co-occurring with losses of the same arm are typically small (median size of ~1.1 Mb). The co-occurrence of these events suggests the presence of a genetic element that is bi-allelically inactivated by these deletions, thereby driving their recurrence. This observation enabled us to narrow down our search to specific regions of the arm and, in many cases, to propose candidate genes that likely contribute to the frequent loss of the arm.

Conversely, our analysis also identified medium-size deletions (median size of ~8.2 Mb) that are prevalent only in the absence of an arm loss. This pattern suggests the presence of a genetic element that is frequently inactivated by either event, driving the recurrence of these losses. This pattern may suggest that only a mono-allelic inactivation of this driver is advantageous for the cancer cell. Alternatively, essential genes located near the candidate driver(s) may prevent bi-allelic loss of the genomic region containing that driver(s) from being viable, as their complete inactivation would be lethal. Our point-mutation analysis supports this explanation, as we found no evidence of a driver gene inherently resistant to bi-allelic inactivation. This suggests that negative selection against bi-allelic deletions is more likely due to the presence of nearby essential genes. Analysis of mutation data identified a class of genes frequently bi-allelically inactivated by an arm loss and a point mutation. *TP53* is the best-known example of this mechanism, and we indeed identified it as a strong driver of Chr17p loss across many tumor types. On the other hand, we see no evidence for the existence of genes whose inactivation by an arm loss and by a gene mutation is mutually exclusive, suggesting that essential genes whose complete loss is not compatible with cell fitness are rarely (if ever) drivers of arm losses. These results suggest that negative selection does not act against the bi-allelic inactivation of driver genes. Instead, the negative selection observed in the FD-AL pattern appears to target the bi-allelic inactivation of nearby passenger genes rather than the driver gene itself.

By comparing the gene expression of cancer samples harboring a specific arm loss to those without it, followed by a GSEA, we identified pathways associated with the arm loss. However, our analysis demonstrates that these gene expression changes are not sufficient in and of themselves for driver identification, as they may be confounded by co-occurring aneuploidies and other genomic alterations. Therefore, an integrative approach that takes into account both gene expression and genetics, like the one we present here, is required. Integrating genetic and expression analyses identified specific pathways whose dysregulation is most likely attributed to the arm loss. This approach also allowed us to link specific driver genes to these pathway alterations, pinpointing specific cellular pathway(s) through which the loss of these genes promotes cancer.

We note that most of our proposed candidates are cancer type-specific, emphasizing the context-specific role of arm loss (reviewed in[1]). A small subset of genes appears across cancer types, though, alluding to their strong and global role as cancer aneuploidy drivers. Additionally, most drivers are recognized in only one of our analyzed genomic patterns, indicating that specific genes tend to prefer a specific mechanism of inactivation, and emphasizing the importance of such integrative analyses.

It is important to note that our work does not contradict previous studies that demonstrated that aneuploidy prevalence could largely be predicted by the density and potency of the cancer genes that reside on each arm.[6] In fact, we were able to improve the correlation between the Charm scores and arm loss prevalence with our new list of candidate drivers (Figure S5). However, the cumulative effect of TSGs on the prevalence of arm loss does not contradict a major contribution of specific driver genes.[6,10] A prominent example for that is TP53, widely recognized as the key driver of 17p arm loss (del17p). While multiple genes on this arm may contribute cumulatively, del17p is clearly a common mechanism for achieving biallelic inactivation of *TP53*.[16,18] Indeed, our results clearly validate the importance of *TP53* in the recurrence of this arm loss, while suggesting that additional genes are also involved in the high prevalence of del17p (Figure 7E).

In sum, our work successfully identified driver genes underlying recurring arm losses. We believe that this study makes an important step toward a systematic, comprehensive characterization of the drivers of aneuploidy in human cancer.

Limitations of the study

There are several limitations to our approach: (1) We only focused here on chromosome-arm losses; future research should expand this approach to include chromosome-arm gains as well. (2) We only considered the most common focal gene inactivation mechanisms, namely mutations and copy number alterations. Incorporating a broader range of inactivation mechanisms (e.g., promoter methylation) into this framework may help capture more diverse driver patterns. (3) Our analysis considered only one event at a time, ignoring the potential driving role of co-occurring aneuploidies.[7,9] Much more data are needed to perform such combinatorial analyses, but with the fast accumulation of genomic information, this will likely become possible within a few years. (4) We only considered protein-coding genes. It will be important to ultimately extend the analysis to consider other genetic elements, such as microRNAs and long non-coding RNAs. (5) Lastly, the CNA data used in this study are based on SNP arrays, which do not provide uniform genome-wide coverage. This may reduce the resolution of focal CNAs, especially in regions sparsely covered by SNP probes, potentially leading to the under-detection of small or poorly mapped deletions.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Uri Ben-David (ubendavid@tauex.tau.ac.il).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This study analyzes publicly available data from TCGA and METABRIC. DOIs and links are provided in the key resources table.
- All original code has been deposited at Zenodo at https://doi.org/10.5281/zenodo.17041168 and is publicly available as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

The authors would like to thank all members of the Ben-David lab and the Shamir lab for their invaluable help and support throughout this study. This work was supported by the European Research Council Starting Grant (grant #945674 to U.B.-D.), the Israel Science Foundation (grants #1805/21 to U.B.-D. and #2206/22 to R. Shamir), the Israel Cancer Research Fund (project grant to U.B.-D.), and the U.S.-Israel Binational Science Foundation (grant #2023236 to U.B.-D.). R. Saad was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. Some figures were created with BioRender.com.

