A GENOME GRAPH PARADIGM FOR CANCER GENOME STRUCTURE

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Xiaotong Yao

December 2021

A GENOME GRAPH PARADIGM FOR CANCER GENOME STRUCTURE

Xiaotong Yao, Ph.D.

Cornell University 2021

Cancer genomes harbor structural variants (SV), producing various driver alterations and reflecting the signatures of mutagenesis. With the fast accumulating whole genome sequencing (WGS) data, many complex patterns of SVs have been discovered. Yet, one of the major obstacles to better SV analysis is the lack of flexible and general framework to represent the complexity of SVs. Here, I present genome graph as a general data structure to represent rearranged genomes, and give a mixed-integer quadratic programming algorithm, Junction Balance Analysis (JaBbA), to infer coherent and accurate integer copy numbers (CN) simultaneously for both DNA segments and junctions from the WGS of a tumor sample. With JaBbA, I discover distinct classes of complex rearrangements from 2778 pan-cancer genome graphs, including *tyfonas*, a massive complex amplification pattern different from previously known double minutes or breakage-fusion-bridge cycles. Tyfonas are more effective at producing highly-expressed fusion genes and are specifically enriched in acral not cutaneous melanoma. Next, I use the genome graph framework to reconstruct the evolving SV events in diverging lineages of human fibroblast cells surviving natural telomere crisis and show that telomere crisis can instigate a wide range of chromosomal aberrations, not limited to BFBs and chromothripsis. Lastly, by integrating genome graph in the analysis of 85 whole genomes of lung adenocarcinoma lacking alterations in the RTK/RAS/RAF pathway, I characterized simple and complex SV events abolishing the expression of various tumor sup-

pressors or enhance that of oncogenes. In sum, I show that genome graph and JaBbA is a general, robust analysis paradigm that elucidates the complexity of SVs in cancer genomes, provides a basis for SV evolutionary trajectory inference, and empowers whole genome analysis of cancers.

BIOGRAPHICAL SKETCH

Xiaotong Yao has had strong interest in biology from an early age and later developed an appreciation for using computational methods to answer biological questions. His interest for computational biology started with systems biology and synthetic biology by participating in the 2011 International Genetically Engineered Machine Competition. Subsequently, he curated lipid metabolic pathways in the *Streptomyces avermitilis* metabolic network. To get formally trained in computational biology, he joined the Master's program in Bioinformatics and Systems Biology at the Biology Department of New York University, where he built predictive models for protein sumoylation from public protein function databases. Following a passion for precision medicine, in 2015, he became a PhD student in the Tri-institutional Program in Computational Biology and Medicine at Weill Cornell Medicine, focusing on cancer genomics, and in 2016 he joined Dr. Marcin Imielinski's lab to pursue his dissertation research in using genome graphs to model complex Structural variants in cancer whole genomes. He aspires to continue the research in cancer genomics in the future towards making a more direct impact on the patient outcomes.

To Rosalind Yao and Shan Huang.

ACKNOWLEDGEMENTS

This thesis is founded by my mentor Dr. Marcin Imielinski. In 2016, I was pondering whether a particular recurrent SV event is just a variant calling artifact, when Marcin generously broke me free from the existing framework of counting junctions like mutations, by revealing a possibility where all SVs in a genome can be unified in one graph to reveal their complexity. The elegance of this formulation immediately captured my interest, but it was not until after numerous afternoons sitting down with Marcin, browsing through thousands of genome graphs, improving our models over and over, that I finally felt I started to grasp the significance of this work. During the challenging times of preparing for manuscript submission, responses to reviewers, Marcin's astuteness has been the constant guidance I held onto. Beyond imparting his unique combination of expertise in computational sciences and medicine, working with Marcin instilled in me a swift yet rigorous approach to data analysis and scientific investigation in general, that can best be summarized by a quote from the famous investor Naval Ravikant – *Impatience with actions, patience with results.* With new data at hand there should be immediate list of things burning to be found out; always ask yourself the next question; fail quick and iterate; battle-test your code; motivate every sentence in a presentation or manuscript with a question – are just a few of the teachings that were cast into my tenets. Besides extraordinary professional and academic mentoring, Marcin actively cared for my physical and mental health. For that, I will be forever grateful for the chance to have done this work with Marcin.

Next, I want to thank my wonderful teammates and fellow graduate students, Kevin Hadi, Julie Behr, Aditya Deshpande, who have been offering support and encouragement when things got tricky, critic and reflection when there is room

for improvement. It is a privilege to have friends who are also role models. I also thank the rest of the Imielinski lab members for feedback on all the projects and unconditional support throughout this period.

In true spirit of science as a collective effort, I want to thank all the collaborators whose names I cannot list in an exhausting manner. Working with these brightest minds has been an honor and I will carry the gifts of this experience onwards. Special thanks to Drs. Sally Dewhurst and Titia de Lange for giving me the opportunity to making discoveries on the SV outcomes of telomere crisis; Dr. Jian Carrot-Zhang, for characterizing the whole genome of lung adenocarcinomas; Dr. Nicolas Robine for many projects together and continuous supportive conversations.

Last but not least, I want to thank the rest of my thesis committee Drs. Christina Leslie (chairperson), Simon Powell, and Iman Hajirasouliha for helping me through the academic training.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 From patterns to mechanisms to etiology

Genomic instability is a hallmark of cancer [1] and genomic DNA sequencing of tumor tissues have proved the ubiquity of genomic variants [2]. Even though the idea that chromosomal abnormalities as an intrinsic property of cancer has been proposed [3] for more than a century, the advent of massively parallel sequencing in the past decade has been revolutionary in expanding the known repertoire of somatic variants with unprecedented detail including single nucleotide variants (SNV), small insertions and deletions (INDEL), and structural variants (SV) [4].

The somatic variants in a cancer genome are result of mutagenesis, DNA repair, and evolution [5]. The goal of studying their patterns is to categorize variant into classes, then facotorize the counts of variant of classes among a population into *signatures* (more discussed below), and eventually infer the indogeneous and exogeneous mechanisms from which specific signatures arise.

Such studies commonly combine various features of observed genomic variants, including substitution type, spatial distribution (strand-coordinated clustered mutations by APOBEC [6]), multi-scale sequence context (tri-nucleotide contexts [7]; mesoscale palindromes preference by APOBEC3A [8]), replication timing and transcription level (higher frequency in late replication region and less transcribed regions [9]), 3-dimensional chromatin structures (mutation rate change associated with topological association domain boundaries [10]; lamina

associated domains accumulate more UV-related mutations [11]), tissue specificity (lineage-specific INDELs in lung adenocarcinoma [12], negative correlation between frequency and cell of origin-specific expression [13]), to elucidate the relationships between the genesis and repair of DNA damages in soma ([14, 15, 16]). They are creating wide-reaching impacts on at least three fronts.

First of all, these studies offer insights into the basic biology of genome maintenance and DNA repair. For example, one of the most critical and extensively studied DNA repair pathways, homologous recombination (HR), was associated with excessive accumulation of SNV mutation signature 3 (will discuss the definition of mutation signatures in later sections), tandem duplications, and microhomology-directed deletions [17]. By adapting an *Escherichia coli* replication fork barrier in human cells, Willis et al. proposed that *BRCA1* not *BRCA2* is responsible for suppressing the replication restart-bypass mechanism that result in microhomology-mediated tandem duplications [18].

Second, more textured understanding of the heterogeneous background somatic mutation rate of the genome in various tissue origins and environments improves the detection of positive selection during the evolution of tumors. As one of, if not the most important goals of cancer genomics, determining the exact set of alterations (or *drivers*) out of a vast sea of somatic variants that are responsible for malignant phenotypes of cells has been heavily guided by the analysis of significantly recurrently mutated sites, regions, or genes. Very early on, the data clearly showed that without a well-calibrated background mutation distribution, an enormous amount of false positives will render cancer genome sequencing efforts futile [19]. Consequently, many statistical models are used to correct for variations in background mutation frequencies (MutSigCV [20],

FishHook [12], dNdScv [21]. As the patterns of variants are linked to their mutagenesis mechanisms and are not homogeneous among different tumors, they can contribute to a more accurate estimate of background mutation rates along the genome and further reduce Type I error of driver discovery.

Last but not least, they are bringing novel and profound clinical impact to the prevention, diagnosis, and treatment of cancer. Using the positive association between smoking behavior of the donors and the observed SNV signature 4 burden, Alexandrov et al. [22] estimated each pack of cigarette smoked per day for a year roughly contribute to 150 somatic mutations accumulated in lung tissue. This confirms and complements decades of public health and epidemiology conclusions about the causal link between smoking and elevated cancer risks. In diagnostics, mutation signatures combined with sophisticated machine learning techniques are shown to predict the tissue of origin with mutation patterns alone [23]. Homologous recombination defects of breast and ovarian tumors can be predicted by combining several feature of SNVs, INDELs, and SVs, and guide the prescription of PARP-inhibitor therapy [24].

Philosophically, the studies of variant patterns take two main routes. One is to crystallize latent signatures of mutagenesis pathways out of observed counts of distinct classes of variants in a collection of unrelated genomes. One of the earliest landmark examples is the extraction of *mutation signatures* from a feature space of tri-nucleotide context (the immediate 5′ and 3′ nucleotide of the mutated nucleotide) of SNVs [25]. The general idea is to count the number of variants in each patient that fall into one of the 96 categories (4 possible choices on 5′ and 3′ each times 6 substitution types) and compile a matrix from a large patient cohort. The factorization of this matrix will produce linearly independent

vectors in the feature space and the top of the list represent the more common 'themes' in the cohort. Many of the original mutation signatures have been associated with specific mutagenesis pathways since discovered. For instance, mutation signature 1 has attributed to the spontaneous or enzymatic deamination of 5-methylcytosine to thymine resulting in C>T transitions, and it is correlated with the age of the patient, analogous to a molecular clock for cell divisions [7]. So far this methodology and the resulting signature repertoire have been instrumental in revealing the active mutational processes in many different cancers [26] and have been revised multiple times to reflect the expanding datasets [27]. As exciting as it is, mutation signatures are simply mathematical abstractions of observed data which always requires further molecular evidence to trace back to the responsible mutational processes, which remains challenging. Plus, the repertoire of the signatures is not yet finalized with respect to source data and algorithmic improvements. Last but not least, such approach ideally requires a non-overlapping and complete taxonomy of variant classes, which is easier for small, local variants (SNVs, small INDELs), and still lacking for complex SVs. I will further expand on this point in the next section.

The other study design is to create controlled mutagenesis experiments, for instance in human cell lines, organoids, mouse models, where a particular mutagen is introduced or a particular DNA repair pathway component is disabled for a sustained period, then identify the variants acquired during, to attribute specific types of variants to the corresponding processes. There are successful stories of evaluating the SNV/INDEL mutation signatures induced by exposure to external mutagens in cell models [14]. Specifically for SVs, Umbreit et al. characterized the consequences of mechanically broken dicentric chromosome bridge [28]. Shoshani et al. captured the continuous instability in extra-

chromosomal circular DNA (eccDNA) in response to therapeutic pressure [29]. Ly et al. constructed an elegant model of inactivated centromere that led to missegregation, which caused various complex SVs including chromothripsis and eccDNA [30]. Apparently, such studies are more powerful in establishing the causal links between mutational processes and signatures, but significantly harder to design, especially in the face of intricate interactions between DNA damage and repair mechanisms, bias introduced by positive or negative selection, and the essentiality of many DNA repair genes. After all, it also relies on an adequate framework to describe the variants.

Despite these challenges, the combination of the these two approaches has rapidly and permanently reshaped this field and will be both indispensable. In this thesis, I will demonstrate the data driven approach with my study discovering complex SV patterns in genome graphs with Chapter 3, and the controlled experimental approach in Chapter 4 as I portray the evolution of SVs after telomere crisis.

1.2 Complex structural variants discovered in cancers

Among somatic variants, structural variants (SV) are of particular interest as they are expected to have a bigger phenotypic impact on cells, more complex etiology, and have been linked to various mechanisms of therapy resistance and metastasis [31, 32]. In recent years, many complex patterns of SVs have been discovered.

In [33], Stephens et al. observed the phenomenon and coined the term *chromothripsis*, theorized to originate from shattering a chromosome arm into pieces

and erroneously rejoined a subset of them in random order. Such processes can generate up to hundreds of junctions at random locations and with random orientations, while leaving an oscillating CN profile due to loss of interspersed segments. Since its initial definition, there have been many claims of its prevalence in different tumor types, even non-cancerous somatic tissues. One recent pan-cancer whole genome analyses estimated chromothripsis-like events exist in as many as 29% of pan-cancer genomes [34].

In prostate carcinoma genomes, Baca et al. [35] discovered chained, long-range, reciprocal junctions that are formed through simultaneous multi-loci (>2) translocations. Even though it is also a complex SV event with clustered junctions, its pattern is qualitatively distinct from chromothripsis, with usually smaller footprint, long-range chains, sometimes involving many chromosomes (>=3), and less loss of genetic materials limited to the *deletion bridges* between adjacent junctions along the chain. Furthermore, there are also experimental evidence that androgen receptors are co-recruited with topoisomerase II beta (*TOP2B*) to sites of TMPRSS-ERG fusion and lead to double strand breaks (DSB) [36], indicating that chromoplexy may have a transcription-related origin than a chromothripsis.

As a prime example of using WGS to refresh and upgrade the understanding of an old chromosomal aberration, double minutes (DM) or extrachromosomal circular DNAs (eccDNA), recently have been investigated in a series of studies [37, 38, 39] presenting its high-resolution, complex structure along with histone markers and 3-D chromatin conformations. In subsets of glioblastoma multiforme (GBM) and other central nervous system cancers, certain cancer genes like epidermal growth factor receptor (*EGFR*) can detach from its original chro-

mosome location and exist on eccDNAs, which replicates with other nucleus DNA yet inherited by daughter cells in a non-symmetric stochastic process [40].

In a more systematic fashion, the Pan-cancer Analysis of Whole Genomes (PCAWG) consortium catalogued WGS variants across >2,500 cases spanning 38 tumor types[2] to identify novel classes of complex SVs and cluster these into signatures, mirroring previous work in the categorization of single nucleotide variants (SNVs) into distinct mutational processes[41, 25, 42, 43]. It also builds a taxonomy of basic repertoire of CNA outcomes given a sequence of junctions, and nominated a new event type *templated insertion cycle/chain* (TIC) attributed to template switching of stalled replication forks.

Despite these extraordinary advancements, the discoveries of SV patterns have been outpaced by that of smaller variants. There are several reasons behind the current status. First, the field has not yet converged to a single algorithmic framework to identify and systematically compare the full spectrum of these patterns in a tumor whole genome sequence. For instance, it is unclear whether some of the clustered rearrangement patterns commonly observed in cancer represent variations on known events (e.g. "amplified" chromothripsis) or as yet uncharacterized event classes [2, 41]. Plus, even though some of these studies are prudent in interpreting SV patterns combining both junctions and CNA, they are not unified in a easily-replicable, data-driven way. Many example events shown in publications clearly contain copy number change points without a consistent junction, which are inconsistent with the fact that every *mappable* copy number change point should have a junction "explaining" it (discussed in more details in [44]). Furthermore, various heuristics are often used

to attribute CNA to nearby junctions, making it hard to evaluate performance or replicate results across labs, studies or datasets.

Second, though copy number is primarily a concept applied the dosage of genomic *intervals*, a junction may also be present in one or more copies per cell, and thus be associated with a *junction copy number* (JCN). Elevated JCN might occur through the focal (e.g. extrachromosomal) amplification [45, 38] or whole chromosome (or genome) duplication of an already rearranged allele (Figure 3.3). We argue the topology of JCNs are important pieces of evidence that all efforts hitherto tend to omit or use implicitly, due to the lack of reliable inference methods.

Third, unlike small variants which can be each counted as an irreducible edit, rearranged genome is not simply the sum of all junctions. When junctions overlap over the reference genome, their interpretations are interdependent. Two junctions can be on the same molecule (in *cis*) or different (in *trans*), and their multiplicity allow for different copies of a junction to have different localizations. To actually guess the final derivative sequence resulted from complex SVs, we need a general paradigm that can represent this complexity, and serve as an intermediate scaffold to integrate long-molecule profiling technologies.

As a result, the taxonomy of complex SV is still under-defined, stagnating the discovery of causal mutational processes, and hindering the determination of genomic consequences of known mutagenesis pathways. In the following section, I review the current common practices in identifying and analyzing SVs from WGS data and explain in more detail why it is not ideal for the complex SV patterns in cancer.

1.3 Characterization of SVs with WGS data

Practically, SVs in cancer genomes have been routinely characterized in two separate ways: aberrant junctions and copy number aberration (CNA). On one hand, junctions are two loci associated with orientations (called *breakends*) in the reference genome that are adjacent in the studied genome. When a junction is not connecting two breakends that are already adjacent in the reference genome, it is called an aberrant junction. For ease of discussion, we simply call it junction. There have been many methods and tools to identify junctions from WGS [46, 47, 48, 49, 50], and the evidence they rely on fall into three main categories: discordant read pairs, split reads, and assembled contigs. Depending on the type or combination of types of evidence supporting a candidate junction, the precision of identification can be as good as single nucleotide level. For somatic junction identifications in cancer genomes, the general principle is to differentiate the junctions with supporting evidence in the tumor and not in the matching normal (usually peripheral blood) tissues.

On the other hand, the amount of reads mapped to a certain reference genomic region, termed *coverage*, is a readout for the absolute number of copies of that DNA segment within an "average" cell of the studied sample. CNAs are parts of the reference genome with copy numbers (CN) deviated from 2 (for autosomes). They are detected as locations of shifting center of coverage data [51, 52, 53, 54, 55, 56], or called *copy number change points*. Of course the variations in coverage over the reference genome are affected by a lot of confounding factors other than CN, for instance, GC-content of the reference genome (due to the GC-bias of short-read sequencers [57]), mappability (uniqueness of reference sequences affects false positive and false negative read alignments) [58]. In can-

cer tissue sequencing, there are also mixture of stromal cells, which are assumed to be near diploid and do not possess the tumor-specific CNAs. Elucidated in [53], the purity (proportion of cancer cells in a sample) and ploidy (DNA segment width-weighted average CN) form an affine transformation (preserved collinearity) from coverage to absolute CN, hence can compare across samples.

Despite being inferred orthogonally, junctions and CNAs are two facets of the same latent genome structure. Based on the structure of DNA, apart from whole chromosome, contig, or genome duplications or loss, any change in the copy number of a genomic region require at least one breakage and formation of the covalent 3'-5' phosphodiester bond, which, if mappable at both breakends, should be detected as a junction. Conversely, at any breakend that forms a junction in a rearranged genome, if no perfect reciprocal junction is connected to the other side of the breakage, then there is bound to be loss of some genetic material.

Conceptually, integrating the two procedures should improve both detections. On one hand, junctions provide much more precise localization for copy number change points, and consistent copy number changes with respect to the junction topology should increase the confidence in the junction calls. Indeed, there have been multiple methods that have made such attempts. A CNA inference method, CONSERTING[59], iteratively matches candidate copy number change points with nearby junctions of the consistent orientations and achieved better segmentation than other methods using only coverage evidence. In terms of junction identification, Pedersen and Quinlan developed Duphold [60], that annotates junction candidates with whether the coverage change at the break-

point is convincingly consistent to greatly improve the specificity of SV calls within germline genomes.

However, these improvements are still confined within the variant detection realm and to truly couple these orthogonal readouts and approach complex rearranged sequences, a fundamentally new paradigm of thinking about SVs are desperately needed.