AUTHOR CONTRIBUTIONS

All authors conceived the study, designed the analyses, interpreted the results, and wrote the manuscript. R. Saad performed the analyses. R. Shamir and U.B.-D. oversaw the study.

DECLARATION OF INTERESTS

U.B.-D. declares receiving consulting fees from Accent Therapeutics. R. Saad is an employee of CytoReason LTD.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2025.116455.

Received: March 12, 2025
Revised: July 31, 2025
Accepted: September 30, 2025
Published: October 22, 2025

REFERENCES

1. Ben-David, U., and Amon, A. (2020). Context is everything: aneuploidy in cancer. Nat. Rev. Genet. *21*, 44–62. https://doi.org/10.1038/s41576-019-0171-x.

2. Taylor, A.M., Shih, J., Ha, G., Gao, G.F., Zhang, X., Berger, A.C., Schumacher, S.E., Wang, C., Hu, H., Liu, J., et al. (2018). Genomic and functional approaches to understanding cancer aneuploidy. Cancer Cell *33*, 676–689.e3. https://doi.org/10.1016/j.ccell.2018.03.007.

3. Sdeor, E., Okada, H., Saad, R., Ben-Yishay, T., and Ben-David, U. (2024). Aneuploidy as a driver of human cancer. Nat. Genet. *56*, 2014–2026. https://doi.org/10.1038/s41588-024-01916-2.

4. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. *45*, 1113–1120. https://doi.org/10.1038/ng.2764.

5. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607. https://doi.org/10.1038/nature11003.

6. Jubran, J., Slutsky, R., Rozenblum, N., Rokach, L., Ben-David, U., and Yeger-Lotem, E. (2024). Machine-learning analysis reveals an important role for negative selection in shaping cancer aneuploidy landscapes. Genome Biol. *25*, 95. https://doi.org/10.1186/s13059-024-03225-7.

7. Prasad, K., Bloomfield, M., Levi, H., Keuper, K., Bernhard, S.V., Baudoin, N. C., Leor, G., Eliezer, Y., Giam, M., Wong, C.K., et al. (2022). Whole-genome duplication shapes the aneuploidy landscape of human cancers. Cancer Res. *82*, 1736–1752. https://doi.org/10.1158/0008-5472.CAN-21-2065.

8. Patkar, S., Heselmeyer-Haddad, K., Auslander, N., Hirsch, D., Camps, J., Bronder, D., Brown, M., Chen, W.-D., Lokanga, R., Wangsa, D., et al. (2021). Hard wiring of normal tissue-specific chromosome-wide gene expression levels is an additional factor driving cancer type-specific aneuploidies. Genome Med. *13*, 93. https://doi.org/10.1186/s13073-021-00905-y.

9. Nair, N.U., Schäffer, A.A., Gertz, E.M., Cheng, K., Zerbib, J., Sahu, A.D., Leor, G., Shulman, E.D., Aldape, K.D., Ben-David, U., and Ruppin, E. (2024). Chromosome 7 gain compensates for chromosome 10 loss in glioma. Cancer Res. *84*, 3464–3477. https://doi.org/10.1158/0008-5472.CAN-24-1366.

10. Watson, E.V., Lee, J.J.-K., Gulhan, D.C., Melloni, G.E.M., Venev, S.V., Magesh, R.Y., Frederick, A., Chiba, K., Wooten, E.C., Naxerova, K., et al. (2024). Chromosome evolution screens recapitulate tissue-specific tumor aneuploidy patterns. Nat. Genet. *56*, 900–912. https://doi.org/10.1038/s41588-024-01665-2.

11. Kuzmin, E., Baker, T.M., Lesluyes, T., Monlong, J., Abe, K.T., Coelho, P.P., Schwartz, M., Del Corpo, J., Zou, D., Morin, G., et al. (2024). Evolution of chromosome-arm aberrations in breast cancer through genetic network rewiring. Cell Rep. *43*, 113988. https://doi.org/10.1016/j.celrep.2024.113988.

12. Shukla, A., Nguyen, T.H.M., Moka, S.B., Ellis, J.J., Grady, J.P., Oey, H., Cristino, A.S., Khanna, K.K., Kroese, D.P., Krause, L., et al. (2020). Chromosome arm aneuploidies shape tumour evolution and drug response. Nat. Commun. *11*, 449. https://doi.org/10.1038/s41467-020-14286-0.

13. Shih, J., Sarmashghi, S., Zhakula-Kostadinova, N., Zhang, S., Georgis, Y., Hoyt, S.H., Cuoco, M.S., Gao, G.F., Spurr, L.F., Berger, A.C., et al. (2023). Cancer aneuploidies are shaped primarily by effects on tumour fitness. Nature *619*, 793–800. https://doi.org/10.1038/s41586-023-06266-3.

14. Girish, V., Lakhani, A.A., Thompson, S.L., Scaduto, C.M., Brown, L.M., Hagenson, R.A., Sausville, E.L., Mendelson, B.E., Kandikuppa, P.K., Lukow, D.A., et al. (2023). Oncogene-like addiction to aneuploidy in human cancers. Science *381*, eadg4521. https://doi.org/10.1126/science.adg4521.

15. Davoli, T., Xu, A.W., Mengwasser, K.E., Sack, L.M., Yoon, J.C., Park, P.J., and Elledge, S.J. (2013). Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell *155*, 948–962. https://doi.org/10.1016/j.cell.2013.10.011.

16. Liu, Y., Chen, C., Xu, Z., Scuoppo, C., Rillahan, C.D., Gao, J., Spitzer, B., Bosbach, B., Kastenhuber, E.R., Baslan, T., et al. (2016). Deletions linked to TP53 loss drive cancer through p53-independent mechanisms. Nature *531*, 471–475. https://doi.org/10.1038/nature17157.

17. Chin, M., Sive, J.I., Allen, C., Roddie, C., Chavda, S.J., Smith, D., Blombery, P., Jones, K., Ryland, G.L., Popat, R., et al. (2017). Prevalence and timing of TP53 mutations in del(17p) myeloma and effect on survival. Blood Cancer J. *7*, e610. https://doi.org/10.1038/bcj.2017.76.

18. Laue, K., Pozzi, S., Cohen-Sharir, Y., Winkler, T., Eliezer, Y., Israeli Dangoor, S., Leikin-Frenkel, A.I., Lange, K., Zerbib, J., Ricci, A.A., et al. (2023). Inactivation of p53 drives breast cancer brain metastasis by altering fatty acid metabolismPreprint. bioRxiv. https://doi.org/10.1101/2023.12.20.572490.

19. Sasaki, H., Zlatescu, M.C., Betensky, R.A., Ino, Y., Cairncross, J.G., and Louis, D.N. (2001). PTEN is a target of chromosome 10q loss in anaplastic oligodendrogliomas and PTEN alterations are associated with poor prognosis. Am. J. Pathol. *159*, 359–367. https://doi.org/10.1016/S0002-9440(10)61702-6.

20. Scarisbrick, J.J., Woolford, A.J., Russell-Jones, R., and Whittaker, S.J. (2000). Loss of heterozygosity on 10q and microsatellite instability in advanced stages of primary cutaneous T-cell lymphoma and possible association with homozygous deletion of PTEN. Blood *95*, 2937–2942. https://doi.org/10.1182/blood.V95.9.2937.009k15_2937_2942.

21. Boström, J., Cobbers, J.M., Wolter, M., Tabatabai, G., Weber, R.G., Lichter, P., Collins, V.P., and Reifenberger, G. (1998). Mutation of the PTEN (MMAC1) tumor suppressor gene in a subset of glioblastomas but not in meningiomas with loss of chromosome arm 10q. Cancer Res. *58*, 29–33.

22. Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. *12*, R41. https://doi.org/10.1186/gb-2011-12-4-r41.

23. Pertesi, M., Ekdahl, L., Palm, A., Johnsson, E., Järvstråt, L., Wihlborg, A.-K., and Nilsson, B. (2019). Essential genes shape cancer genomes through linear limitation of homozygous deletions. Commun. Biol. *2*, 262. https://doi.org/10.1038/s42003-019-0517-0.

24. Tuysuz, E.C., Mourati, E., Rosberg, R., Moskal, A., Gialeli, C., Johansson, E., Governa, V., Belting, M., Pietras, A., and Blom, A.M. (2024). Tumor suppressor role of the complement inhibitor CSMD1 and its role in TNF-induced neuroinflammation in gliomas. J. Exp. Clin. Cancer Res. *43*, 98. https://doi.org/10.1186/s13046-024-03019-6.

25. Ma, C., Quesnelle, K.M., Sparano, A., Rao, S., Park, M.S., Cohen, M.A., Wang, Y., Samanta, M., Kumar, M.S., Aziz, M.U., et al. (2009). Characterization *CSMD1* in a large set of primary lung, head and neck, breast and skin cancer tissues. Cancer Biol. Ther. *8*, 907–916. https://doi.org/10.4161/cbt.8.10.8132.

26. Escudero-Esparza, A., Bartoschek, M., Gialeli, C., Okroj, M., Owen, S., Jirström, K., Orimo, A., Jiang, W.G., Pietras, K., and Blom, A.M. (2016). Complement inhibitor CSMD1 acts as tumor suppressor in human breast cancer. Oncotarget *7*, 76920–76933. https://doi.org/10.18632/oncotarget.12729.

27. Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L. A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495–501. https://doi.org/10.1038/nature12912.

28. Kaelin, W.G. (2004). The Von Hippel-Lindau tumor suppressor gene and kidney cancer. Clin. Cancer Res. *10*, 6290S–6295S. https://doi.org/10.1158/1078-0432.CCR-sup-040025.

29. Liang, X., Gao, X., Wang, F., Li, S., Zhou, Y., Guo, P., Meng, Y., and Lu, T. (2023). Clinical characteristics and prognostic analysis of SMARCA4 -deficient non-small cell lung cancer. Cancer Med. *12*, 14171–14182. https://doi.org/10.1002/cam4.6083.

30. Dinstag, G., and Shamir, R. (2020). PRODIGY: personalized prioritization of driver genes. Bioinformatics *36*, 1831–1839. https://doi.org/10.1093/bioinformatics/btz815.

31. Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat. Rev. Cancer *18*, 696–705. https://doi.org/10.1038/s41568-018-0060-1.

32. Zhao, M., Kim, P., Mitra, R., Zhao, J., and Zhao, Z. (2016). TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. Nucleic Acids Res. *44*, D1023–D1031. https://doi.org/10.1093/nar/gkv1268.

33. Zhou, Z., Xu, S., Jiang, L., Tan, Z., and Wang, J. (2022). A systematic pan-cancer analysis of CASP3 as a potential target for immunotherapy. Front. Mol. Biosci. *9*, 776808. https://doi.org/10.3389/fmolb.2022.776808.

34. Chen, Z.G., Saba, N.F., and Teng, Y. (2022). The diverse functions of FAT1 in cancer progression: good, bad, or ugly? J. Exp. Clin. Cancer Res. *41*, 248. https://doi.org/10.1186/s13046-022-02461-8.