1.4 Complexity of somatic SVs in cancer genomes

For legacy reasons, most analyses on genomic structural variants have been represented with data structures designed for smaller, local variants (like SNVs, short INDELs, and simple duplication and deletion junctions [61]) where a variant is represented as an sequence edit at one or a pair specific locations in the genome. It considers each junction or CNA as an independent edit of a reference locus [62, 63]. This works fine most of the time for germline genomes, whose SVs are dominated by simple classes, namely tandem duplications, simple deletions, and inversions [64]. Tandem duplications and simple deletions are composed of only one junction while inversions and balanced translocations (exceedingly rare in germline but common in cancer) a pair of reciprocal junctions. Although more complex patterns in germline genomes have also been characterized [65, 64], they are, as expected with larger phenotypic effect and lower rate of emergence, rarer than the simpler ones. This approach has apparent advantages including its simplicity, availability of robust tool set like VCF formats and VCFtools [61], and each junction or CNA can be analyzed as

an independent unit, yet it quickly becomes inadequate when the SVs become complex in cancer genomes.

Such school of thought to treat each junction as an independent unit of SV is also influencing the extraction of SV signatures. Many attempts have been published to represent each junction or CNA with a unified feature space like the 96-dimension tri-nucleotide context of SNVs or INDELs. While this approach is valid to the extent of local sequence context of the breakends, and most simple SVs, they are less suitable to describe the properties of complex events which can contain many junctions per event. For example, while HRDetect [24] uses counts of several types of junctions to help predict HRD status, however in the final model only the simple junction counts contributing to the prediction (tandem duplications and deletions of various sizes and breakend microhomology status). The rearrangement signature six which was associated with clustered inversion-like junctions and deletions was not explored or utilized. In another attempt at integrated mutational signature inference, Funnel et al. [66] adpated multi-model correlated topic models to avoid the pitfall of relative sparsity of SVs to SNVs and showed the combined signature achieve better stratification of ovarian cancer patients. One of their important SV features is characterized by fold-back inversions (FBI), a hallmark of BFB cycle-like processes. Nevertheless, the framework cannot provide further details on the exact events that are producing these junctions.

The limitation of this approach in dealing with complex SV events is visible even in the simplest form of complexity. Consider the toy model in Figure 2 of [67], where a unrearranged genomic contig *ABCDE* underwent two nested tandem duplications. When mapped to the reference genome, the second junc-

tion and its assciated CNA appeared to be consistent with a deletion when only seen from the perspective of the reference genome. If we have perfect long-read sequencing for any rearranged allele as such, we may directly observe its structure and could have avoided this illusion. Yet, in practical short-read WGS, when the B/C/D segments are (easily) longer than the hundred basepairs reads, all we can rely on are the local junctions and the change of copy numbers. Thus, a more general data structure is called for to encapsulate all the observed junctions and CNAs without making any assumption about the actual underlying event.

1.5 Representation of SVs with genome graphs

Graphs have been widely used to model genomic sequences for decades, with *de novo* assembly as one of the fields most heavily reliant on diverse yet related data structures like de Bruijn graphs [68], string graphs [69], unitig graphs [70]. Motivated by rapidly increasing resequencing data and pool of known variants, Garrison et al. developed *variation graphs* (vg) [71], compressing all sequence variants found in a population, allowing for variant-aware read mapping for individual samples, and achieving more accurate variant calling and genotyping [72, 73]. Vg creates 'bubbles' (alternative paths) to represent polymorphic sequences and efficiently maps sequencing reads by generalized compressed suffix arrays to avoid biases of aligning to a single linear reference genome. More recently, Li et al. [74] developed *minigraph*, the first method to construct reference pangenome graph from multiple genome sequences of the same species while retaining the coordinates of a linear reference. They showed that this ap-

proach can compactly annotate highly complex and polymorphic regions like the Human Leukocyte Antigen locus.

One commonality of them is that vertices stores the sequence of a contiguous DNA segment (contig) and a path (or cycle) constitute longer sequence and eventually approach a genome. In resequencing studies, we map reads to locations in an existing reference genome, and identify variants based on the discrepancy between reads and the mapped reference sequence. Nevertheless, we can apply a similar idea of modeling, now replacing actual DNA sequences to the genomic intervals they map to in the reference.

There have been several studies employing the concept of *interval graph* or *breakpoint graph* to model the somatic SVs in cancer genomes. Originally, breakpoint graphs are built to solve ancestral genome reconstruction problems based on synteny across species [75, 76]. One of the earliest example of applying a reference genome-based genome graph was from Greenman et al. [77], where vertices represent allelic specific double-stranded segments and edges the adjacencies among them, while each vertex (segment) must have a left and a right boundary and an edge must specify which boundaries of the incident vertices are incident. In another early effort, Oesper et al. [78] mark the left and right boundaries of each double stranded genomic intervals with two vertices and categorize the adjacencies between vertices as either constituting the interval (interval edge) or junction (reference and variant edges). Later, Dzamba et al. 2017 [79], Deshpande et al. [80], Aganezov et al. 2019 [81], Lee and Lee 2021 [82], also adopted this second implementation.

In our own publication [83] and more expansively Chapter 2, we have formulated a directed skew-symmetric graph data structure, called *genome graph*, that

is dual to the variations of interval graphs described above. Like the previous work, we will show that genome graphs can represent genomic rearrangements based on a reference genome and serve as a scaffold to inferring the derivative sequences. Though these implementations are functionally equivalent and only subtly different, a skew-symmetric directed graph formulation made the concept, in our opinions, much more intuitive and immediately compatible with numerous algorithms that are general to directed graphs.

To our knowledge, there has not been a published software package that facilitates the analysis of genome graphs. One that can freely construct any genome graphs, parse the results from the aforementioned methods, serves as the foundation to build useful operations (e.g. subgraph, walk, shortest paths, partition, max/min flow), and characterize biological meaningful events. In Chapter 2, we set out to build such an interface in R programming language using the state-of-the-art genomic interval operations [84] and graph computation (https://igraph.org/). We aim to provide a robust, practical, and versatile tool to popluarize the usage of genome graphs in the whole-genome analysis of cancers. As an example, Chapter 5 will show our collaboration with the Cancer Genome Atlas to characterize the whole genomes of lung adenocarcinomas lacking known alterations in the RTK/RAS/RAF pathway by whole exome sequencing (WES).

Another major obstacle between genome graphs and its wider adoption by the research community is the lack of proper visualization tools. Projection of data onto reference genome coordinates, or genome browsers, has been the quintessential tool of genomics [85, 86]. There have also been several tools designed to visualize SVs at different stages of the analysis. Samplot is a Python

command line tool that generates static images of junction-supporting reads along with coverage data in the vicinity of the junction breakends [87]. Specifically for graph genomes like vg, sequence tube maps [88] takes an creative turn and visualize reads aligned to graph genomes in a similar way to transport networks. Later, MoMI-G [89] attempted to deliver a comprehensive suite for long-read data aligned at complex SV sites with sequence tube maps and circular plots as modular units.

However, for complex structural variants, they are too constrained within one or two regions at a time to convey the full picture. For example, chromoplexy can involve many different chromosomes or very distant loci. One particular tool, Ribbon[90, 91], is designed with SV complexity in mind and is able to put more than two discontiguous windows in one image, yet it is specific for long-read alignments and cannot visualize general genome graphs. To this end, we built a GPU-powered Javascript application *gGnome.js* to interactively browse any genome graph mapped to its reference genome.

1.6 Reconstruction of junction-balanced genome graphs

To represent coherent double stranded genomic DNA, or karyotype, we need to annotate genome graphs (both vertices and edges) with non-negative integer copy numbers, and because of the linear structure of DNA, we find that these copy numbers need to follow Eq. 2.1. Such graphs are named junction-balanced genome graphs. In Chapter 2, we will show that junction balance arises naturally from compressing all the linear alleles.

Despite its clear conceptual advantage, genome graph has not been ubiqui-

tously adopted as the mainstream framework for SV analysis. The main reason is that the numeric quality of the copy numbers in the reconstructed graphs have not shown high enough fidelity for the heterogeneous landscape of complex somatic SVs, as well as varying sample purity.

There have been several approaches proposed to reconstruct junction-balance genome graphs with different objectives, underlying models, inference methods, and scopes (Table 2.1, [78, 92, 93, 79, 80, 41, 82, 81]). In Chapter 2, I will present a mixed-integer quadratic programming method, Junction Balance Analysis (JaBbA) to infer coherent and optimal copy numbers of a genome graph from WGS data. Besides, I will show a comprehensive benchmarking of some of the aforementioned methods and discuss the potential reasons for their strengths and limitations. With recent improvements and proof of application in large pan-cancer WGS cohorts [83], it shows a promising trend that genome graph-based SV analyses are becoming robust enough to power general SV analysis.

1.7 SV pattern discovery guided by junction copy numbers

The inference of JCN in WGS involves the fitting of a genome graph to read depth and junction data through the application of a junction balance constraint, which requires the CN of every genomic interval (i.e. vertex) to be in balance with the JCN of its neighboring junctions (i.e. edges) (Figure 2.1, 2.3)[94, 77, 78, 93, 79, 92]. We hypothesized that the topology and CN of both vertices and edges on junction-balanced genome graphs might reveal novel classes of complex SV events and mutational processes.

In particular, high junction copy numbers (JCN) must arise through subsequent alterations that replicates existing junctions, and hence are an indication of complex amplification events. As touched upon before, there are well known mechanisms like double minutes and breakage-fusion-bridge cycles active in various cancers and WGS has painted detailed structures of them. Yet, the same mechanism can generate a wide range of complexity [80] and it is unknown if there are other amplification patterns. One such example is neochromosomes found in several liposarcoma cell lines that do not fit in either DM or BFB cycles [95]. Hence, a systematic, data-driven classification of complex amplicons in a comprehensive dataset is crucial to answer these questions.

In the first half of Chapter 3, I will show the discovery of 3 stable classes arising from hierarchical clustering of amplified subgraphs from out pan-cancer genome graphs and show that tyfonas is a distinct amplification pattern.

1.8 Stratifying pan-cancer patient cohort by burdens of SV events

Using SV events as a biomarker has always been a top desired application of WGS. Early efforts includes using the burdens of tandem duplication and deletions of various sizes as a features to predict homologous recombination deficiency (HRD, [24]). More recently, the existence of extrachromosomal circular DNA (eccDNA) has been linked to excessive replication stress in the cells. One necessary indication that SV events may possess such discriminating power is to show that non-random grouping patterns among patients can arise based solely on their burdens. For instance, like mentioned above, in [35], almost half

of studied prostate adenocarcinoma contain chromoplexy patterns, then if we can see a subgroup of patients differentiated by chromoplexy presence from the others, and it happens to enrich in prostate adenocarcinoma, it should be a positive indication that our patterns are confirming previous results and add to our confidence about new results.

In the second half of Chapter 3, I show our effort in using the 13 simple to complex SV event types to reveal 13 groups of patients in the pan-cancer cohort and their associations with tumor type, genotype, and overall survival.

1.9 Evolution of SVs after telomere crisis

The studies of the relationship between variant outcomes and mutational processes usually take one of two directions. Most of the studies including our efforts described above are searching for common patterns from (hopefully unbiased) observations about genomes, and trace back to their possible mechanisms. However, more traditionally, the genome maintenance and DNA repair research community has always taken the more hypothesis-driven approach to create controlled experiments of mutagenesis. However, there is significant gap between the most advanced variant profiling technology and pattern recognition methods and the delicately designed experimental systems where we can clearly attribute any DNA changes to a particular mutagenesis process. Furthermore, this gap can be complicated by the fact that the same "type" or "class" of variants can be generated by multiple mechanisms, and a mechanism leads to diverse outcomes. Therefore, the integration between advanced analysis and finely controlled experimental systems are the key moving forward.

At this point, we have demonstrated the power of our genome graph paradigm in delineating SV events from large observational pan-cancer WGS cohorts. Beginning the next chapter 4, I will use it to investigate the chromosome structural consequences of an crucial and (nearly) universal mutagenesis process, telomere crisis.

In Chapter 4 I will dive deep into this detailed characterization of the structural outcomes of telomere crisis, and discuss about the implications of adopting our genome graph framework to resolve evolving genome structures among multiple related samples in the Concluding Remarks 6.

1.10 Genome graphs facilitates whole genome analysis of tumors

The most important goal of sequencing cancer genomes is to pinpoint the somatic alterations responsible for the malignant phenotype, or *drivers*. To achieve that, the main signal to rely on is the recurrent alteration of a particular region of the genome from independent patients tumors. To put simply, each patient's tumor can be regarded an independent clonal expansion, if a particular genomic region is altered in a cohort more than expected from a null distribution where the region is not positively selected, it can be nominated as selected in these cancers. In theory, whole genome sequencing (WGS) is a more ideal tool to profiling the cancer genomes for such purpose since it samples reads from the whole genome, while whole exome sequencing (WES) only mainly captures the known exonic sequences. Yet, limited by higher cost and difficulty to analyze, WGS was not the primary choice of earlier cancer genome profiling efforts. For example, the Cancer Genome Atlas (TCGA), as one of the earliest large-scale cancer sequencing projects, relied mainly on WES, while powerful at discovering recurrent coding mutations, also misses by design non-coding mutations, and the vast majority of SVs.

Having discussed how our genome graph paradigm can represent complex

genome structures and be used to study SV patterns, I will take it a step further and show how it enables a more comprehensive whole genome analysis of tumors by better modeling of the rearrangements. More specifically, as part of the next phase of TCGA effort to push for whole genome characterizations, we focused on a subset lung adenocarcinomas (LUAD) in which WES and RNA-seq (for fusion transcripts) did not assign a clear driver. In the last Chapter 5,

CHAPTER 2

JUNCTION-BALANCED GENOME GRAPHS REPRESENT STRUCTURALLY ALTERED GENOMES *

In this chapter, I give formal definitions of a genome graph, a junction balanced genome graph, and Junction Balance Analysis to infer integer copy numbers from short-read whole genome sequencing data. Taking advantage of our novel interactive genome browser *gGnome.js*, I show that our complete tool set can serve as an intuitive portal for researchers to incorporate genome graphs in WGS studies.

Part of this chapter is described in our published article [83]*, led by Kevin Hadi, Julie Behr, Marcin Imielinski, and myself.

Individual contributions: *Marcin Imielinski and I conceptualized, formulated, and implemented JaBbA for Junction Balance Analysis. Kevin Hadi and I collected and processed the WGS data from simulations and HCC1954, HCC1143 cell lines. I carried out the benchmarking experiments of JaBbA. Marcin Imielinski and I proposed and developed gGnome, and Kevin Hadi, Julie Behr, Zi-ning Choo, Alon Shaiber, Joe DeRose have contributed. Charalampos Xanthoupolakis, Marcin Imielinski, and I designed the gGnome.js genome browser for genome graphs.*

*Hadi, K., Yao, X., Behr, J.M., et al. Distinct classes of complex structural variation uncovered across thousands of cancer genome graphs. Cell, 183(1):197–210, 2020.

2.1 Genome graph as a general data structure to represent rearranged genome

2.1.1 Reference genome

Let the reference genome C comprise c pairs of strings, labeled C^i and $C^{-i}, i \in 1, \ldots, c$. Each string pair $C^{\pm i} = \{C^i, C^{-i}\}, ie \in 1, \ldots, c$ is called a *chromosome*, and each string in that pair is called a *strand*. We use C^i and C^{-i} to refer to "positive" and "negative" strands of chromosome i, each having length $L_i \in \mathbb{N}$. We use brackets to denote substrings on these strands. For example, $C^i[q, r]$ refers to the substring of C^i beginning at position q and ending at position r (inclusive) where $q \leq r \in 1, \ldots, L_i$. We also use C^i_q as a shorthand for $C^i[q, q]$. In the remainder of this dissertation, we call $C^i[q, r], i \in 1, \ldots, c, q \leq r \in 1, \ldots, L_i$ a signed *genomic interval*. The width of interval $C^{-i}[q, r]$ is $r - q + 1$.

Every signed genomic interval has a "reverse complement" $C^{-i}[q, r]$ and together they make a double stranded interval $C^{\pm i}[q, r]$, termed *segment*. Two segments $C^{i_1}[q_1, r_1], C^{i_2}[q_2, r_2]$ are said to *overlap* if $i_1 = i_2$ and either $q_2 \leq r_1$ or $q_1 \leq r_2$, and a set of intervals are said to be *disjoint* if none of its members overlap. As a trivial example, one chromosome $C^{\pm i}$ is a segment of width L_i.

As is the convention in defining reference genomes, on the positive strand $C^i, i > 0$, the coordinate increases from 5′ end (left) to 3′ end (right). In other words, C^i_q is the 5′ end of $C^i[q, r]$, and C^i_r its 3′ end. Conversely, C^{-i}_r is the 5′ end of $C^{-i}[q, r]$, and C^{-i}_q the 3′ end. We term the 3′ end the outgoing end b_o and the 5′e end receiving end b_r. Based on skew symmetry, $b_o(v) = b_r(\bar{v})$ and $b_r(v) = b_o(\bar{v})$. A pair of reverse complement interval ends of the opposite sides,

for example, $b = \{b_r(v), b_o(\bar{v})\}$ or $b = \{b_o(v), b_r(\bar{v})\}$ is termed a *breakend*. When marking the location of a breakend $b = \{b_r(v), b_o(\bar{v})\}, v = C^i[q, r], i > 0$, we simply use a width 1 signed interval $b = C^i_q$, here the positive sign of i indicates the breakend belongs to the interval continuing onto the right of the coordinate q, while for breakend $b = \{b_o(v), b_r(\bar{v})\}$, $b = C^{-i}_r$.

Taken the above definitions together, each segment $s = C^{\pm i}[q, r]$ has two breakends, left (smaller coordinate) $b_L(s) = \{b_r(v), b_o(\bar{v})\} = C^i_q$ and right (larger coordinate) $b_R(s) = \{b_o(v), b_r(\bar{v})\} = C^{-i}_r, v = C^i[q, r], i > 0$.

2.1.2 Genome graph

Based on a reference genome C, we define genome graph $G = (V, E)$, where the vertices V is a multi-set of signed intervals, and edges E is a set of directed adjacencies representing the 3′-5′ phosphodiester bonds between vertices. Since genomic DNA is a pair of reverse complement strands, we restrict G to be skew-symmetric [96] with the mapping function "reverse complement" denoted with a bar over, such that for any vertex $v \in V$, there exists its *symmetric* vertex $\bar{v} \in V$, and for any edge $e = (v_1, v_2) \in E$, $\bar{e} = (\bar{v_2}, \bar{v_1}) \in E$. Intuitively, each reverse complement pair of vertices represents a double stranded DNA segment. Since any vertex maps to a genomic interval, we mark its location in the reference genome with $v = C^i[q, r]$.