35. ROSENBAUM, P.R., and RUBIN, D.B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika *70*, 41–55. https://doi.org/10.1093/biomet/70.1.41.

36. Kandoth, C., Schultz, N., Cherniack, A.D., Akbani, R., Liu, Y., Shen, H., Robertson, A.G., Pashtan, I., Shen, R., Benz, C.C., et al. (2013). Integrated genomic characterization of endometrial carcinoma. Nature *497*, 67–73. https://doi.org/10.1038/nature12113.

37. Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330–337. https://doi.org/10.1038/nature11252.

38. Hernandez, A.L., Young, C.D., Wang, J.H., and Wang, X.J. (2019). Lessons learned from *SMAD4* loss in squamous cell carcinomas. Mol. Carcinog. *58*, 1648–1655. https://doi.org/10.1002/mc.23049.

39. Knudsen, E.S., Nambiar, R., Rosario, S.R., Smiraglia, D.J., Goodrich, D. W., and Witkiewicz, A.K. (2020). Pan-cancer molecular analysis of the RB tumor suppressor pathway. Commun. Biol. *3*, 158. https://doi.org/10.1038/s42003-020-0873-9.

40. Gurumayum, S., Jiang, P., Hao, X., Campos, T.L., Young, N.D., Korhonen, P.K., Gasser, R.B., Bork, P., Zhao, X.-M., He, L.J., and Chen, W.H. (2021). OGEE v3: Online GEne Essentiality database with increased coverage of organisms and human cell lines. Nucleic Acids Res. *49*, D998–D1003. https://doi.org/10.1093/nar/gkaa884.

41. Chakravarty, D., Gao, J., Phillips, S.M., Kundra, R., Zhang, H., Wang, J., Rudolph, J.E., Yaeger, R., Soumerai, T., Nissan, M.H., et al. (2017). OncoKB: A precision oncology knowledge base. JCO Precis. Oncol. *2017*, 1–16. https://doi.org/10.1200/PO.17.00011.

42. Pereira, B., Chin, S.-F., Rueda, O.M., Vollan, H.-K.M., Provenzano, E., Bardwell, H.A., Pugh, M., Jones, L., Russell, R., Sammut, S.-J., et al. (2016). The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. Nat. Commun. *7*, 11479. https://doi.org/10.1038/ncomms11479.

43. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550. https://doi.org/10.1073/pnas.0506580102.

44. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. Cell Syst. *1*, 417–425. https://doi.org/10.1016/j.cels.2015.12.004.

45. Cohen-Sharir, Y., McFarland, J.M., Abdusamad, M., Marquis, C., Bernhard, S.V., Kazachkova, M., Tang, H., Ippolito, M.R., Laue, K., Zerbib, J., et al. (2021). Aneuploidy renders cancer cells vulnerable to mitotic checkpoint inhibition. Nature *590*, 486–491. https://doi.org/10.1038/s41586-020-03114-6.

46. Zerbib, J., Ippolito, M.R., Eliezer, Y., De Feudis, G., Reuveni, E., Savir Kadmon, A., Martin, S., Viganò, S., Leor, G., Berstler, J., et al. (2024). Human aneuploid cells depend on the RAF/MEK/ERK pathway for overcoming increased DNA damage. Nat. Commun. *15*, 7772. https://doi.org/10.1038/s41467-024-52176-x.

47. Ippolito, M.R., Zerbib, J., Eliezer, Y., Reuveni, E., Viganò, S., De Feudis, G., Shulman, E.D., Savir Kadmon, A., Slutsky, R., Chang, T., et al. (2024). Increased RNA and Protein Degradation Is Required for Counteracting Transcriptional Burden and Proteotoxic Stress in Human Aneuploid Cells. Cancer Discov. *14*, 2532–2553, OF1–OF22. https://doi.org/10.1158/2159-8290.CD-23-0309.

48. Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G.R., Wu, G.R., Matthews, L., et al. (2005). Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. *33*, D428–D432. https://doi.org/10.1093/nar/gki072.

49. Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T.S., Malta, T.M., Pagnotta, S.M., Castiglioni, I., et al. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. *44*, e71. https://doi.org/10.1093/nar/gkv1507.

50. Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discov. *2*, 401–404. https://doi.org/10.1158/2159-8290.CD-12-0095.

51. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

52. Chesnaye, N.C., Stel, V.S., Tripepi, G., Dekker, F.W., Fu, E.L., Zoccali, C., and Jager, K.J. (2022). An introduction to inverse probability of treatment weighting in observational research. Clin. Kidney J. *15*, 14–20. https://doi.org/10.1093/ckj/sfab158.

53. Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287. https://doi.org/10.1089/omi.2011.0118.

54. Broad Institute TCGA Genome Data Analysis Center (2016). SNP6 Copy Number Analysis (GISTIC2) (Broad Institute of MIT and Harvard). https://doi.org/10.7908/C1S181WN.

55. Akhmedov, M., Kedaigle, A., Chong, R.E., Montemanni, R., Bertoni, F., Fraenkel, E., and Kwee, I. (2017). PCSF: An R-package for network-based interpretation of high-throughput data. PLoS Comput. Biol. *13*, e1005694. https://doi.org/10.1371/journal.pcbi.1005694.

56. Snel, B., Lehmann, G., Bork, P., and Huynen, M.A. (2000). STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res. *28*, 3442–3444. https://doi.org/10.1093/nar/28.18.3442.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
TCGA RNASeq and mutation data	TCGAbiolinks 49	https://doi.org/10.18129/B9.bioc.TCGAbiolinks
TCGA.SEG files	Broad Institute's FireBrowse website	http://firebrowse.org/
TCGA aneuploidy data	Taylor et al. 2	https://doi.org/10.1016/j.ccell.2018.03.007
CGC tumor suppressors	COSMIC website	https://cancer.sanger.ac.uk/cosmic/census
TSGene tumor suppressors	TSGene website	https://bioinfo.uth.edu/
OGEE essential genes	OGEE website	https://v3.ogee.info/
OncoKB oncogenes	OncoKB website	https://www.oncokb.org/
METABRIC data	cBioPortal	https://www.cbioportal.org/
Software and algorithms		
R version 4.4.2	The R Project for Statistical Computing	https://www.r-project.org/
TCGAbiolinks R package	Bioconductor	https://doi.org/10.18129/B9.bioc.TCGAbiolinks
clusterProfiler R package	Bioconductor	https://doi.org/10.18129/B9.bioc.clusterProfiler
DESeq2 R package	Bioconductor	https://doi.org/10.18129/B9.bioc.DESeq2
GISTIC2 docker image	Docker Hub	https://hub.docker.com/r/shixiangwang/gistic
MutSigCV docker image	Docker Hub	https://hub.docker.com/r/genepattern/mutsigcv
BioRender	BioRender.com	https://biorender.com/
Custom code used for all analysis	This paper	https://doi.org/10.5281/zenodo.17041168

METHOD DETAILS

Data collection and processing

Data from 7,503 TCGA samples spanning 20 cancer types were obtained as follows: Gene expression and mutation data from data release version 35 were downloaded using the R package TCGAbiolinks,[49] while segmentation (.SEG) files from TCGA data version 2016_01_28 were obtained from the Broad Institute's FireBrowse website. Arm-level status for each sample and aneuploidy scores (defined as the total number of arm copy-number alterations) were extracted from Table S2 of Taylor et al.[2] A list of 1,166 known tumor suppressor genes was compiled from the Cancer Gene Census[31] and the Tumor suppressor gene database v2.0.[32] A list of 585 cross-cancer essential genes (core-essential genes) was obtained from the Online Gene Essentiality (OGEE) database,[40] and a list of 341 oncogenes was obtained from OncoKB.[41] An arm-loss event was defined as recurring in a specific cancer type if it occurred in 20% or more of the samples. Microarray gene expression data and gene-level copy number data for METABRIC samples were downloaded from cBioPortal.[50]

Differential gene expression analysis

Two types of differential gene expression analyses were performed to evaluate the impact of all recurrent arm losses on gene expression within each cancer type. The first analysis compared cancer samples with the arm loss to normal samples, using DESeq2.[51] The second analysis compared cancer samples with the arm loss to those without it, by estimating the log fold-change of each gene, while accounting for differences in aneuploidy levels between the two groups. Since the sample's aneuploidy level and the arm deletion status are correlated, inverse probability of treatment weighting (IPTW)[52] was used to account for the effect of the sample's aneuploidy levels.

Each type of differential gene expression analysis was followed by GSEA,[43] using the R clusterProfiler package,[53] applied to the list of all differentially expressed genes using the "GSEA" function with default parameters. Two collections of gene sets were queried in each analysis: the MSigDB 'Hallmark' gene sets (50 sets)[44] and the Reactome database,[48] containing 1615 gene sets (only the1293 gene sets with $10 \leq size \leq 500$ were used).

FD + AL analysis

For each cancer type in which a particular arm was recurrently lost, we focused on the subset of tumors harboring the arm loss. Chromosomal regions in that arm affected by focal deletions co-occurring with the arm loss were identified by applying GISTIC2.0[22] with the parameters used in.[54] Out of the genes within each identified region we excluded those deleted in <5 samples or in <5% of the samples with the respective arm loss. In regions containing $\leq$3 genes after filtering, genes with the highest deletion rate were nominated as putative drivers. In regions with >3 genes, a gene was nominated as a putative driver only if it was also identified as a driver independently in another cancer type.

PM + AL analysis

For each cancer type in which a particular arm was recurrently lost, we focused on the subset of tumors harboring the arm loss. Genes affected by point mutations co-occurring with the arm loss were identified by applying MutSig2CV[27] using a q-value threshold of 0.25. Out of the results only genes located on the corresponding arm were taken.

FD-AL analysis

For each cancer type in which a particular arm was recurrently lost, we focused on the subset of tumors lacking the arm loss. We identified chromosomal regions that are frequently lost only when the other copy of the arm they belong to is not lost, by applying GISTIC2.0[22] with the parameters used in.[54] To focus only on focal loss events that are mutually exclusive to the arm loss, regions that have $\geq$50% overlap with a region identified in the focal deletions co-occurrence analysis above were excluded. Frequent losses spanning both arms or $\geq$90% of the arm were excluded as well.