Because the phosphodiester bonds are always connecting the outgoing (3′) end of a source vertex $v_1 = C^{i_1}[q_1, r_1]$ and the receiving (5′) end of a sink vertex $v_2 = C^{i_2}[q_2, r_2]$, we can denote the genomic location of an edge $e = (v_1, v_2)$ with a tuple of the connected breakends $e = (b_o(v_1), b_r(v_2))$. For a node v we name the

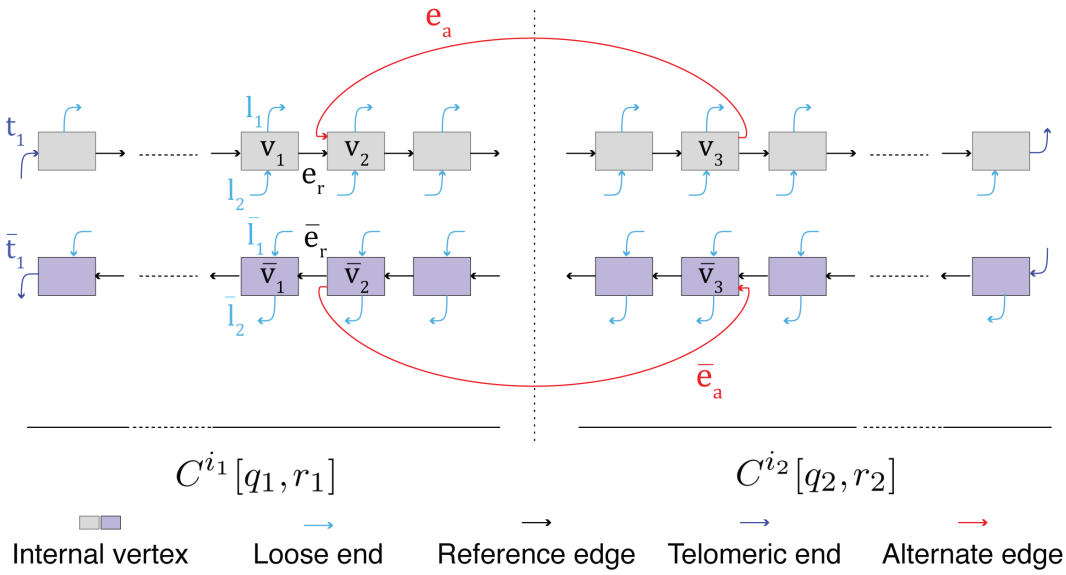

Figure 2.1: **Schematic of a genome graph.**

Grey and purple rectangles are positive and negative strand intervals. Links are edges of different types: black, reference; red, alternate or aberrant; blue, loose ends; dark blue, telomeric ends.

set of incoming edges to its 5′ end $E_{-}(v)$ and the set of outgoing edges from its 3′ end $E_{+}(v)$.

Following this definition, there are at least two types of edges. The first type, *REF* edges are the ones that are connecting breakends adjacent in the reference genome hence always with location $e_r = (C^i_q, C^i_{q+1})$, if $i > 0$, or $(C^i_{q_1+1}, C^i_{q_1})$, if $i < 0, q \in [1, L_i)$. In contrast, the second type *ALT* edges are neo-adjacencies that are not present in the reference genome and resulting from rearrangements (**Figure 2.1**).

Like a reverse complement pair of vertices $\{v, \bar{v}\}$ compose a double-stranded DNA segment, a reverse complement pair of edges $\{e, \bar{e}\}$ compose a *junction*. Trivially, e and $\bar{e}$ will always be the same type of edge, so we also have two

types of junctions, REF and ALT. Each junction $a = \{e, \bar{e}\}, v_1 = C^{i_1}[q_1, r_1], v_2 = C^{i_2}[q_2, r_2]$ can be represented with the two breakends it connects, namely $b_1 = \{b_o(v_1), b_r(\bar{v_1})\} = C^{i_1}_{q_1[1-I(i_1)]+r_1 I(i_1)}$, and similarly $b_2 = C^{i_2}_{q_2[1-I(i_2)]+r_2 I(i_2)}$, where indicator function $I : \mathbb{R} \rightarrow \{0, 1\}, I(x) = 1, if\, x \geq 0, I(x) = 0, if\, x < 0$.

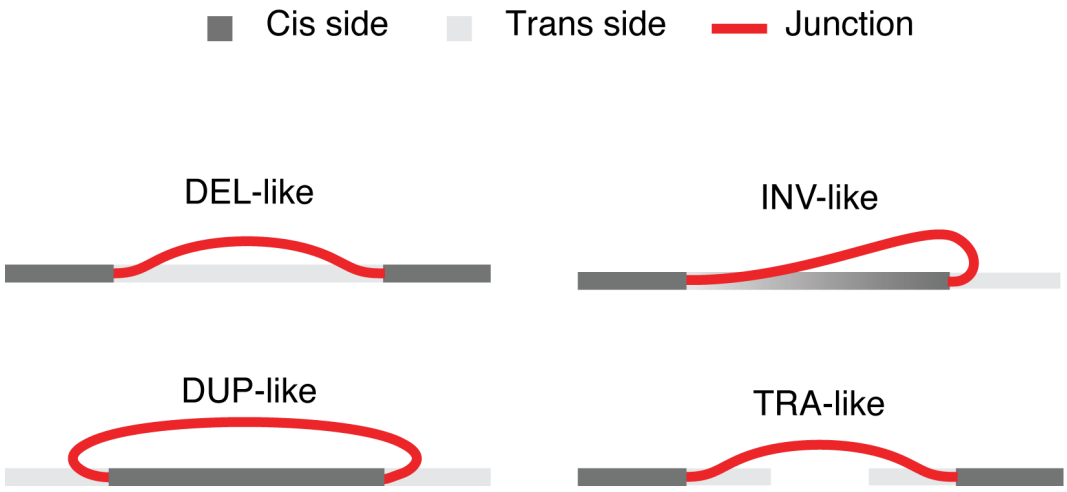

Figure 2.2: **Four basic classes of junctions.**

Dark gray indicates the *cis* side and light gray the *trans* side of a breakend.

For a breakend location in the middle of a chromosome, only one of its two sides is incident to the junction, we call this side the *cis* side for it is the part of the genome that will be on the derivative molecule after introducing the junction, and the other *trans* side. Based on the relative locations and orientations of its two connected breakends, we can classify a junction into four basic classes (**Figure 2.2**): DUP-like (tandem duplication), DEL-like (deletion), INV-like (inversion), TRA-like (translocation), named after the most common and simplest SV events that can give rise to such junctions. In WGS studies, these junctions are the direct products of SV calling methods like SvABA [97], GRIDSS [47], Delly [48], Lumpy [50], and many more. In the following section we show a basic algorithm to build a genome graph from a set of breakends and junctions.

2.1.3 Build genome graph from breakends and junctions

To describe a rearranged and copy number altered reference genome, we partition C according to a collection of *breakends* $\mathcal{B}$. We also define a set of *junctions* $\mathcal{A}$ representing alternative adjacencies between a set of the breakends in $\mathcal{B}$.

Each $b^i = C^{\in}\mathcal{B},\ i \in 1,\ldots,c$ is an ordered and unique sequence of integer coordinates $B^i = (B^i_k), 1 \leq B^i_k \leq L_i$ on chromosome i, where $B^i_1 = 1$ and $B^i_{|B^i|} = L^i$. Each junction $A \in \mathcal{A}$ is a tuple $(i_1, r_1, i_2, r_2),\ r_1 \in B^{|i_1|}, r_2 \in B^{|i_2|}, |i_1|, |i_2| \in 1,\ldots,c$ representing a (3′-5′ phosphodiester) bond between the position $r_1 + \frac{-sgn(i_1)+1}{2}$ on chromosome / strand C^{i_1} and position $r_2 + \frac{sgn(i_2)+1}{2}$ on chromosome / strand C^{i_2}. For every adjacency $A = (i_1, r_1, i_2, r_2) \in \mathcal{A}$ we require $\mathcal{A}$ to contain the reverse complement adjacency $\bar{A} = (-i_2, r_2, -i_1, r_1)$. The adjacencies in $\mathcal{A}$ are "alternative" relative to a set of "reference adjacencies" $\mathcal{R}$ implied by B^i, comprising tuples (i, B^i_k, i, B^i_k) and $(-i, B^i_k, -i, B^i_k)$ for each breakend $B^i_k, k \in 1,\ldots,|B^i|$ in each chromosome $i \in 1,\ldots,c$.

Algorithm 1: BuildGraph

procedure BUILDGRAPH($\mathcal{A}$, $\mathcal{B}$) ▷ Returns genome graph $G = (V, E, \psi, \phi)$

  $G \leftarrow emptyGraph$; $\phi, \psi \leftarrow emptyDictionary$

  $\mathcal{B} = \mathcal{B} \cup getBreakPoints(\mathcal{A})$ ▷ Extract additional breakends from junctions in $\mathcal{A}$

  for $B^i_k,\ k \in 1, \ldots, |B^i| - 1,\ c \in 1, \ldots, |C|/2$ **do**

    $v, \bar{v} \leftarrow newVertices(2, \psi, "I")$ ▷ $newVertices$ updates ψ to map v and $\bar{v}$ to "I"

    $\phi(v) = (i, B^i_k, B^i_{k+1} - 1)$

    $\phi(\bar{v}) = (-i, B^i_k, B^i_{k+1} - 1)$

    $l_1, \bar{l}_1, l_2, \bar{l}_2 \leftarrow newVertices(4, \psi, "L")$

    $e_1, \bar{e_1}, e_2, \bar{e_2} \leftarrow newEdges(\{(l_1, v), (\bar{v}, \bar{l}_1), (v, l_2), (\bar{l}_2, \bar{v})\}, \psi, "L")$

    if $k = 1$ **then** $\psi(\{l_1, e_1, \bar{l}_1, \bar{e_1}\}) \leftarrow "T"$ ▷ Label telomeric loose ends with "T"

    if $k = |B^i| - 1$ **then** $\psi(\{l_2, e_2, \bar{l}_2, \bar{e_2}\}) \leftarrow "T"$

    $V(G) \leftarrow V(G) \cup \{v, \bar{v}, l_1, \bar{l}_1, l_2, \bar{l}_2\})$

    $E(G) \leftarrow E(G) \cup \{e_1, \bar{e_1}, e_2, \bar{e_2}\})$

  end for

2.1.4 Walks on a genome graph map to rearranged karyotype

Analogous to an assembly graph, on genome graph G, traveling through a walk $h = (v_1, e_1, v_2, e_2, \ldots, e_{k-1}, v_k), k \in \mathbb{N}^+$, where each edge $e_i = (v_i, v_{i+1}), i \in 1, 2, \ldots, k-1$, is equivalent to generating a substring of DNA sequence, simply by concatenating the reference sequences of the vertices. If some of the edges along the walk belong to ALT edges E_A, then the walk represents a pos-

Algorithm 1 (Continued)

$\mathcal{R} = referenceAdjacencies(\mathcal{B})$ ▷ Get implied reference adjacencies from breakends
for $A = (i_1, r_1, i_2, r_2) \in \mathcal{A} \cup \mathcal{R}$ **do**
  if $sgn(i_1) > 0$ **then**
    $v_1 \leftarrow getVertex(\{\hat{v} \mid \phi(v) = (i_1, q, r_1 + \frac{-sgn(i)+1}{2}), v \in V(G), q \in \mathbb{N}\})$
  else
    $v_1 \leftarrow getVertex(\{\hat{v} \mid \phi(\hat{v}) = (i_1, r_1 + \frac{-sgn(i)+1}{2}, q), v \in V(G), q \in \mathbb{N}\})$
  end if
  if $sgn(i_2) > 0$ **then**
    $v_2 \leftarrow getVertex(\{\hat{v} \mid \phi(v) = (i_2, r_2 + \frac{sgn(i_2)+1}{2}, q), v \in V(G), q \in \mathbb{N}\})$
  else
    $v_2 \leftarrow getVertex(\{\hat{v} \mid \phi(\hat{v}) = (i_2, q, r_2 + \frac{sgn(i_2)+1}{2}), v \in V(G), q \in \mathbb{N}\})$
  end if
  if $A \in \mathcal{R}$ **then** $e = newEdges((v_1, v_2), \psi, "R")$ **else** $e = newEdges((v_1, v_2), \psi, "A")$
  $E(G) \leftarrow E(G) \cup \{e\}$
end for
return $(V(G), E(G), \psi, \phi)$
end procedure

sible derivative sequence resulted from the SVs that gave rise to these edges. Since G is skew-symmetric, every walk h has a reverse complement walk $\bar{h} = (\bar{v_k}, \bar{e_{k-1}}, \ldots, \bar{e_2}, \bar{v_2}, \bar{e_1}, \bar{v_1})$. We also denote the number of times a walk h travel through a node v or an edge e using $\delta : (V \cup E, H) \rightarrow \mathbb{N}$. Based on this definition, we have two important observations.

First, given a tuple of genomic intervals $(v_1, v_2, \ldots, v_k)$ whose order implies sequential covalent bonding, we can easily produce a genome graph and represent the tuple as a walk on that graph, by filling in edges between each consecutive pair of intervals based on their corresponding incident breakends, then add the same for the reverse complement walk. For example, the edge joining (v_1, v_2) should be $(b_o(v_1), b_r(v_2))$. This observation implicated that any DNA sequence, as long as it can be constructed from pasting substrings of the reference genome, can be converted to a walk and a genome graph.

Second, every interval has exactly one upstream neighbor (incident to its 5′ breakend, except for the first interval), and one downstream neighbor (incident to its 3′ breakend, except for the last interval), as genomic DNA is known to be linear. Thus, we can associate a walk with a non-negative integer $\kappa(h) \in \mathbb{N}$ as the multiplicity of such molecule present in a genome. Consequently, the graph arise from these walks must posses be *junction balanced*, a property we are going to define in Subsection 2.1.5.

2.1.5 Junction-balanced genome graph

We define a mapping $\kappa : \{V \cup E\} \rightarrow \mathbb{N}$ of non-negative integer copy number (CN) to vertices and edges of G, where $\kappa(v), v \in V$ and $\kappa(e), e \in E$ represent the CN of vertex v and edge e, respectively. For a graph $G = (V, E)$ that arise from the aforementioned procedure from a set of tuples of genomic intervals, the following junction balance constraint is always true:

$$\kappa(v) = \sum_{e \in E^{-}(v)} \kappa(e) = \sum_{e \in E^{+}(v)} \kappa(e) \tag{2.1}$$

This is because to travel into v once one must travel through one of the edges in $E_{-}(v)$, and vice versa for the outgoing edges. In plain language, the principle of *junction balance* constrains the CN of every vertex to be equal to the sum of its incoming edges and the sum of its outgoing edges. Formally,

$$\begin{aligned} \forall v \in V, \kappa(v) &= \sum_{h \in H} \kappa(h)\delta(h, v) \\ \forall e \in E, \kappa(e) &= \sum_{h \in H} \kappa(h)\delta(h, e) \end{aligned} \tag{2.2}$$

in which $\delta : \{H, V \cup E\} \rightarrow \mathbb{N}$ is the number of times h travels through a vertex or an edge $\sum_{j \in 1,\ldots,k} I(v = v_j)$, with the only exception that if $h = (v_1, e_1, v_2, e_2, \ldots, e_{k-1}, v_k), v_1 = v_k, k \in \mathbb{N}^+$ is circular $\delta(h, v_k) = 0$, so that v_k is not double counted.

In addition, since double-stranded DNA require both strands to have the same CN, we require the CN κ to obey *skew-symmetry*, which means that every vertex or edge must have the same copy number as its reverse complement.

$$\kappa(v) = \kappa(\bar{v}), \ \forall_{v \in V} \quad \kappa(e) = \kappa(\bar{e}), \ \forall_{e \in E} \tag{2.3}$$

We call the combination (G, κ) for which κ satisfies Eqs. 2.1-2.3 a *junction-balanced genome graph* (JBGG).

2.1.6 Decomposing a JBGG to walks

In a perfect scenario where we have infinitely long sequencing reads, we can read out any sequence directly and fully (telomere to telomere), for any rearranged karyotype. However, in practice, sequencing reads from existing technologies are fragmented and limited in length, so the walks representing complete and true karyotypes are always latent. Despite current long-read sequencing provides us a better chance at obtaining the rearranged allele, it is still "local" relative to the scale of some of the most long-range complex SV events like tyfonas [83] (see also Chapter 3). Furthermore, due to the differences in cost, accuracy, throughput, and analytical robustness, the overwhelming majority of cancer WGS data is still based on massive parallel short-read sequencing, with which SV events are detected as copy number aberrations and junctions. De-

spite certain challenges, for instance with SVs in repetitive regions, short-read WGS has been found to approach the complete SV repertoire well [44].

Though we cannot directly read the derivative sequence after rearrangement, from junctions we can always build a genome graph (see Algorithm 1 in later sections). If we have the complete knowledge on all junctions, then this graph will encapsulate every possible karyotype that could result from them, hence must contain the set of walks representing the true sequence (proof not shown). The enumeration of all possible walks from all the source nodes to all the sink nodes is analogous to elementary flux mode of a stoichiometric matrix [98], and produces a set of paths and cycles H where the latter can be further embedded to form more complex walks.

However, such genome graphs only deal with the topology of junctions, and for a coherent double-stranded DNA model, the vertices and edges' dosages need also be constrained by Eq. 2.1 and Eq. 2.3. With that we can define the reverse problem of Eq. 2.2,

$$\begin{aligned} &\text{Given } G = (V, E), \kappa(G) \\ &\quad \text{find } \kappa(h) : H \to \mathbb{N} \\ &\qquad s.t. \forall v \in V, \kappa(v) = \sum_{h \in H} \delta(h, v)\kappa(h) \\ &\qquad\quad \forall e \in E, \kappa(e) = \sum_{h \in H} \delta(h, e)\kappa(h) \\ &\qquad\quad \kappa(h) = \kappa(\bar{h}) \end{aligned} \tag{2.4}$$

Since in its essence, this is equivalent to a flow decomposition problem, the junction balance constraints 2.1 guarantee that it is feasible, usually with a huge number of possible solutions. Upon this pool of possible karyotypes, we can

further define different loss functions, such as parsimony (number of unique walks with $\kappa(h) > 0$), to reach some optimal guesses at what the actual rearranged allele is that constitute the input JBGG. Taking the most naive formulation as an example, to minimize the total number of unique walks to reconstitute an input JBGG, formally the question is defined as,

$$
\begin{aligned}
&\text{Given a JBGG } G = (V, E), \kappa(G) \\
&\quad \underset{\kappa(h): H \to \mathbb{N}}{\text{minimize}} \sum_{h \in H} [\![\kappa(h)]\!] \\
&\qquad s.t. \forall v \in V, \kappa(v) = \sum_{h \in H} \delta(h, v)\kappa(h) \\
&\qquad\quad \forall e \in E, \kappa(e) = \sum_{h \in H} \delta(h, e)\kappa(h) \\
&\qquad\quad \kappa(h) = \kappa(\bar{h})
\end{aligned}
\tag{2.5}
$$

We shall see an example of solving a variation of this problem in Chapter 4.

When combined with read coverage data, we can further infer JBGG by filling in an optimal set of CNs for all the vertices and edges. We call this problem the inference or *reconstruction of JBGG*, to which we will present one solution in the next section.

2.2 Junction Balance Analysis infers copy numbers on genome graphs from short-read whole genome sequencing

We developed an algorithm, JaBbA, to accurately infer junction-balanced genome graphs from WGS data. In the following three subsections, we will 1) formally define this algorithm *junction balance analysis*, 2) describe the pipeline implementing the algorithm, and 3) comprehensively compare its features and performance to other methods of similar purpose.

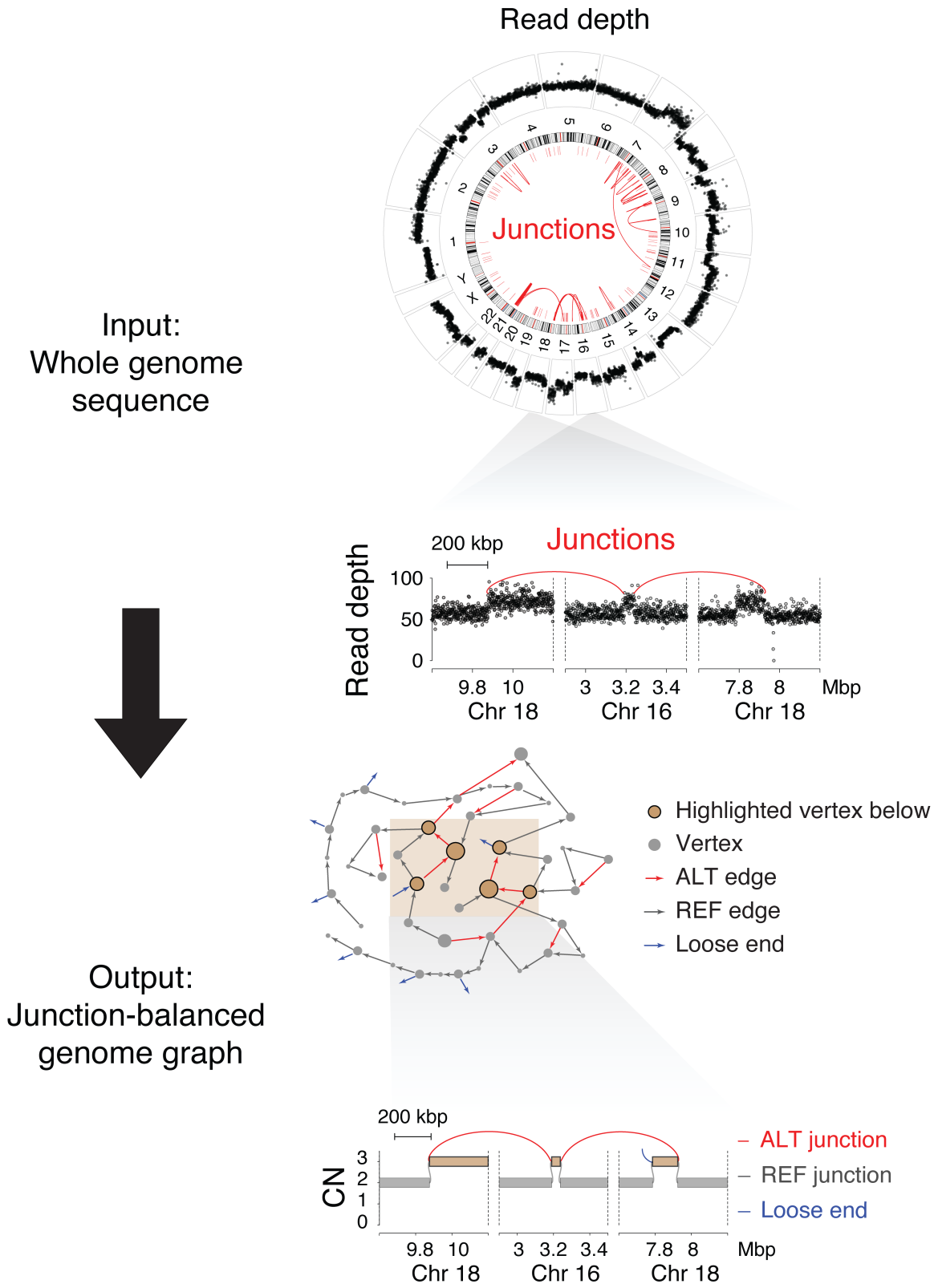

Figure 2.3: **Inputs and outputs of Junction Balance Analysis**

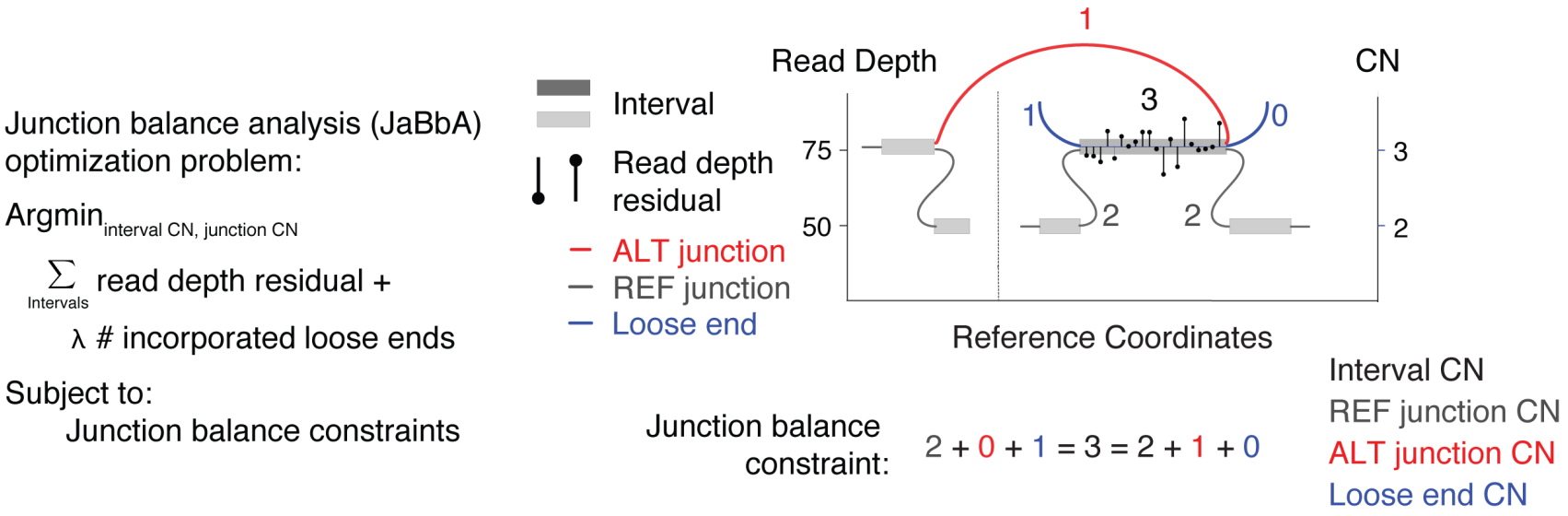

Figure 2.4: **A schematic of the mixed integer quadratic program optimization problem that JaBbA solves.**

2.2.1 Formulation of the mixed-integer quadratic programming problem

We infer JBGGs from a genome graph G and binned, normalized, and purity / ploidy-transformed read depth data $x \in \mathbb{R}^n$ across n genomic bins (see below for read depth transformation details) through the solution of a mixed integer quadratic program (MIQP), which assigns an integer CN $\kappa : V_I \cup E \rightarrow \mathbb{N}$ to the vertices and edges of G. The genome graph G is generated, as above, from a set of breakends $\mathcal{B}_{seg}$ obtained from a preliminary segmentation of genome-wide read depth (i.e. via segmentation software such as CBS) and a set of junctions $\mathcal{A}$ (i.e. from a junction caller such as SvABA or DELLY) (Figure 2.3). In Figure 2.3: Top left, read depth on genomic bins and junctions are the two required input data summarized from a WGS. Bottom left, an example locus showing read depth as scatter plots over reference coordinates and input junctions as red curves connecting the breakends. Top right, the network layout of the genome (sub)graph of the highlighted locus built from the junctions. Bottom right, the junction balanced genome graph (JBGG) output with integer CN for segments on the Y axis and junction CN implied.

Each vertex $v \in V_I(G)$ is associated with a partition of bins $J(v) \subseteq \{1, \ldots, n\}$ (based on genomic coordinate overlap) and a mean bin value $\rho(v) = \frac{1}{|J(v)|} \sum_{j \in J(v)} x_j$. We model each bin subset $x_{J(v)}$ as an i.i.d. sample from a Gaussian distribution with standard deviation $\sigma(v)$ and mean $\kappa(v)$. The log likelihood is

$$logP(x_{J(v)}|\kappa(v), \sigma(v)) = \sum_{j \in J(v)} log\ \mathcal{N}(x_j|\kappa(v), \sigma(v)^2) = -\mathcal{V}(v, \kappa, x, J) + Const(\kappa) \quad (2.6)$$

where $\mathcal{N}(\mu, \sigma^2)$ is the Gaussian probability density function with mean μ and

variance σ^2 and $\mathcal{V}(v, \kappa, x, J) = \frac{|J(v)|}{2\sigma(v)^2}(\rho(v) - \kappa(v))^2$ is the *read depth residual* of vertex v. The variance $\sigma^2(v)$ is a κ-independent parameter that models read depth noise and is computed directly from the data. The simplest noise model is a constant, where this parameter is set to the genome-wide sample variance of the read depth around each vertex mean: $\sigma^2(v) = \hat{\sigma}^2 = \frac{1}{n-1}\sum_{v \in V_I}\sum_{j \in J(v)}(x_j - \rho(v))^2$. In practice, we apply a vertex-specific variance estimate $\sigma^2(v)$ to account for heteroscedasticity in the read depth data (see "JaBbA model fitting" section below).

Given this model, the joint log-likelihood of the read depth data x across the graph given copy number assignment κ is

$$logP(x|\kappa) = -\sum_{v \in V_I} \mathcal{V}(v, \kappa, x, J) + Const(\kappa) \tag{2.7}$$

We also refer to $V(G, \kappa, x, J) = \sum_{v \in V_I} V(v, \kappa(v), x, J)$ as the *read depth residual* of the JBGG (G, κ) relative to data x.

The satisfaction of junction balance and skew-symmetry constraints in Eq. 2.1-2.3 may place nonzero copy number at one or more loose end edges. Each loose end in the input graph represents a slack variable that allows the junction balance constraint to be relaxed at specific internal vertices, allowing the data to be fit even when junctions are missing from the input (e.g. due to low mappability, sequencing depth, or purity). Only loose ends that are given nonzero CN, are considered to be "used" in the final graph. To penalize solutions that require the use of many loose ends, we add an exponential prior with decay parameter λ on the loose end CN in (G, κ), which makes models with many missing junctions unlikely. This prior has log likelihood

$$logP(\kappa|G, \lambda) = -|V_I|log\lambda - \lambda\mathcal{R}(G, \kappa) \tag{2.8}$$

where

$$\mathcal{R}(G, \kappa) = \sum_{v \in V_I} \mathcal{R}(v, \kappa) = \sum_{v \in V_I} \kappa(E_L^-(v)) + \kappa(E_L^+(v)) \tag{2.9}$$

is a *complexity penalty*. Adding the log likelihood in Eq. 2.7 to the prior in Eq. 2.8 yields a penalized log likelihood for the data with regularization parameter λ. Under this model, the maximum a posteriori probability (MAP) estimate of κ

$$f(G, \kappa, x, J, \lambda) = \mathcal{V}(G, \kappa, x, J) + \lambda \mathcal{R}(G, \kappa) \tag{2.10}$$

which combines the quadratic read depth residual $\mathcal{V}$ and ℓ_1-norm complexity penalty $\mathcal{R}$ into a single quadratic objective. In practice, we apply models that penalize the number of loose ends with nonzero copy number, i.e. applying an ℓ_0-norm penalty $\mathcal{R}(\kappa) = \sum_{v \in V_I} [\kappa(E_L^-(v)) > 0] + [(\kappa(E_L^+(v)) > 0])$. We use f to define a MIQP, which we solve to infer a MAP estimate for κ given data x and genome graph G:

$$\begin{aligned} \underset{\kappa: V_I \cup E \to \mathbb{N}}{\text{minimize}} \quad & f(G, \kappa, x, J, \lambda) \\ \text{subject to} \quad & \kappa(v) = \kappa(\bar{v}), \ \forall_{v \in V_I} \\ & \kappa(e) = \kappa(\bar{e}), \ \forall_{e \in E} \\ & \kappa(v) = \sum_{e \in E^-(v)} \kappa(e) \quad = \sum_{e \in E^+(v)} \kappa(e), \forall_{v \in V_I} \end{aligned} \tag{2.11}$$

The resulting MAP estimate $\hat{\kappa}$ defines the JBGG $(G, \hat{\kappa})$ which is outputted and returned to the user (Figure 2.4).

2.2.2 JaBbA pipeline

Overall, the pipeline can be split into 3 parts, 1) building a primitive genome graph from the primary segmentation of coverage, 2) fitting integer CN for vertices and edges by solving the MIQP problem, and 3) simplify the output graph and report in the form of gGraph of gGnome package (see section 2.3).

Junction Balance Analysis (JaBbA, `https://github.com/mskilab/JaBbA`) is an R package freely available under the MIT license. The required inputs to JaBbA are binned (e.g. 200bp) and normalized read depth data y, purity α, ploidy τ, a set of junctions $\mathcal{A}$ (see mathematical formulation above), and a hyperparameter λ. The key output is a `gGraph` object representing a junction-balanced genome graph (i.e. solution to **Eq. 2.11**) which can be queried and analyzed using downstream algorithms, e.g. SV event classification algorithms in the `gGnome` package (`https://github.com/mskilab/gGnome`, see "Structural variant event classification" section below). The workflow of JaBbA is composed of four phases: read depth preprocessing, graph building, JaBbA model fitting, and postprocessing.

Read coverage preprocessing, primary segmentation, and purity-ploidy estimation

Raw input read depth data are generated for tumor and matched normal (i.e. constitutional) samples by tallying the count of midpoints of well aligned (MAPQ > 0) proper read pairs in a vector $z \in \mathbb{Z}_+^n$ (e.g. $n \approx 15$ million for 200 bp bins across hg19). These are further GC-corrected and mappability normalized through an iterative LOESS fitting procedure similar to [99] and implemented

in the fragCounter R package (https://github.com/mskilab/fragCounter). Briefly, a local estimated scatterplot smoothing (LOESS) function f_{GC} is fitted (R `stats` package `loess` function) to a random (100,000) subsample of bins to predict read depth as a function of GC content. Each bin's read depth $z_j,\ j \in \{1, ...n\}$ is then divided by its fitted value $f_{GC}(z_j)$ to yield the normalized value z_j^{GC}. A similar LOESS-based procedure is then applied to correct each GC-corrected read depth z_j^{GC} with respect to its 100mer mappability score to yield the GC and mappability corrected read depth z_j^{GC}. This procedure is applied separately for tumor and matched normal (if applicable).

The ratio of the GC and mappability corrected read depth profiles for the tumor z^T and normal z^N is then used to derive the normalized read depth profile $y \in \mathbb{R}_+^n$ where $y_j = \frac{z_j^T m_j^N}{z_j^N},\ j \in \{1, ...n\}$ and m_j^N is the chromosomal-wide median in the normal read depth data for the chromosome containing bin j. Correction of the tumor / normal ratio by m_j^N maintains a linear relationship between read depth and locus abundance in the original sample. Without this final step, bins from the X and Y chromosomes in a diploid male tumor will show the same read depth as the autosomes. Samples without a matched normal (e.g. cell lines) are divided by a "universal" (i.e. average) normal averaged across 943 GC and mappability corrected TCGA normal read depth profiles and then chromosome balanced.

We note that the read depth normalizations, described above, are formally performed upstream of the JaBbA pipeline (e.g. via fragCounter, R / Bioconductor). The goal of these normalizations is to reduce biases, with the goal of rendering the binned read depth y_j proportional (± noise) to the abundance of locus j in the tumor sample (which may include admixed normal cells). The

final (purity / ploidy) transformation, included formally as part of the JaBbA pipeline, will renders the normalized read depth approximately equal (± noise) to its clonal integer CN in the tumor (given an accurate purity / ploidy estimate). Given a tumor sample with purity $\alpha \in [0, 1]$, ploidy $\tau \in \mathbb{R}_+$, and normalized read depth data y, the read depth y_j, tumor CN $\kappa(j)$, and constitutional CN $\kappa_N(j)$ will obey:

$$\frac{y_j}{\sum_{\hat{j}=1}^{n} y_{\hat{j}}} \approx \frac{\alpha\kappa(j) + (1-\alpha)\kappa_N(j)}{\sum_{\hat{j}=1}^{n} \alpha\kappa(\hat{j}) + (1-\alpha)\kappa_N(\hat{j})} \tag{2.12}$$

Given $\bar{y} = \frac{1}{n}\sum_{\hat{j}=1}^{n} y_{\hat{j}}$, $\tau \approx \frac{1}{n}\sum_{\hat{j}=1}^{n} \kappa(\hat{j})$, and a near diploid constitutional sample (i.e. in which $\frac{1}{n}\sum_{\hat{j}=1}^{n} \kappa_N(\hat{j}) \approx 2$),

$$y_j \approx \frac{\alpha\kappa(j) + (1-\alpha)\kappa_N(j)}{\alpha\tau + 2(1-\alpha)}\bar{y} \tag{2.13}$$

Defining

$$\begin{aligned} \beta &= \frac{\bar{y}\alpha}{\alpha\tau + 2(1-\alpha)} \\ \gamma &= \frac{2\bar{y}(1-\alpha)}{\alpha\tau + 2(1-\alpha)} \end{aligned} \tag{2.14}$$

yields the following equation:

$$\kappa(j) \approx x_j = \frac{2y_j - \kappa_N(j)\gamma}{2\beta} \tag{2.15}$$

Given known constitutional copy number (in practice, $\kappa_N(j) = 2$ for all bins with the exception of X and Y chromosome in males, where $\kappa_N(j) = 1$), the right hand side of Eq. 2.15 is computed directly from the (GC and mappability corrected, tumor / normal ratio transformed) read depth data, purity, and ploidy and represents the binned read depth input x to JaBbA.

Graph building

To build the genome graph G, the binned read depth x is segmented (e.g. using CBS) to yield the breakends $\mathcal{B}$. These breakends are combined via **Algorithm 1** with the junctions $\mathcal{A}$ to yield G. Briefly, this procedure 1) divides the reference genome into internal vertices $v_I \in V_I$ each of which will be assumed to have a coherent CN, and 2) establishes edges $e \in E = E_R \cup E_A \cup E_L$ that each represents an adjacency consistent with the reference genome (REF edge, E_R), created by an rearrangement junction (ALT edge, E_A), or an unmatched breakend (loose end, E_L). The ends of the vertices are the union of junction breakends $getBreakPoints(\mathcal{A})$ and a primary segmentation of the genome $\mathcal{B}_{seg}$ using the Circular Binary Segmentation algorithm [100] (CBS), or from user input. The mapping $J(v)$ (see above) is constructed to associate internal vertices $v \in V_I(G)$ with partitions of bin indices.

JaBbA model fitting

Given a genome graph G and GC / mappability corrected, normalized, purity / ploidy transformed read depth data $x \in \mathbb{R}^n$, and bin set $J(v) \subseteq \{1, \ldots, n\}$, $v \in V_I(G)$, we infer a junction balanced genome graph (G, κ) using the procedure outlined above ("Inferring junction-balanced genome graphs" section) with the following modifications: Instead of the arithmetic mean $\rho(v)$, we use the sample geometric mean $\rho(v) = exp(\frac{1}{|J(v)|} \sum_{j \in J(v)} ln(x_j))$, which is more robust to outliers. To model heteroscedasticity in read depth, we compute a vertex specific variance function $\sigma^2(v)$. Specifically, we use LOESS to fit a smooth non-linear function f that links copy number to variance by fitting the sample mean $\rho(v)$ and sample variance $\hat{\sigma}^2(v) = \frac{1}{|J(v)|-1} \sum_{x \in J(v)} (x_j - \rho(v))^2$ across all $v \in V_I$. This function is then used to

determine the segment-specific variance term as $\sigma^2(v) = f(\rho(v))$. These $\sigma^2(v)$ values are then used to populate the read depth residual sum of squares $\mathcal{V}(v, \kappa, x, J)$ portion of the objective function in **Eq. 2.6**. Combining with the user-defined loose end penalty hyperparameter λ (default 100, tuned for 200 bp binned tumor / normal pairs, see above), the MIQP problem (**Eq. 2.11**) is solved on the genome graph using an ℓ_0 complexity penalty. The solver is `CPLEX` (v12.6.2, IBM).