To identify the most plausible drivers in each region, we adapted PRODIGY,[30] a tool originally designed for patient-specific ranking of cancer driving mutations, to pinpoint drivers of large focal deletions. This tool ranks candidate genes by estimating their cumulative effect on dysregulated pathways, as identified by differential expression analysis. Specifically, given a protein-protein interaction network $G = (V, E, W)$, where W are edge weights representing interaction confidence, a set of differentially expressed genes DEG and a list of dysregulated pathways for a pathway p represented by the graph $G_p = (V_p, E_p)$ the effect of each candidate gene g is estimated as follows:

(1) Constructing a new graph $G_{p,g} = (V_{p,g}, E_{p,g}, W_{p,g}, P_{p,g})$, where: $V_{p,g} = V_p \cup \{g\} \cup N(V_p) \cup N(\{g\})$
$E_{p,g} = E_p \cup \{(u,v) | u,v \in V_{p,g} \text{ and } (u,v) \in E\}$
$$W_{p,g}(u,v) = \begin{cases} 0.1, & u,v \in V_{p,g} \\ 1 - W(u,v), & otherwise \end{cases}$$
$$P_{p,g}(v) = \begin{cases} log(|FoldChange(v)|), & v \in DEG \cap V_p \\ -degree(v)^{\alpha}, & otherwise \end{cases}$$

Here $N(S)$ is the list of direct neighbors of set of vertices S in G and α is a parameter that controls the penalty assigned to nodes with high degree. A value of $\alpha = 0.05$ was used, following the recommendation in the original paper.

(2) Finding an approximate solution to the rooted prize collecting Steiner tree problem,[55] i.e., a sub-tree $G^{`} = (V^{`}, E^{`})$ rooted at g maximizing the score:

$$S = \sum_{v \in V'} P(v) - \sum_{e \in E'} W(e)$$

(3) Assigning a normalized effect score by dividing the score S of the solution by the maximal possible score $S_{max} = \sum_{v \in DEG \cap V_p{}'} log(|FoldChange(v)|)$

The following steps were applied to each chromosomal region.

(1) First, differential expression analysis was conducted, comparing cancer samples with a focal or arm copy-number loss of the region to all other samples.
(2) Second, pathways from the Reactome pathway database[48] enriched for differentially expressed genes were identified by the hypergeometric test.
(3) Third, PRODIGY was used to rank the genes in the region by their cumulative effect on the dysregulated pathways, using the STRING protein–protein interaction network.[56].
(4) Known TSGs appearing in the top 5 ranking genes were proposed as putative drivers.

Driver-based Charm score

For each gene the number of cancer types in which it was nominated as a candidate driver was counted. The new $Charm^{drivers}$ score of an arm was defined as the sum of this count across all of that arm-residing genes.

Pearson's correlations between the arm loss prevalence and both the new Charm score and the original score from Davoli et al.[15] were calculated using the 25 chromosome-arms for which candidate driver genes were identified in our analysis.

Analysis of METABRIC data

Gene level copy number data were used to identify arm loss events and recurrently lost genes and segments: an arm was considered lost if at least 90% of its genes exhibited deletion; significantly lost genes were identified using a binomial test, comparing each gene's loss rate against the background gene loss rate across the genome, and adjacent recurrently-lost genes were subsequently merged into continuous segments. Specifically, for the FD + AL analysis, genes that were lost biallelically at significantly higher rates than background were merged into segments. In segments containing three or fewer genes, the gene(s) with the highest loss prevalence were nominated as candidate drivers. For larger segments, candidate genes were nominated only if independently identified in other cancer types. For the FD-AL analysis, genes that were lost significantly more often than the background rate were merged into segments. PRODIGY was applied to each significantly lost segment. Differentially expressed genes (DEGs) were calculated by comparing gene expression between samples harboring a loss of at least one gene within a segment to samples without any losses in that segment. For the PM + AL analysis, drivers were identified by applying MutSig2CV[27] to all samples harboring the relevant arm loss. Genes located on the corresponding arm with q-value of 0.25 or lower were nominated as candidate drivers.

QUANTIFICATION AND STATISTICAL ANALYSIS

Multiple test correction

Unless stated otherwise, all statistical tests were adjusted for multiple testing using the Benjamini–Hochberg procedure to control the false discovery rate (FDR), with significance defined as FDR <0.05.

Differential expression analysis using IPTW

Weights were calculated using the samples' aneuploidy scores: first, the probability of a sample to harbor an arm loss, $P(\textit{has arm loss})$, was estimated from the data. Then, a logistic regression model $L(p|s)$, estimating the probability p of a sample to have the arm loss given its aneuploidy score s, was trained. Finally, the sample's weight was calculated by dividing the observed probability of a sample to harbor the arm loss by the estimated probability of harboring the arm loss event given the sample's aneuploidy score: $w = \frac{P(\textit{has arm loss})}{L(\textit{has arm loss}|\textit{aneuploidy score})}$. For each gene the log fold-change in gene expression between the two groups was estimated using a weighted version of Cohen's d statistic, applied to $log_{10}(FPKM)$ expression values and using the calculated sample weights: $d = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{{s_w}^2}}$, where $\overline{X}$ is the weighted mean of each group and ${s_w}^2$ is the weighted pooled standard deviation.

Gene set enrichment analysis

For the analysis presented in Figures 7A and 7B, and Data S5 and S6, which incorporate results from multiple GSEA runs, p-values were first corrected across all results using the Benjamini-Hochberg procedure and then only results with FDR <0.05 were taken.

Mutations mutually exclusive to arm losses

For each cancer type in which a particular arm was recurrently lost, a Fisher's Exact test was applied to each gene located on the affected arm. Specifically, the frequency of point mutations in tumors with the arm loss was compared to that in tumors without the arm loss. This test was designed to identify genes mutated at a significantly lower rate in the presence of arm loss, consistent with mutual exclusivity. The p-values from all tests for each CA-CT pair were corrected for multiple testing. While a more precise version of this test would account for the reduced number of alleles for genes in tumors harboring the arm loss, no notable results were identified even with this less stringent approach.