To increase sensitivity for missed junctions, the JaBbA pipeline employs several iterations of MIQP inference when two tiers of junctions (high and low confidence) are provided. Briefly, we fit a model using only high-confidence (e.g. FILTER=PASS in the unfiltered `SvABA` output VCF) junctions as input. In subsequent iterations, we fit a modified model that includes additional low-confidence junctions (e.g. those with low read support) that occur near (<1 kbp) a loose end in the previous iteration. This procedure is run for a maximum of four iterations, or until no additional low confidence junctions near loose ends from the previous iteration have been found. Such a procedure particularly improves junction sensitivity in the setting of low purity and / or read depth.

Post processing and allelic CN fitting

After the optimization with `CPLEX`, we further simplify the solution by merging neighboring vertices of the same total CN that are connected only through REF edges. If ALT and REF read counts at constitutional heterozygous SNP sites are provided as (optional) input to JaBbA, we also infer the tumor allelic CN $\kappa_h(v)$ and $\kappa_l(v)$ corresponding to the "high" (h) and "low" (l) allele, where $\kappa_h(v) + \kappa_l(v) = \kappa(v)$ and $\kappa_l(v) \leq \kappa_h(v)$, for each internal vertex $v \in V_I$ that overlaps a heterozygous

SNP. (Only biallelic heterozygous SNPs with two constitutional states, i.e. ALT and REF are considered.)

Briefly, we model the "low" count at each heterozygous SNP ("low" can be REF or ALT, depending on which allele has fewer reads) overlapping v as a Poisson distribution with mean $\beta\kappa_l(v) + \frac{1}{2}\kappa_{lN}(v)\gamma$. Here, $\kappa_{lN}(v)$ is the constitutional allelic CN of the lower CN allele at each vertex v (e.g. 0 for X and Y chromosome in males, 1 elsewhere in the genome), and β and γ are defined as for (non-allelic) read depth (**Eq. 2.14**) with the following adjustments: genome-wide average $\bar{y}$ is the mean allele-specific read count across all heterozygous SNPs, and τ is the mean non-allelic CN across heterozygous sites only. An analogous Poisson model is defined for the "high" allele (i.e. REF or ALT, depending which has more reads) at each heterozygous SNP associated with v. A joint Poisson log likelihood for the allelic CN of a given vertex v, $\kappa_h(v)$ and $\kappa_l(v)$, is then computed by summing all log likelihoods across both high and low alleles of all heterozygous SNPs mapped to v. The maximum likelihood configuration for $\kappa_h(v)$ and $\kappa_l(v)$ is chosen and reported. The process is repeated across all internal vertices $v \in V_I(G)$ with at least one heterozygous SNP.

The final output is then saved into a gGraph object that can be manipulated, visualized, and analyzed with the `gGnome` R package (`https://github.com/mskilab/gGnome`) and `gGnome.js` browser (`https://github.com/mskilab/gGnome.js`).

2.2.3 JaBbA robustly produce accurate copy numbers for DNA segments and junctions

Having described JaBbA algorithm and pipeline, we next set out to comprehensively evaluate its performance against other tools of similar purposes.

Table 2.1: **Comparison of JBGG reconstruction methods.**

Except for AmpliconArchitect and CouGaR, which only reconstruct complex amplicons, most methods build genome-wide JBGGs. MIQP, mixed-integer quadratic programming; ILP, integer linear programming; LBP, loopy belief propagation; MILP, mixed-integer linear programming.

Method	Scope	Solver	Publication
JaBbA	genome	MIQP	[83]
PREGO	genome	ILP	[78]
ReMixT	genome	Variational Inference	[92]
Weaver	genome	LBP	[101]
RCK	genome	MILP	[102]
InfoGenomeR	genome	ILP	[82]
AmpliconArchitect	amplicon	ILP	[80]
CouGaR	amplicon	ILP	[79]

Comparison to previous genome graph reconstruction methods

To my best knowledge, there have been 8 published methods to reconstruct JBGGs from WGS data (Table 2.1). AmpliconArchitect and CouGaR work only within amplified regions of the genome, while the other methods all infer JBGG genome-wide. Most adopt a undirected variation of the interval graph structure, while JaBbA is built upon a skew-symmetric directed graph. Weaver and ReMixT solve the problem with probabilistic graphical models, and the others utilize different variations of integer programming.

JaBbA differs primarily from previous genome graph methods in its robust

modeling of WGS read depth noise and missing (i.e. false negative) junctions within a MIQP framework. First, while previous approaches (PREGO [78], CouGaR [94], ReMixT [92], and Weaver [93]) directly model read counts as a Poisson (or gamma-Poisson) random variable, JaBbA gleans non-linear mean-variance relationships empirically from high-resolution fixed bin (200 bp) read depth data. This allows JaBbA to utilize transformed and normalized read depth data (e.g. via GC and mappability correction, tumor / normal ratio, purity and ploidy transformation). Such data violates Poisson assumptions (e.g. integer variable, linear or constant relationship between mean and variance) but more faithfully reflects CN in stromally admixed and aneuploid tumor genomes.

Second, JaBbA is robust to missing junctions by allowing (but penalizing) the use of *loose ends* in its reconstructions. Loose ends arise when a CN change is unaccompanied by a nearby junction. Such an event is a common occurrence in WGS data (e.g. due to low coverage, algorithmic filters, rearrangements in repetitive regions, etc.). Methods that do not explicitly model loose ends (e.g. PREGO, Weaver) suffer from either under- (PREGO) or over-segmentation (Weaver). Finally, by posing its inference as a MIP, JaBbA avoids analytic limitations inherent to PGM inference (ReMixT, Weaver), including the inability to identify global (expectation maximization, ReMixT) or local (loopy belief propagation, Weaver) optima or model very high-level CN states (ReMixT). In contrast, JaBbA's MIP-based framework allows for the discovery of globally optimal model fits while allowing for an unrestricted range of CN states.

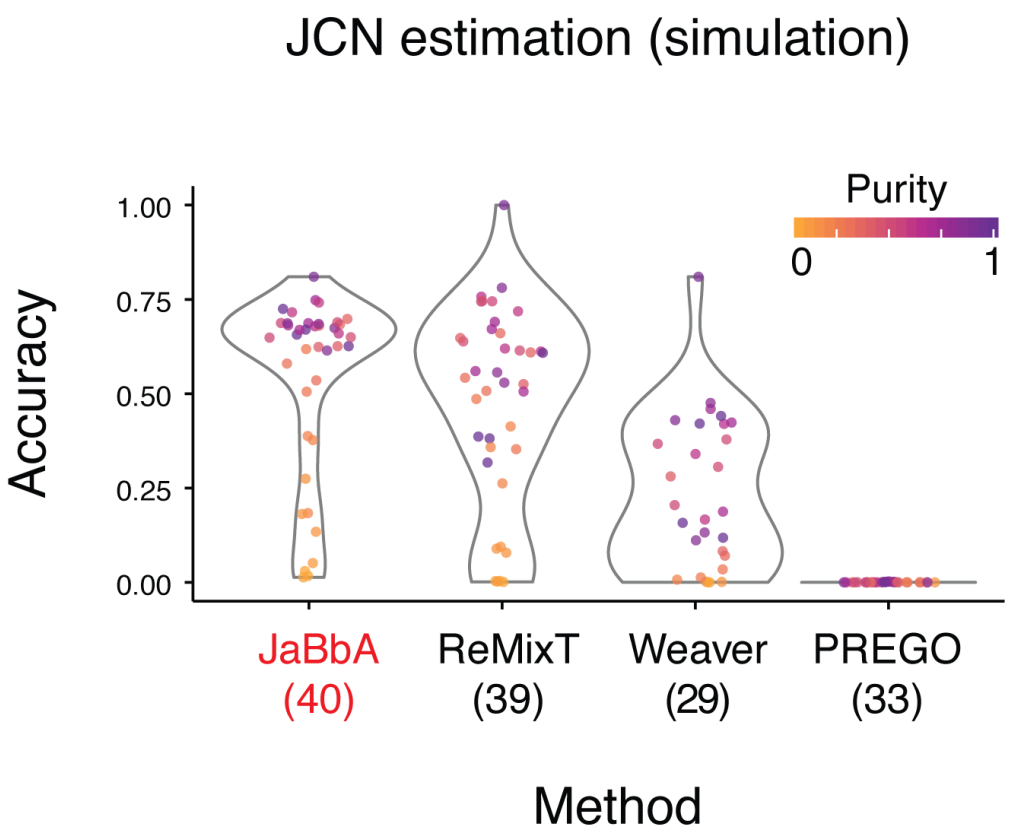

Figure 2.5: **Junction copy number accuracy.**

Benchmarking across 40 simulated genomes and HCC1954, with four technical replicates at ten different sample purity levels by admixing with matched normal genome. The numbers in parentheses after the names of the methods indicate the numbers of successful runs out of 40. F1 score is the harmonic mean of precision and recall. Same for Figure 2.6, 2.7, 2.8, 2.9. This figure shows the accuracy of JCN estimation for simulated genomes only (fraction of junctions with correctly estimated JCN).

Benchmarking JaBbA

In the simplest terms, reconstructing genome graphs consists of two tasks: estimating junction copy numbers and DNA segment copy numbers. A junction

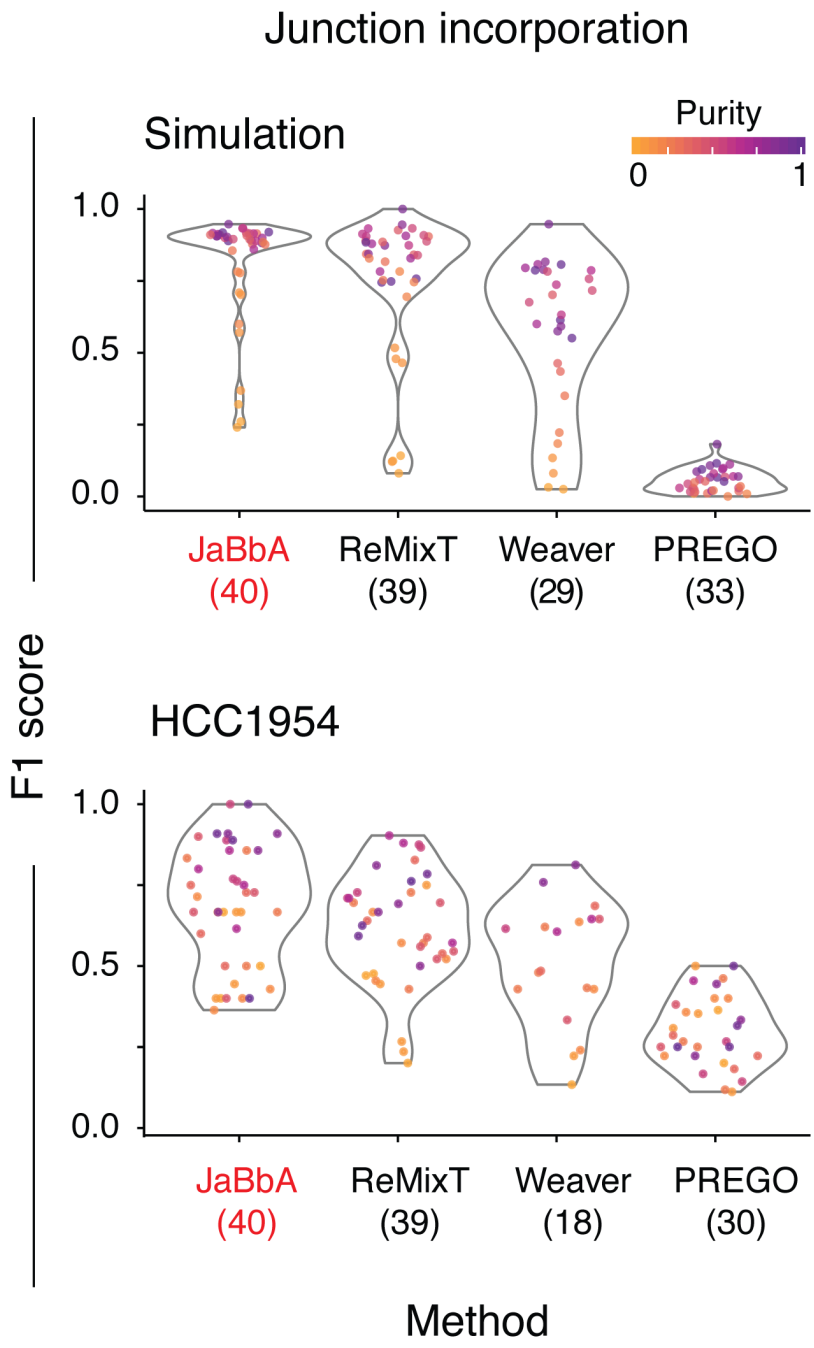

Figure 2.6: **F1 score of junction incorporation.**

is said to be incorporated in the genome graph if it is assigned non-zero copy number. Thus, a genome graph reconstruction method's performance can be evaluated from these two main aspects, 1) incorporating correct junctions and

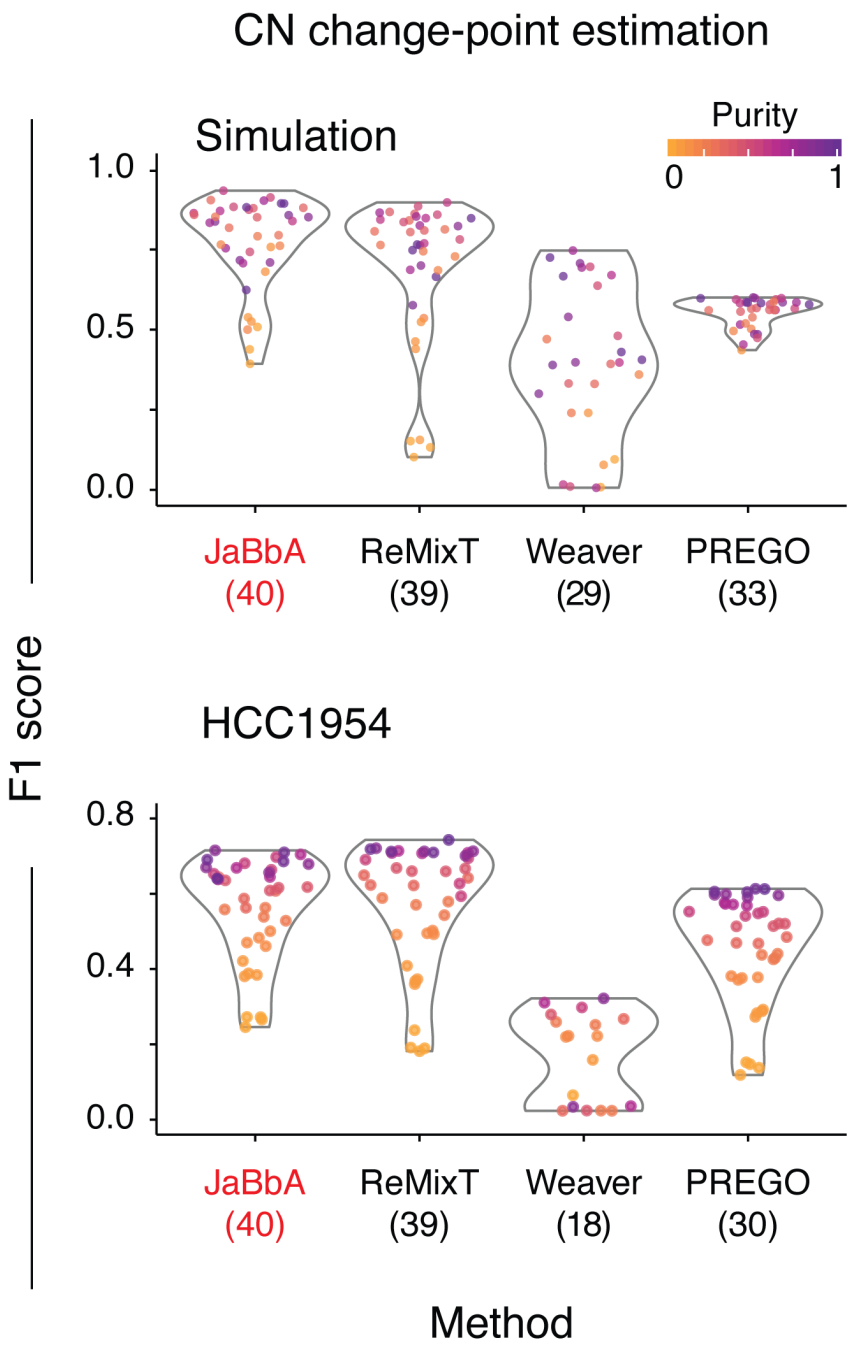

Figure 2.7: **F1 score of copy number change point placement.**

estimating the correct JCNs, and 2) faithfully segmenting the genome and estimating their CNs. We compared JaBbA's performance in both aspects against three other genome graph reconstruction methods (PREGO [78], ReMixT [92],

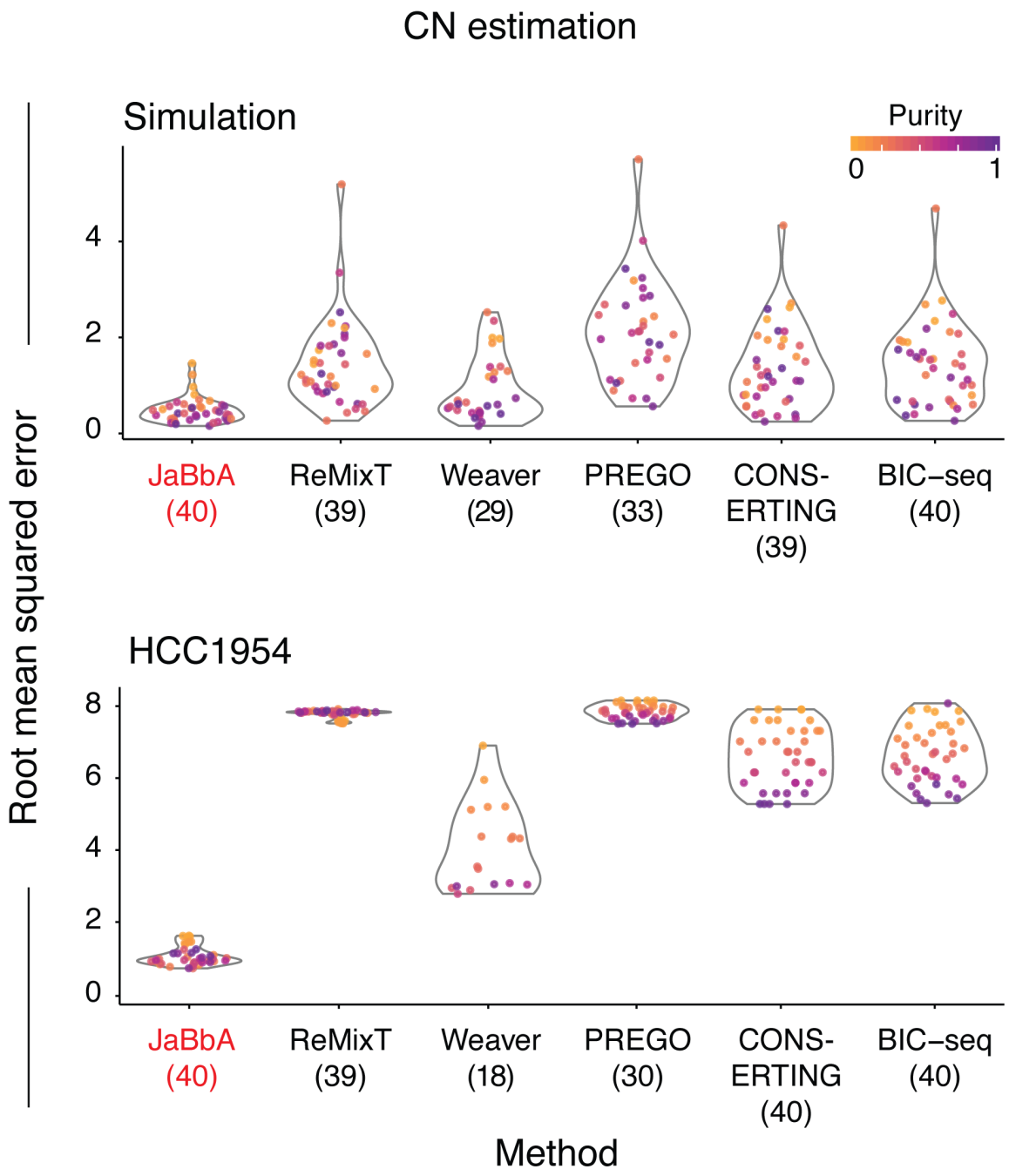

Figure 2.8: **Root mean square error of estimated CN from gold standard.**

Weaver [93]), and specifically in segmental CN estimation with two extra somatic CNA callers that do not infer genome graphs (BIC-seq [52], CONSERT-

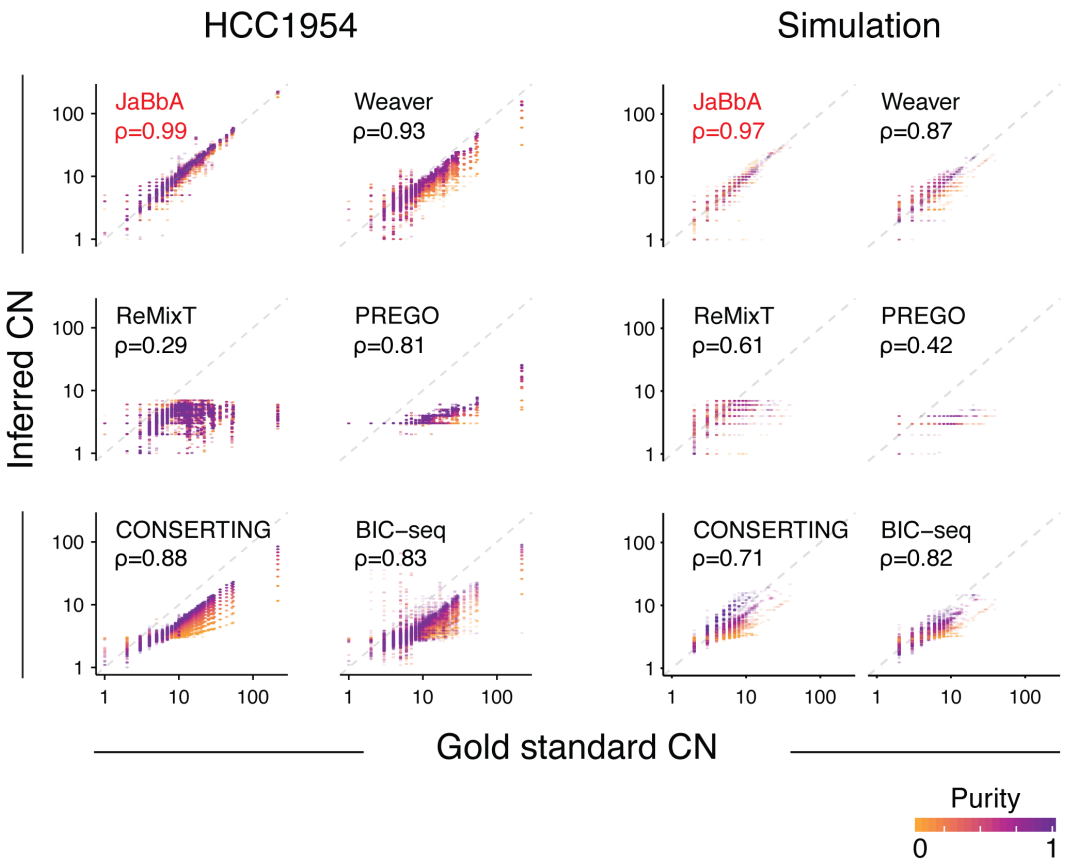

Figure 2.9: **Estimated versus gold standard CN in all successful runs.**

Pearson correlation coefficients ρ using all points are labeled.

ING [59]). Another genome graph reconstruction method, CouGaR, is limited to amplified regions, so we only applied it to one replicate of the HCC1954 to demonstrate its difference from the other methods at a putative BFB cycles locus (Figure 2.10).

The input junctions to all genome graph reconstruction methods were their respective default recommendations, namely SvABA unfiltered candidates for JaBbA; internal junction callers for Weaver, CouGaR; and the union of 3 junction calls (SvABA filtered [46], Delly [48], Novobreak [103]) for PREGO, ReMixT, and CONSERTING as they make no default recommendations. As a note, we

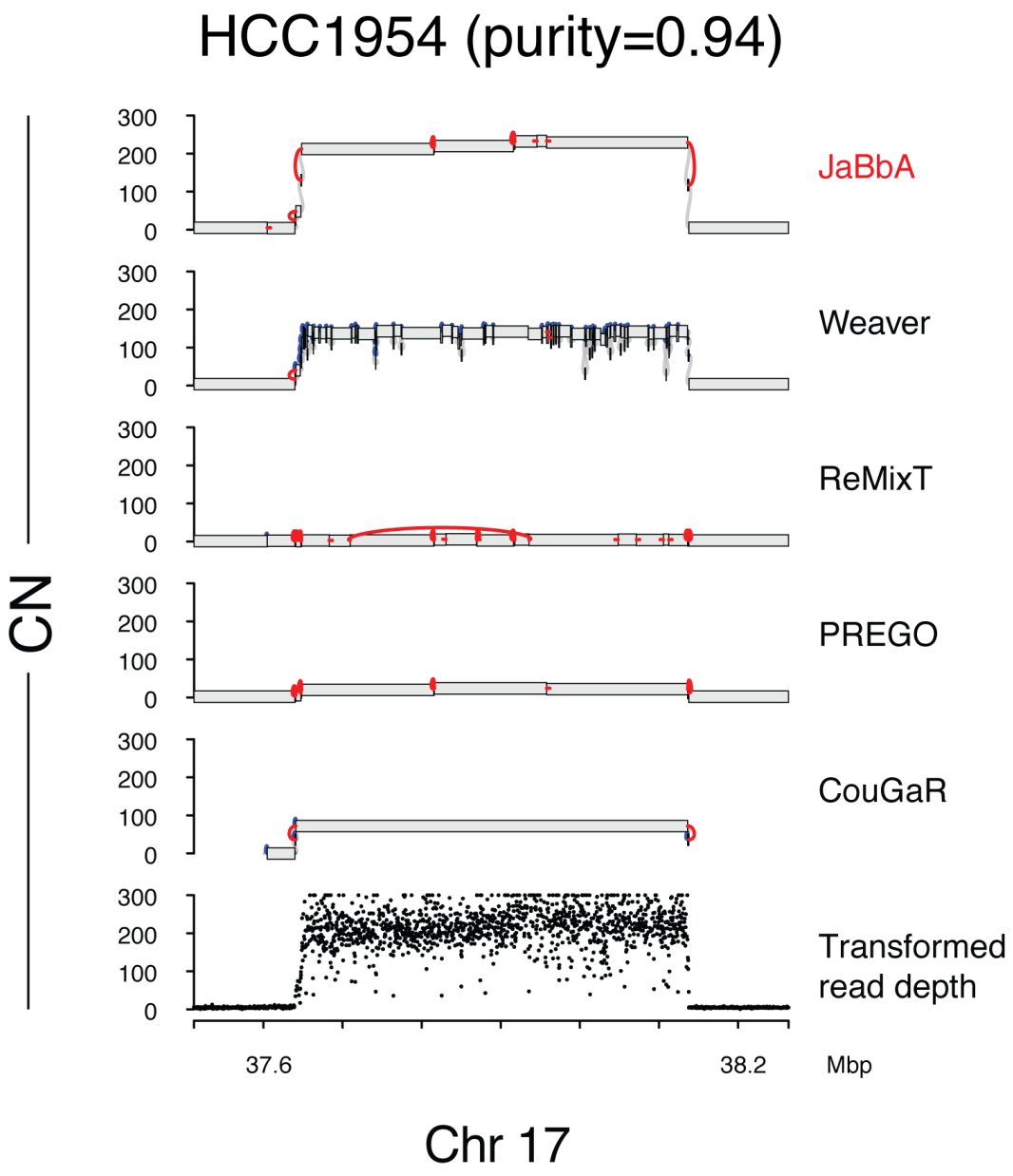

Figure 2.10: **Reconstructed genome graphs around *HER2* in HCC1954.**

Example of a putative BFBC event in HCC1954 and the reconstruction by five different methods (top five) and the binned read coverage transformed in CN space (bottom).

saw very similar performance with JaBbA using standard settings (SvABA unfiltered junctions) and the union junction set (data not shown).

Junction incorporation and JCN estimation

To evaluate junction incorporation and JCN estimation, the incorporated junctions were approximately matched to gold standard junctions, e.g. both breakends were within 1 kbp from the corresponding gold standard junctions with the same orientations. True positives were the number of matches, false positives were incorporated junctions without a match, and false negatives were gold standard junctions without a match. An F1 score of the incorporated junctions, the harmonic mean of precision and recall, was then computed for each benchmarking sample (Figure 2.6). Gold standard junctions for HCC1954 were defined as the consensus set inferred from the original high depth data (junctions identified by at least two callers out of SvABA, Delly, Novobreak).

Inferred JCNs of the graph reconstruction were compared to the matching gold standard junctions, if any, from the simulated dataset. The proportion of correctly fitted JCNs out of all incorporated junctions times recall of incorporation represented the accuracy or completeness of JCN estimation (Figure 2.5).

Segmental CN estimation

When evaluating CNA inference performance, two metrics were considered. One was the correct placement of CN change points. Analogous to matching junction breakends, we considered an inferred CN change-point to match a gold standard if their distance was within 1 kbp and they have the same direction of CN change, e.g. increasing or decreasing CN from the side with smaller coordinates to the larger. To prevent spurious matching between inferred and gold standard CN change points, occurring in cases with excessive hyperseg-

mentation, gold standard CN change points were only matched to the nearest identically oriented inferred CN change point. Based on the matching, F1 scores were computed and shown in Figure 2.7. For HCC1954, the gold standard CN change-points were defined as the consensus junction breakends.

The other set of metrics reflects the concordance between the inferred CN profile with the gold standard. For each segment in the gold standard, the overlapping inferred segments were identified. The inferred CN for that gold standard segment was then defined as the overlap width-weighted average of inferred CN. Subsequently, the root mean squared error (RMSE) $\sqrt{\frac{\sum(\hat{x}-x_{goldstandard})^2}{n}}$ was computed across all gold standard segments for each sample (Figure 2.8). The gold standard for HCC1954 CN profile was derived from published microarray-based segmentation downloaded from the CCLE data portal. High-depth WGS coverage data was mapped onto these segments and transformed into copy number space using the known ploidy of 4.5 and purity of 0.99. To compactly show the relationship between inferred CN and gold standard from all runs, we drew scatter plots across all gold standard segments in all successful runs for each method, and the Pearson ρ of each method was calculated from all data points pooled together (Figure 2.9).

2.3 Implementation of genome graphs in gGnome package

To facilitate the analyses of genome graphs, we built an R package *gGnome* which allows fast and useful creation, exploration, visualization, and computation around genome graphs. The basic design principle is to keep the integrity of genome graphs, in particular the skew-symmetric property, on the back end

with private fields and methods in R6 classes, while expose the most intuitive inputs and outputs like junctions and segments to the users.

We constructed two main classes, `gGraph` and `gWalk`, corresponding to genome graph and walks. In the most basic form, a genome graph can be initiated with a reference genome alone, where each node is a chromosome and no edges exist (`gG()`). On top of that, we can segment the whole genome based on a set of breakpoint coordinates, by automatically connecting two consecutive nodes with a (pair of reverse complement) REF edge (`gG(breaks = breakpoints)`). When a ALT junction emerges, it joins two breakends that are not adjacent in reference genome to form new (pair of reverse complement) ALT edges (`gG(juncs = junctions)`). One can create any genome graph from the input data of segmentation (in a format equivalent to `GRanges`), and/or a set of junctions (in a format equivalent to `GRangesList`). More often in practical scenarios, we start from a copy number-annotated genome graph inferred from WGS. gGnome allows one to parse the results from most of the existing genome graph reconstruction methods (JaBbA, ReMixT, Weaver, PREGO, RCK, CouGaR, AmpliconArchitect).

With a genome graph (supposedly stored in variable `gg`), we can make a series of queries based on vertex and edge metadata, with the graph subsetting operator "[,]", where the expression before the comma filters the vertices and the latter filters the edges. For example, when looking up amplified subgraphs where the vertex CN is larger than twice the ploidy (supposedly stored in variable `pl`), one can execute `gg[cn>pl*2,]`.

To quickly explore any part of a genome graph, we integrated our static genome browser-style plotting package `gTrack`. With any genome graph, one

can simply call `gg$gtrack` to get the `gTrack` object to then plot using `plot` function. All the optional arguments of `gTrack` can be passed down natively from this interface. For example, if one wants to plot copy number stored in the metadata column "cn" on the Y-axis, it is as simple as `gg$gtrack(y.field = "cn")`. As a more powerful alternative, we also provide interactive visualization through *gGnome.js*, described in the next section. Here, with any `gGraph` object, one can simply generate the required JSON format with `gg$json()`.

Given a genome graph, can also calculate a lower bound shortest derivative (in the rearranged genome) genomic distance between any two genomic loci after rearrangements, by summing up the width of vertices the path traversed (encoded in the gGraph method `gg$dist`). Taking advantage of the optimal implementation of Floyd-Warshall by the *igraph* R interface, this can be immediately scaled up to find all shortest paths from one set of genomic loci another set, encoded in the function `proximity`.

One of the more interesting topics in SV pattern discovery is the clustered junctions. We can easily cluster vertices and/or edges based on their connectivity, implemented in `gg$ecluster` for edges and `gg$cluster` for vertices. The output cluster membership (weakly connected or strongly connected), will be recorded in a metadata column and can be used to retrieve the corresponding subgraph, for instance, `gg$ecluster(); gg[,ecluster==1]`.

Besides, with a non-negative numeric field associated with vertices and edges, for example, copy numbers, we can also regard that as *capacity* in traditional flow problems. For example, the *max flow* problem with CN means the "maximum number of copies of such DNA molecules that connects genomic locus A and B".

Next, we provide functions to replicate our algorithms in [83] and this dissertation for finding the subgraphs of 13 different SV events. The `events(gg)` function will automatically search the given graph for these patterns with the default parameters and annotate any findings in the metadata columns with the name of the event, like "chromothripsis". Plus, after running the function, there will be a new data table of all the findings appended to the field of `gg$meta$events`.

2.4 Interactive visualization of genome graphs in arbitrary genomic windows

To visualize the genome graphs we developed `gGnome.js` (source code at `https://github.com/mskilab/gGnome.js`), a Javascript-based genome browser that renders genome graphs and allows for interactive browsing across discontiguous genomic intervals accompanied by various genomic annotations. Here, we describe several important and unique features of gGnome.js.

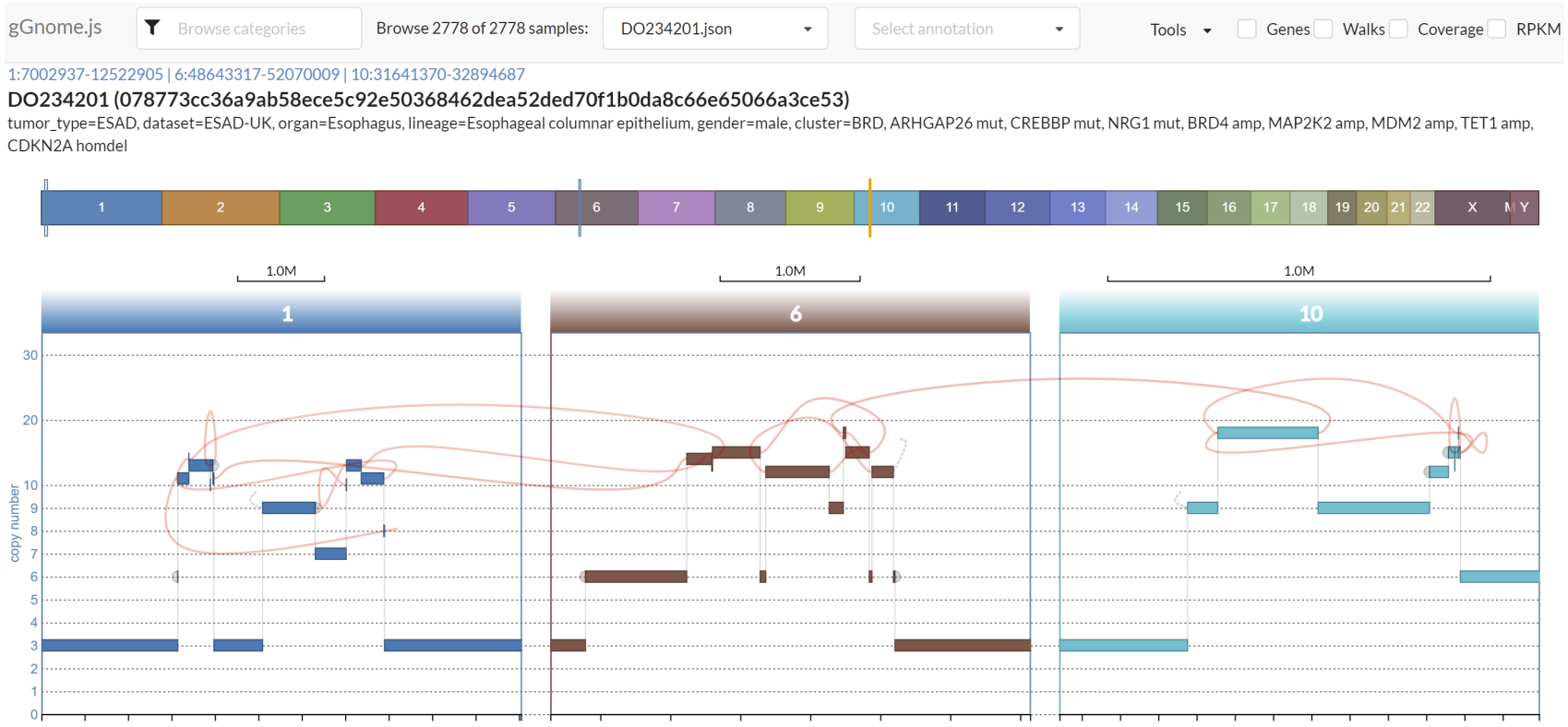

Figure 2.11: **A snapshot of the gGnome.js interface of a BFB cycles event in an esophageal adenocarcinoma sample.**

First and foremost, the flexibility of arbitrary window browsing is crucial for complex rearrangement events, which can vary in scale and involve many distal loci. Annotations of events (or any set of vertices/edges) were read by the browser and can be searched. For example, in Figure 2.11 we can visualize a multi-chromosomal breakage-fusion-bridge cycle (BFB cycles) event in 6 discontiguous windows spanning 4 different chromosomes and locate this event instantly with the `annotation` selector on the top right corner. In Figure 2.11, from the bottom to the top: X axis is the reference genome. Y axis is mapped to JaBbA-inferred copy number. Colored bars are vertices and links edges. Viewing windows can be created by selecting regions on the whole genome bar above the plotting region. Sample information is displayed along with the currently viewing windows. Toolbar at the very top allows for quick search of samples, focus on an annotated region, locate a specific range by coordinates or gene name, as well as controlling visible tracks in the browser.