Drivers' enrichment in dysregulated pathways

For each CA-CT pair, differentially expressed genes between the cancer samples with the arm loss and without it were computed as described above, followed by GSEA using 1,293 pathways from the Reactome database. This analysis generated a collection Θ of 33,412 PT-CA-CT triplets, with each triplet representing a pathway (PT) that is dysregulated when comparing samples with a particular chromosome arm (CA) loss to those without the loss in a specific cancer type (CT). This collection was utilized to create a gene-level list Ω consisting of 1,060,326 GN-CA-CT triplets, where a triplet GN-CA-CT with gene GN belongs to Ω if GN belongs to a pathway PT for some triplet PT-CA-CT in Θ. Similarly, a list Δ of 186 DR-CA-CT triplets was created, where DR-CA-CT belongs to Δ if DR was nominated as a driver of that CA-CT pair in our prior analysis and DR-CT-CT belongs to.

As a background set $\Omega^{`}$ for $\Omega$, we used all possible triplets. Specifically, $\Omega^{`}$ consists of all GN-CA-CT such that GN is one of the 10,646 genes appearing in at least one of the 1293 pathways from the Reactome database, and CA-CT is one of the 230 CA-CT pairs. This set contained 2,448,580 triplets. Similarly, a background set $\Delta^{`}$ for $\Delta$ was created, where DR-CA-CT belongs to $\Delta^{`}$ if DR was nominated as a driver of that CA-CT pair in our prior analysis and DR-CA-CT belongs to W ‘ ($\Delta^{`}$ does not contain all 322 identified DR-CA-CT since not all drivers appear in the Reactome database).

Fisher's exact test was applied to test for enrichment of drivers within the dysregulated pathways, comparing Ω and Δ to the background set. An analogous procedure was performed to test for enrichment of drivers within the leading-edge subsets of the GSEA results.

Analysis of METABRIC data

Significantly lost genes were determined with binomial tests (FDR <0.005 for FD + AL; FDR <0.05 for FD–AL).

Chapter 4: Concluding remarks

The results of our study were discussed in the Discussion section of the published manuscript. Here we briefly describe several possible future directions of research.

Extending the framework to identify drivers of chromosome-arm gains

Our current methodology is centered around identifying drivers of arm losses, based on three inactivation patterns. To generalize the approach to arm gains, several adaptations are required. Two of the existing patterns, point mutations co-occurring with chromosome-arm loss (PM + AL) and focal deletions mutually exclusive with chromosome-arm losses (FD - AL), are conceptually extendable with minimal modifications. For gains, the PM + AL pattern could be reformulated by considering oncogenic mutations and arm gains and FD - AL by using focal gains and arm gains.

The pattern of focal deletion co-occurring with chromosome-arm loss (FD + AL), however, relies on the observation that focal deletions co-occurring with arm loss tend to be small, possibly due to the deleterious effects of homozygous deletion of essential passenger elements. This rationale may not hold for gains, where copy-number increases may not impose comparable constraints.

Future work may therefore need to define amplification-specific patterns that exploit unique features of arm gains.

Incorporating additional modes of gene inactivation:

The present framework examines two forms of gene inactivation based on mutations and deletions. Epigenetic silencing, particularly promoter hypermethylation, is another important mechanism by which tumor suppressors are inactivated[23]. Integrating DNA methylation data could enable the identification of genes whose epigenetic repression co-occurs or is mutually exclusive to arm deletions, thereby broadening the set of candidate drivers. This will require utilizing tools for promoter methylation analysis, while also taking into account the relationship with arm deletions.

Extending the approach to non-coding genes

Currently, we only consider the role of coding genes in driving recurring arm deletions. Non-coding genes, including microRNAs, long non-coding RNAs (lncRNAs), and other regulatory elements are increasingly recognized as key contributors to tumor initiation and progression[24]. Some of the tools used in our farmwork are designed only for coding genes, so incorporating non-coding genes will require the use of other tools developed particularly for non-coding genes. Specifically, to incorporate non-coding genes in the mutation-based patterns, one would need to use a tool that assesses the impact of non-coding mutations. To incorporate non-coding genes in the FD - AL pattern, which utilizes PRODIGY and relies on protein-protein interaction networks, one would need tools capable of quantifying the influence of non-coding gene silencing on dysregulated pathways and processes.

References

1. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). The Universal Features of Cells on Earth. In Molecular Biology of the Cell, B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, eds. (Garland Science).

2. Rosenberg, L.E., and Rosenberg, D.D. (2012). The Genetics of Cancer. In Human Genes and Genomes (Elsevier), pp. 259–288. https://doi.org/10.1016/B978-0-12-385212-0.00016-0.

3. Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719–724. https://doi.org/10.1038/nature07943.

4. Taylor, A.M., Shih, J., Ha, G., Gao, G.F., Zhang, X., Berger, A.C., Schumacher, S.E., Wang, C., Hu, H., Liu, J., et al. (2018). Genomic and Functional Approaches to Understanding Cancer Aneuploidy. Cancer Cell *33*, 676-689.e3. https://doi.org/10.1016/j.ccell.2018.03.007.

5. Holland, A.J., and Cleveland, D.W. (2009). Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. Nat Rev Mol Cell Biol *10*, 478–487. https://doi.org/10.1038/nrm2718.

6. Ben-David, U., and Amon, A. (2020). Context is everything: aneuploidy in cancer. Nat Rev Genet *21*, 44–62. https://doi.org/10.1038/s41576-019-0171-x.

7. Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., and Kinzler, K.W. (2013). Cancer Genome Landscapes. Science (1979) *339*, 1546–1558. https://doi.org/10.1126/science.1235122.