Annotations corresponding to segment-level and also junction-level information were included in the browser. Metadata for segments includes, but is not limited to, total copy number, allele-specific copy number (for samples and intervals in which heterozygous SNP information is available, i.e. cases with both tumor and normal pairs), and correspondence to event type if the segment was identified as part of an event. Edge metadata includes information output by the SV caller used in this study, SvABA, including number of reads supporting the SV and junction topology. These metadata can all be presented when the cursor hover over a specific element.

Second, the `gGnome.js` browser enables visualizing any genome graph, either with CN as inferred by other graph callers, or even without CN in the case of plain genome graphs.

As a meaningful default, we plot integer copy numbers of the segments on the Y axis, that are inferred from cancer whole genome sequencing (WGS) with JaBbA (described in the next section), making it ideal for quick browsing copy number aberrations and immediately reveal the architecture of the complex events.

Third, when visualizing a larger cohort an integrated tag filter helps narrow down desired samples by any metadata string, exemplified by primary cancer types, data source projects, mutation status of genes.

Fourth, gGnome.js also supports visualizing additional genome tracks in the format of scatter plot, bar plot, and gene annotations, whose visibility can be controlled by the functionnalities in the top right corner.

2.5 Discussion

In this chapter, I defined a skew-symmetric direct graph representation of a rearranged genome, termed genome graph. Genome graphs are general enough to encapsulate any junction that has corresponding coordinates within the reference genome, and thus can represent SV events of arbitrary complexity.

Compared to variation graphs [72] such formulation is almost equally general, except for foreign sequences and variants that cannot be associated with a coordinate (e.g. small non-templated insertions), yet exceedingly light-weight

as the vertices are stored as coordinates instead of the actual sequences. In fact, for SV events involving foreign sequences, for instance, viral genome integration, simply concatenating new sequences to the reference genome will allow them to have a coordinate to be represented. Although here I only described the usage of genome graphs in representing each cancer genomes with SVs, it is straightforward to extend to germline SVs. [73, 104]. Following this thought it is apparent that as long as there is a reference genome, the SVs arising in that genome can be represented using gGnome, immediately applicable to other organisms. In fact, we have built genome graphs for the SVs detected in SARS-Cov-2 viruses sampled from COVID-19 patients sequenced at Weill Cornell Medicine with the original strain as a reference [105].

More importantly, this formulation does not assume the represented genome must be one genome or that the vertices coordinates are disjoint. The user can append extra metadata to expand the graph in creative ways. There are at least two important applications of this flexibility.

First, we can use the framework to represent fully or partially phased genome graphs. In fully phased genomes (to either parental alleles, not somatic phase), we would have two pairs of vertices representing the same DNA segment for each parental phase assginment. Same goes with edges, so that the edges of one allele will never be crossing over to the vertices of the other allele, following the reasonable assumption that there is no somatic recombination. In fact, Weaver, ReMixT, RCK, and InfoGenomeR have attempted to infer partially phased genome graphs from bulk WGS. gGnome should be compatible to represent these results and such feature will be released in the near future. The clear advantage of inferring germline phase of junctions is that we can tease

apart SV events that are happening on different alleles yet overlapping on the reference coordinates. For example, in pyrgo instances, the final derivative sequences can be very different depending on whether some DUP-like junctions are in *cis* versus *trans* [83].

Second, gGnome allows the joint analysis of multiple genome graphs. Such situation can arise in multi-sample study designs such as diverging cell clones (shown in detail in Chapter 4), spatio-temporal sampling of tissue [106], primary and metastatic sampling, or even germline genome sampled from populations like 1000 Genome Project [107]. Simply annotate the source identity in a metadata field, concatenate (extend) the each graph, and disjoin the graph to reduce to a pan-genome graph. This allows for coupling of junctions, loose ends, or vertices across multiple samples and jointly infer their structures, which is very powerful representing related but changing SV events.

Beyond genome graphs, we also implemented walks over a genome graph that directly map to linear DNA sequences that a graph can encapsulate. The conversion between reducing walks to graphs, and generating walks from graphs have very profound impact on the analysis of SVs. For the former, any array of genomic regions, including arbitrarily constructed *de novo*, simulated from certain SV evolution scenario, sequencing read mapped to multiple loci, can be natively regarded as a walk, then trivially converted to a graph based on the junctions implied between consecutive elements in that array. Conversely, we can decompose a JBGG following the constraints in Eq. 2.4 to infer an optimal karyotype that constitute such a genome graph by minimizing some form of loss function like Eq. 2.5.

Using this template, genome graphs and walks are ideal data structures to

bridge the fragmented yet high-throughput short-read sequencing, and the long-range, long-read sequencing or optical mapping techniques. Junction-balanced genome graphs (JBGGs) are stoichiometrically plausible compressions of all possible derivative sequences that can result from the junctions they contain, and can be used to generate a pool of walks which provide substrates for finding an optimal linear combination of alleles like that maximizes parsimony or likelihood of generating an observed long-range pattern [83]. In [83], we gave a form of the algorithm 2.4 where the objective is to find the combination of walks that maximizes the amount of support from long-molecule profiling. In the paper we used 10X Genomics linked-read sequencing and Bionano Optical Mapping technologies, but we note that this approach can be extend to other platforms and even combine orthogonal evidence.

Even without the exact inference of linear alleles, genome graphs can produce useful bounds for many biological applications. Because when the genome graph is complete, the true walk must be contained within, and then we can approximate and bound the various quantities between two reference coordinates in the face of SVs. For example, the shortest possible distances bewteen two reference regions in a graph is the sum of traversed nucleotides along their shortest paths. In PCAWG thyroid carcinoma [108], I recaptured the 5 THCA samples with the THADA-IGF2BP3 enhancer hijacking event reported by PCAWG [109], and further discovered a case where the juxtaposition of THADA near IGF2BP3 is achieved by two junctions from a more complex event, which also elevated the expression of IGF2BP3 like in the cases found in. In another example, two gene with distance zero in the genome graph is putatively fused.

The last major obstacle before realizing the great potential to analyze complex

SVs by genome graphs is the reliable inference of JBGGs from real data. Thanks to methods like JaBbA [83], InfoGenomeR [82], the problem of inferring a JBGG from a short-read whole genome sequencing is a generically tractable and robust to most sample purities. JaBbA differs primarily from previous genome graph methods in its robust modeling of WGS read depth noise and explicit accounting of false negative junctions, also called loose ends (see Section 2.2.1). In our benchmarking experiments, JaBbA inferred JCN with consistently higher fidelity than published genome graph-based methods (ReMixT [92], Weaver [93], PREGO [78]) across a wide range of tumor purities (Figure 2.5).

Of particular interest, in the HCC1954 experiment, JaBbA is the only caller that has a wide enough dynamic range to accurately capture the extremely high CN of the *ERBB2* complex amplicon (Figures 3.2, 2.9) while correctly placing the junctions. Because of that, we successfully identified the underlying BFB event the led to the pathogenic amplification. This is one of the benefits of using a linear programming formulation over probabilistic graphical models.

In addition, JaBbA qualitatively outperformed genome graph-based methods and even classic (i.e. non-graph based) CNA callers (BIC-seq [52] and CONSERTING [59]) in estimating interval CN and change point locations across a wide range of tumor purities (Figure 2.7) as it utilizes the extra precision from the junction topologies. These results show that JaBbA is the first genome graph SV caller to accurately infer the topology of JCN while approaching (or exceeding) the fidelity of classic CNA callers.

We argue that one of the major reasons behind JaBbA's superior performance is the use of loose end. Loose ends allow for flexibility in fitting CN closer to the center of the coverage data even when the corresponding junction is not de-

tected at the breakpoint. Loose end penalty λarise from an exponential prior and is a single hyperparameter that changes JaBbA's fundamental behavior. When it approaches zero, loose ends are free and the output will end up hyper-segmented with wrong CN fitted due to coverage noise. When it approaches infinite, loose ends overrides the attraction of CN to the center of the coverage data and the output will end up in a rigid single segment. Properly set loose end penalty will only allow the algorithm place one when the coverage change is compelling and might reflect real CN change point without an identified junction. These *bona fide* loose ends are abundant in our pan-cancer cohort, and they represent the detection limits of junctions with short-read data. Behr et al. [44] made a deep dive into the underlying reasons of these loose ends and found both the technical limitation of junction detection introduced by non-unique read mapping, and real biological events like viral integration and neo-telomeres.

As in its present form, JaBbA have three major limitations. One, the main product out of the pipeline is total CN, and the allelic CN is done conditioned upon inferred total CN and independently for each vertex based on maximum likelihood of the observed the allelic read counts at heterozygosity sites. In fact, assuming infinite site, at each breakend, only one of the two parental alleles can be incident to an ALT junction and it is a feasible computational problem to put two alleles in *cis* or *trans* when there is allelic imbalance, thus producing partially phased genome graphs even from short-read data. This is also attempted by Weaver, ReMixT, RCK, and InfoGenomeR, but from the former two methods we benchmarked, it often comes at the cost of inaccuracy in total CN estimation. This point is independently implicated in the InfoGenomeR publication [82], where JaBbA is the second most accurate estimator behind InfoGenomeR

in all tested cases. Plus, as shown in the next chapter, even with total CN alone, we can already rigorously discover novel classes of complex SV events from pan-cancer cohorts.

Two, over complicated graphs (unusually high number of segments and junctions) rapidly increase the difficulty to reach the exact global optima. We currently mitigate this by decoupling the whole genome graph into partitions, where the vertices from different partitions are disconnected or only connected through a set of vertices whose CN is fixed, due to very small variance in their maximum likelihood estimated CN (or plainly, long segments with enough data to have a nearly definite CN). However, there is not a particular reason why this quadratic loss (L2 norm) should not be replaced with absolute values (L1 norm) and be solved as a MILP with much higher efficiency. We leave that for the next phase of development.

Three, the graph inferred by JaBbA is a consensus representation over the whole sample, which does not directly model intra-tumoral heterogeneity. The inferred integer copy numbers can be interpreted as "average copies per cancer cell" and when there is heterogeneity, which is found to be prevalent [110], the true values of this quantity could deviate from integers. The inference of subclonal integer CNAs with clonal fractions directly from bulk short-read WGS remains a challenging problem [92] while maintaining faithful CN estimation. However, as we will discuss again in the next chapter, most of the SV patterns we can observe from bulk JBGGs are ancestral or from major clone. Besides, with isogenic cloning designs [111] or the emerging single-cell WGS technologies [112, 113, 114], JaBbA stays a reliable and promising tool to reconstruct high-fidelity JBGGs from clonal populations.

We also implemented the genome graph data structure in an open source R package gGnome, and to our knowledge, it is the first in class API to genome graphs that can 1) allow creation, combination, arithmetic of genome graphs, 2) support various algorithms including but not limited to decomposition to linear alleles, finding shortest paths, cluster vertices and edges, 3) parse the results of all the JBGG inference methods into a centralized object, 4) visualize genome graphs statically and interactively.

The main goal of gGnome is to provide the basic operations needed to build practical applications and analysis. It hides the skew-symmetry implementation from the user using private fields and methods, and present the more straightforward form similar to that of [77], so that the construction, dissection, combination of graph objects are more intuitive and easily adaptable to various common inputs in bioinformatics, including segmentation in SEG, BED formats and junctions in BEDPE, BND VCF (v4.2) formats. The users can immediately combine the low-level functions to achieve biological meaningful procedures. For example, in the next chapter, our graph-based SV event taxonomy typically start with filtering the nodes or edges by a field (like CN above certain threshold), followed by partitioning the whole genome graph into candidate subgraphs, and then extract features from the subgraphs to make the final call, which is extra annotation in the vertex/edge metadata fields.

gGnome.js is also the first genome browser that allows viewing arbitrary windows along the reference, which is a critical feature to browse large-scale complex SV events. It has been instrumental in discovering the complex event taxonomy in the next chapter, as we were able to go through thousands of genome graphs rapidly and notice recurring patterns thanks to the ultrafast modern

web-based graphics libraries (WebGL). It also greatly benefits the whole genome analysis in oncology where we can pinpoint the potential SV drivers almost instantly, either by dynamically changing the zoom level to see the CNA-affected regions, or directly browser through annotated events, or even search the graph by gene of interest. We realize the versatility of not only representing one graph per tumor, and is building the next version to visualize pan-genome graphs and several related genome graphs together, named Pan Genome Viewer (PGV, http://github.com/mskilab/pgv).

We believe this suite of packages including gGnome/gGnome.js/JaBbA should greatly help the research community routinely incorporate graph-based analysis of SVs into the analysis of whole genome sequencing.

CHAPTER 3

DISTINCT CLASSES OF COMPLEX AMPLICONS UNCOVERED ACROSS THOUSANDS OF CANCER GENOME GRAPHS *

In this chapter I will describe the classification of three distinct types of complex amplification events from 2778 genome graphs built with JaBbA from a pan-cancer cohort of over 30 types of cancers. Furthermore, I will show that the patients in this cohort can be stratified based on a total of 13 simple to complex genome graph-based SV patterns to clusters of differential overall survival, tumor type enrichment, and association with specific germline or somatic alterations.

This work is part of our publication [83]*, led by Kevin Hadi, Julie Behr, Marcin Imielinski, and myself.

Individual contributions: *Kevin Hadi and I collected and processed the WGS and associated biological or clinical data for the pan-cancer cohort. I executed and controlled the quality of the output genome graphs. I contributed to the classification of amplicons and carried out the analysis of fusion genes. I contributed to the algorithms identifying the 13 SV event types. I carried out the patient clustering analysis and with Kevin Hadi discovered the associations with the tumor types, genotypes, and overall survival.*

*Hadi, K., Yao, X., Behr, J.M., et al. Distinct classes of complex structural variation uncovered across thousands of cancer genome graphs. Cell, 183(1):197–210, 2020.

Table 3.1: **Pan-cancer WGS tumor type composition.**

Abbreviation	Tumor Type	Number	Cell lines
AML	Acute myeloid leukemia	56	6
BE	Barrett's esophagus	340	0
BLCA	Bladder urothelial carcinoma	36	7
BRCA	Breast invasive carcinoma	257	35
CESC	Cervical squamous cell carcinoma	18	0
COAD	Colorectal adenocarcinoma	101	22
ESAD	Esophageal adenocarcinoma	432	1
ESSC	Esophageal squamous cell carcinoma	19	8
GBM	Glioblastoma multiforme	76	7
HNSC	Head and neck squamous cell carcinoma	62	6
KICH	Kidney chromophobe	50	0
KIRC	Kidney renal clear cell carcinoma	61	7
KIRP	Kidney renal papillary cell carcinoma	40	0
LGG	Lower grade glioma	57	5
LIHC	Liver hepatocellular carcinoma	66	10
LUAD	Lung adenocarcinoma	144	29
LUSC	Lung squamous cell carcinoma	68	15
MALY	Malignant lymphoma	112	4
MELA	Melanoma	256	34
MESO	Mesothelioma	4	2
MM	Multiple myeloma	4	4
OTHER	Other	10	8

Table 3.1 (Continued)

Abbreviation	Tumor Type	Number	Cell lines
OV	Ovarian serous cystadenocarcinoma	73	23
PACA	Pancreatic carcinoma	27	9
PBCA	Pediatric brain carcinoma	4	4
PRAD	Prostate adenocarcinoma	115	3
SARC	Sarcoma	58	8
SCLC	Small cell lung carcinoma	46	46
STAD	Stomach adenocarcinoma	64	21
THCA	Thyroid carcinoma	62	2
UCEC	Uterine corpus endometrial carcinoma	60	11
Total		2778	337

End of Table

Table 3.2: **Pan-cancer WGS datasets and sources.**

Abbreviation	Total Number	Attempted JaBbA	Analyzed	Reference	Repository
BARRETTS	347	340	340	[106]	TBD
CA	190	116	115	[83]	TBD
IPM	140	83	80	[83]	TBD
CCLE	330	326	322	[115]	SRA: PRJNA523380
ESAD-UK	423	422	422	[116]	EGA: EGAD00001004137
KLUAD	49	49	49	[117]	EGA: EGAD00001004793
MALY	101	100	100	[118]	EGA: EGAD00001002123
MELA	183	183	183	[119]	EGA: EGAD00001003388
MI	29	0	0	[120]	dbGaP: phs000488.v1.p1
NZ	131	122	122	[17]	EGA: EGAS00001001178
TCGA	1021	1017	990	[121]	dbGaP: phs000178.v10.p8
BACA	57	55	55	[35]	dbGaP: phs000447.v1.p1
Total	3001	2813	2778		

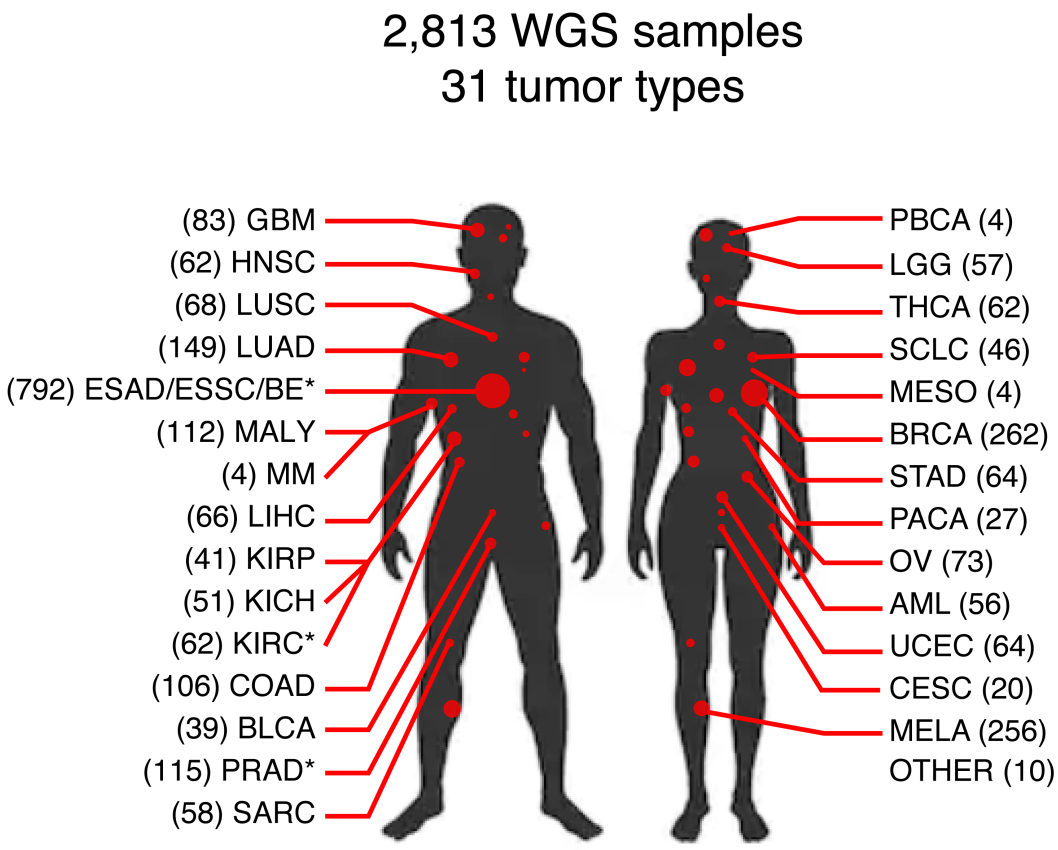

Figure 3.1: **Pan-cancer WGS tumor types.**

See Table 3.1 for abbreviations. $*$ marks datasets with multiple samples per patient.