8. Sdeor, E., Okada, H., Saad, R., Ben-Yishay, T., and Ben-David, U. (2024). Aneuploidy as a driver of human cancer. Nat Genet *56*, 2014–2026. https://doi.org/10.1038/s41588-024-01916-2.

9. Jones, L., Wei, G., Sevcikova, S., Phan, V., Jain, S., Shieh, A., Wong, J.C.Y., Li, M., Dubansky, J., Maunakea, M.L., et al. (2010). Gain of MYC underlies recurrent trisomy of the MYC chromosome in acute promyelocytic leukemia. Journal of Experimental Medicine *207*, 2581–2594. https://doi.org/10.1084/jem.20091071.

10. Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet *17*, 333–351. https://doi.org/10.1038/nrg.2016.49.

11. LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Res *37*, 4181–4193. https://doi.org/10.1093/nar/gkp552.

12. Karampetsou, E., Morrogh, D., and Chitty, L. (2014). Microarray Technology for the Diagnosis of Fetal Chromosomal Aberrations: Which Platform Should We Use? J Clin Med *3*, 663–678. https://doi.org/10.3390/jcm3020663.

13. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods *5*, 621–628. https://doi.org/10.1038/nmeth.1226.

14. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

15. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology. Nat Genet *25*, 25–29. https://doi.org/10.1038/75556.

16. Kanehisa, M. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res *28*, 27–30. https://doi.org/10.1093/nar/28.1.27.

17. Joshi-Tope, G. (2004). Reactome: a knowledgebase of biological pathways. Nucleic Acids Res *33*, D428–D432. https://doi.org/10.1093/nar/gki072.

18. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences *102*, 15545–15550. https://doi.org/10.1073/pnas.0506580102.

19. Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol *12*, R41. https://doi.org/10.1186/gb-2011-12-4-r41.

20. Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495–501. https://doi.org/10.1038/nature12912.

21. Dinstag, G., and Shamir, R. (2020). PRODIGY: personalized prioritization of driver genes. Bioinformatics *36*, 1831–1839. https://doi.org/10.1093/bioinformatics/btz815.

22. Akhmedov, M., Kedaigle, A., Chong, R.E., Montemanni, R., Bertoni, F., Fraenkel, E., and Kwee, I. (2017). PCSF: An R-package for network-based interpretation of high-throughput data. PLoS Comput Biol *13*, e1005694. https://doi.org/10.1371/journal.pcbi.1005694.

23. Jones, P.A., and Baylin, S.B. (2007). The Epigenomics of Cancer. Cell *128*, 683–692. https://doi.org/10.1016/j.cell.2007.01.029.

24. Esteller, M. (2011). Non-coding RNAs in human disease. Nat Rev Genet *12*, 861–874. https://doi.org/10.1038/nrg3074.

תקציר

סרטן מתפתח כתוצאה מהצטברות הדרגתית של שינויים גנטיים ואפיגנטיים. אנאופלואידיה, המוגדרת כאובדן או הכפלה של כרומוזום או זרוע כרומוזומאלית, היא אחד מהשינויים הגנטיים הנפוצות ביותר בסרטן. למרות זאת, תפקידיה בקידום התהליך הסרטני עדיין אינם מובנים היטב.

בעבודה זו אנו מציגים גישה המשלבת מספר סוגי נתונים על גנים: מוטציות, מספר עותקים וביטוי, כדי לזהות גנים שעשויים להיות אחראים לאובדן תדיר של זרועות כרומוזומליות ספציפיות בסוגי סרטן שונים. הניתוח התבצע על נתונים מ-20 סוגי סרטן שונים תוך שימוש בכ־7,500 גידולים מ-The Cancer Genome Atlas. באמצעות ניתוח דפוסי ההישנות של אובדנים ממוקדים ומוטציות נקודתיות עם אירועי אובדן של זרועות כרומוזומליות שלמות, זיהינו 322 גנים פוטנציאליים הקשורים ל-159 אירועים חוזרים של אובדן זרוע כרומוזומלית. הגישה שלנו מזהה גנים שדווחו בעבר כקשורים לאובדן הזרוע הכרומוזומלית בה הם ממוקמים, כגון *TP53* ו-*PTEN*, ובמקביל חושפת מועמדים נוספים רבים, כולל גנים מדכאי גידול שלא קושרו בעבר לאנאופלואידיה.

בנוסף, אנו משתמשים במידע על שינויים בביטוי גנים המקושרים לאובדן של זרועות כרומוזומליות כדי לזהות שינויים ברמת המסלולים הביולוגיים התורמים להתקדמות הסרטן. שילוב הממצאים הנ"ל מדגיש את חשיבותם של גנים מניעי-סרטן מרכזיים העומדים בבסיס השינויים בביטוי, ומחזק את הרלוונטיות הביולוגית שלהם. בנוסף, אנו מספקים קטלוג מקיף של גנים שעשויים להיות אחראים לאובדן תדיר של זרועות כרומוזומליות בסרטן.

אוניברסיטת תל אביב

הפקולטה למדעים מדוייקים ע"ש ריימונד ובברלי סאקלר

בית הספר למדעי המחשב ובינה מלאכותית ע"ש בלווטניק

זיהוי גנים המניעים אובדן תדיר של זרועות כרומוזומליות באמצעות שילוב נתוני מוטציות, מספר עותקים וביטוי גנים

חיבור זה הוגש כעבודת גמר לתואר "מוסמך אוניברסיטה"

בבית הספר למדעי המחשב ובינה מלאכותית

על ידי

רון סעד

בהנחיית

פרופ' רון שמיר

פרופ' אורי בן-דוד

דצמבר 25