3.1 Analysis of pan-cancer junction-balanced genome graphs

To investigate the topology of JCN across cancers, we assembled a dataset comprising 2,813 short-read WGS tumor or cell line samples spanning 31 primary tumor types (Table 3.1), including WGS for 539 previously unpublished cases (Table 3.2). In total, our analysis included 1,648 WGS samples not included in the Pan-Cancer Analysis of Whole Genomes (PCAWG) effort [2]. Application of harmonized pipelines followed by JaBbA (Figure 2.3) yielded 2,778 high quality genome graphs (Figure 3.1, see Appendix 6.5 for sample characteristics and exclusion criteria).

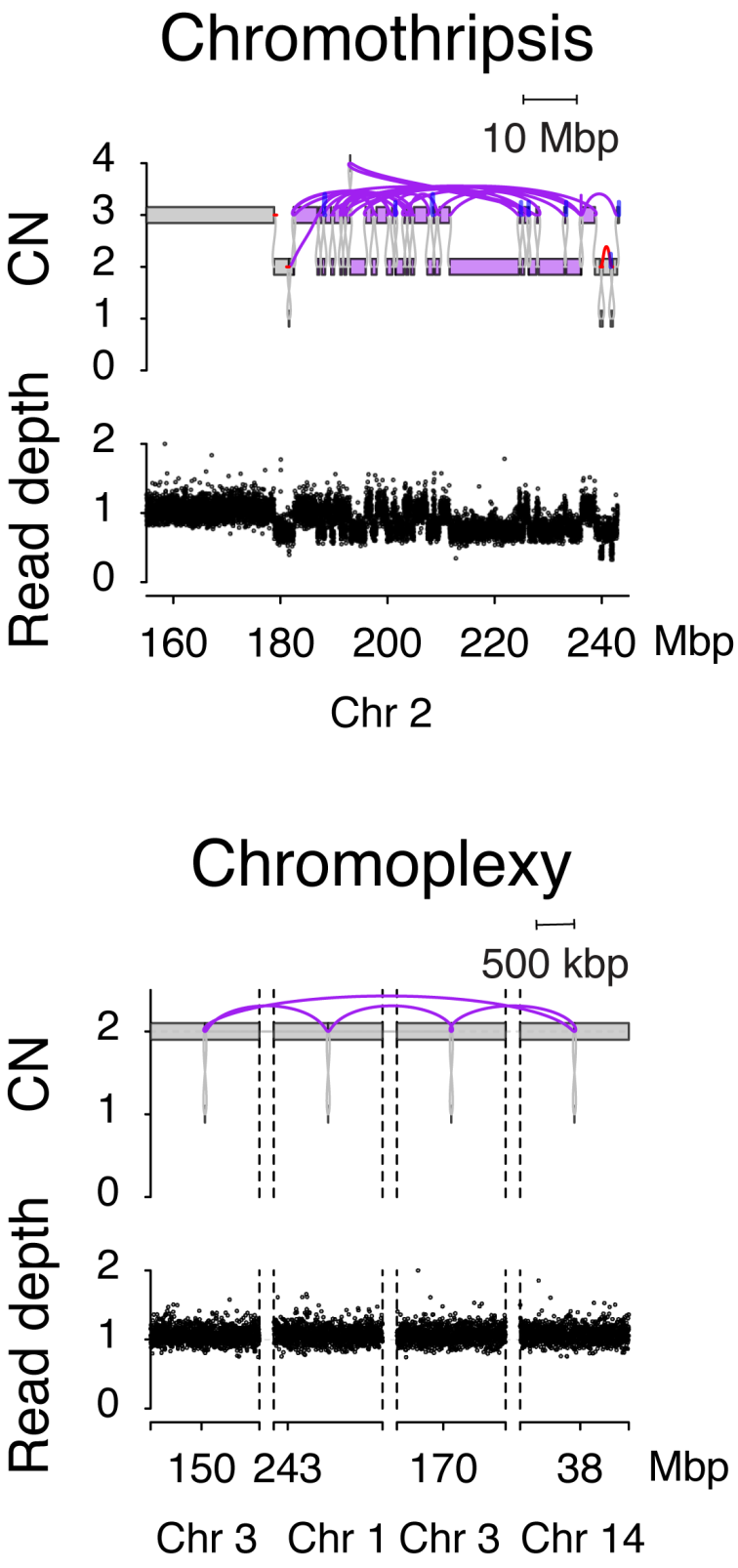

Figure 3.2: **Examples of chromothripsis and chromoplexy.**

Analyzing junction-balanced genome graph topology, we identified subgraphs associated with previously identified complex rearrangement patterns such as chromothripsis, hromoplexy, and TICs (Figure 3.2) implementing criteria described in previous publications within our framework (see Appendix 6.5).

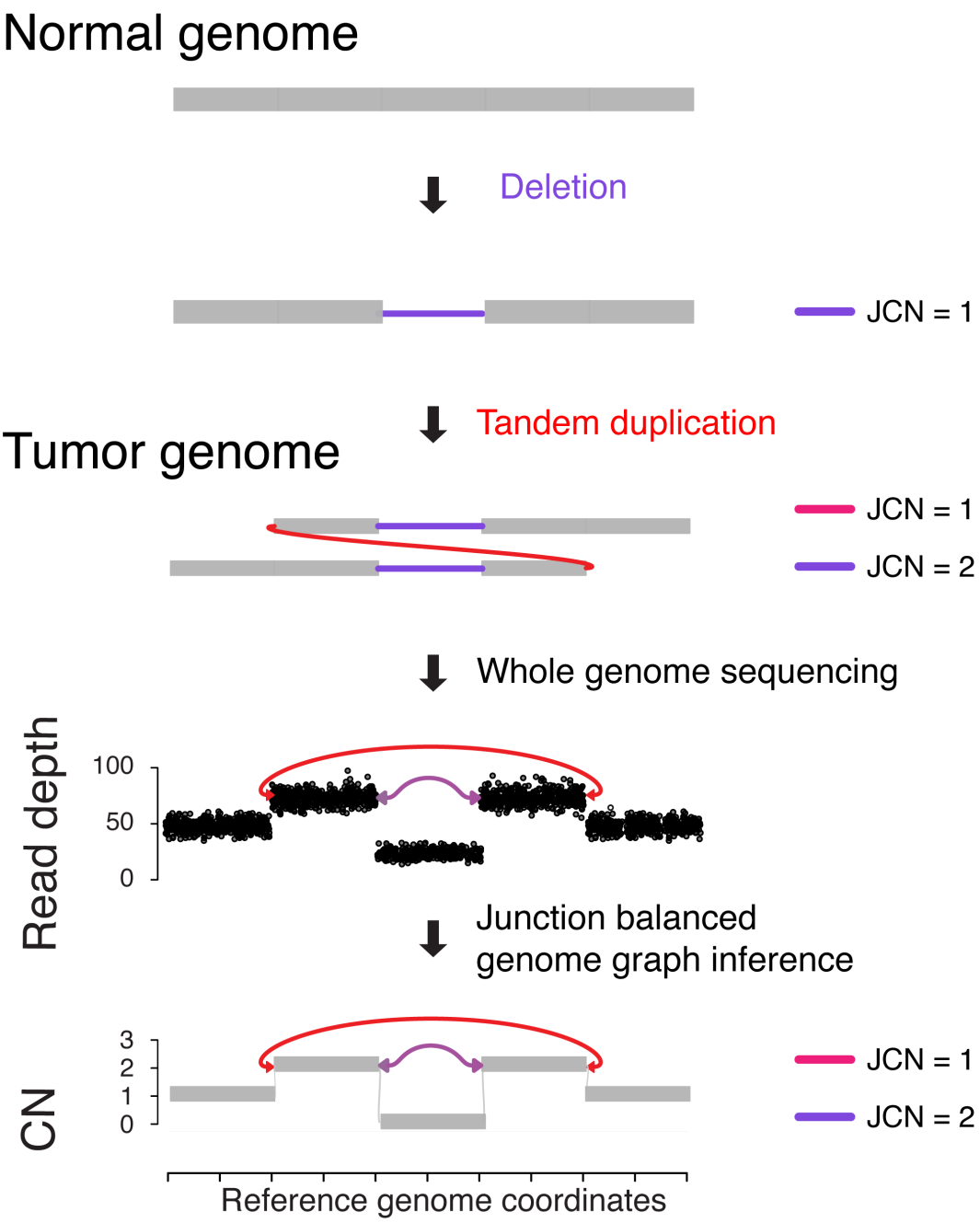

Figure 3.3: **Illustration of elevated JCN.**

Elevated junction copy number (JCN=2) arises from the duplication of an allele harboring a DEL-like junction, resulting in a characteristic read depth and junction pattern from which a junction-balanced genome graph can be reconstructed.

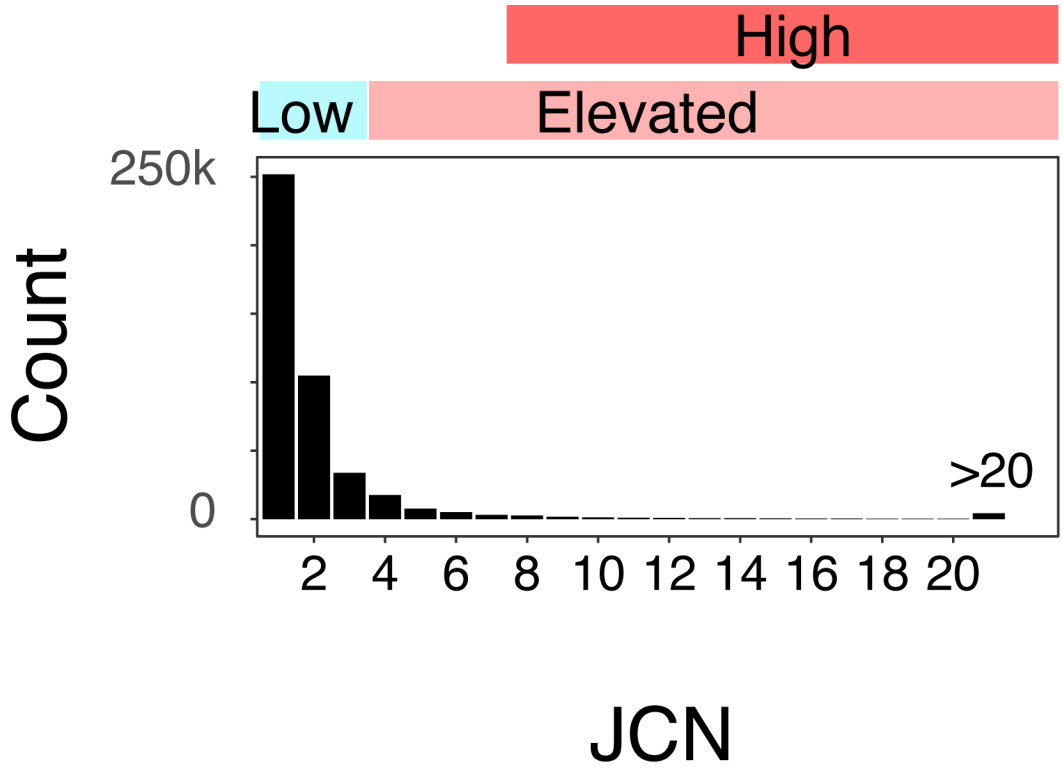

Figure 3.4: **Pan-cancer junction copy number (JCN) distribution.**

Colored bars show categories in the cohort. Low JCN, 1 to 3; elevated JCN, >3; high JCN, >7

While the vast majority of junctions demonstrated low-JCN (JCN ≤ 3), we observed a long tail of junctions with elevated– (JCN > 3) and high-JCN (JCN > 7) (Figure 3.4).

3.2 Tyfonas is massively rearranged amplicon associated with elevated number of fusion genes

We then sought to investigate the rearrangement patterns associated with high-JCN junctions (JCN > 7) in our genome graphs. *A priori*, a junction at such an extreme of JCN may evolve through a double minute, BFB cycles, or an as yet undescribed mechanism for duplicating already rearranged DNA. To characterize classes of amplification events associated with these high-JCN junctions, we first identified 12,588 subgraphs harboring an interval CN of at least twice ploidy among the 2,487 unique genome graphs (by patient) (Figure 3.5), then identified among these amplicons (amplified clusters within a genome) those that harbor at least one junction with JCN > 7. We annotated the resulting 1,703 high-JCN amplicons according to three features: 1) the maximum JCN normalized by the maximal interval CN, 2) the *sum* of fold back inversion JCNs (INV-like junctions that terminate and begin at nearly the same location in the genome) relative to the maximal interval CN, and 3) the number of junctions with elevated JCN (JCN>3) (see Appendix 6.5).

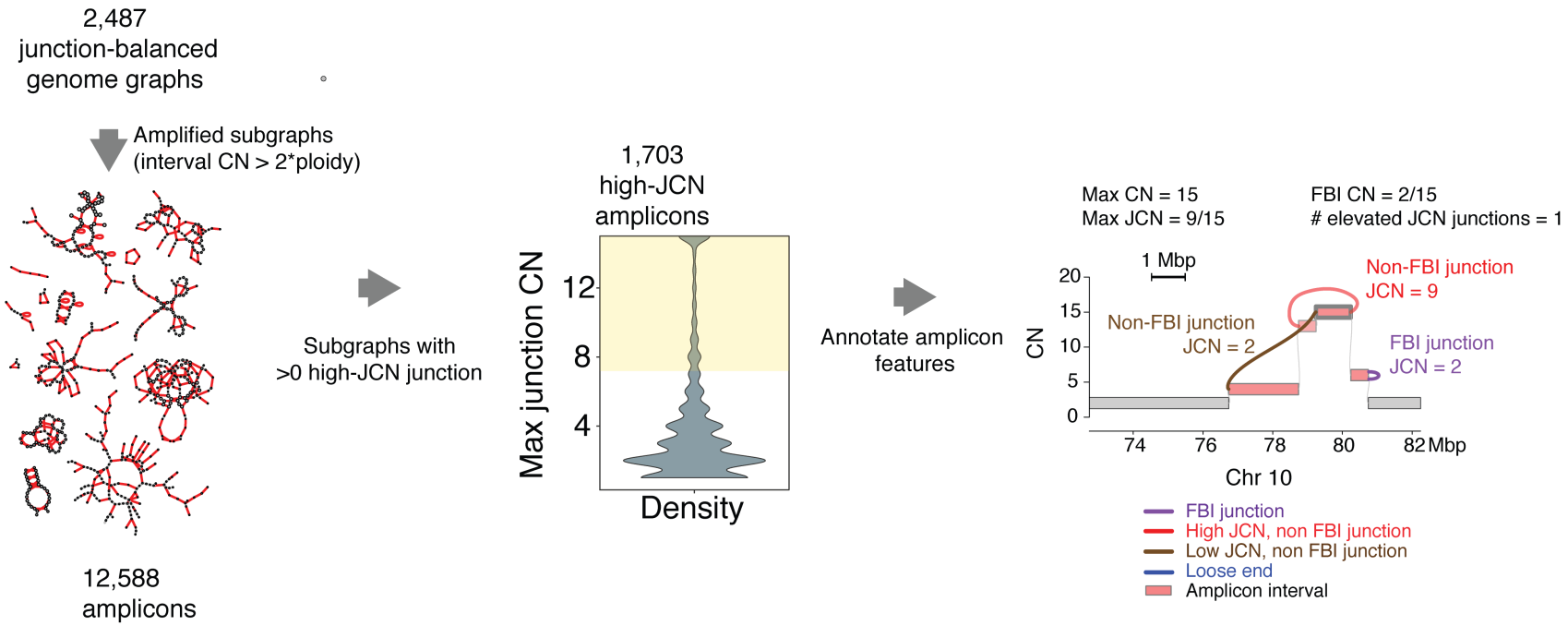

Figure 3.5: **Candidate amplicon subgraphs and the feature space construction.**

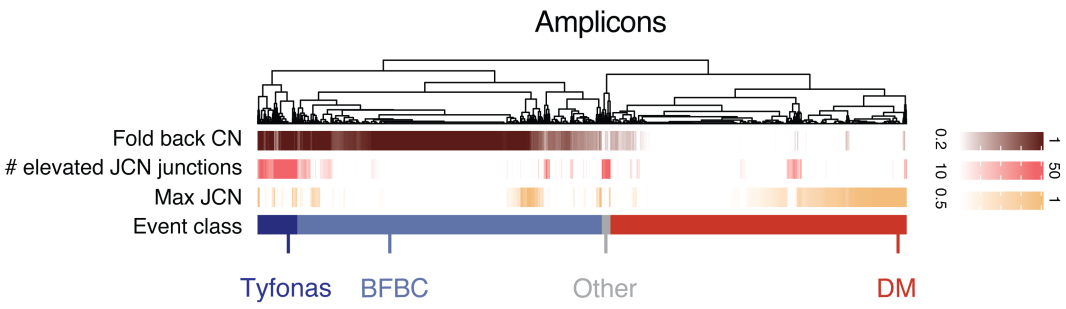

Figure 3.6: **Classification of amplicons with high-JCN junctions.**

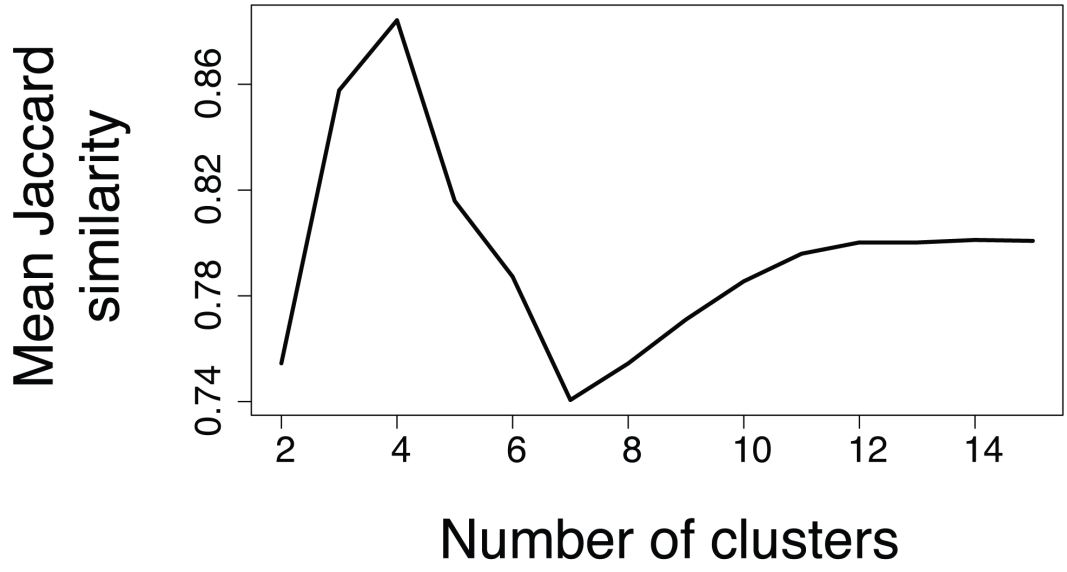

Figure 3.7: **Assessment of stability of amplicon clustering.**

For a given setting of k, cluster stability was computed as the mean Jaccard similarity between each observed cluster and the most similar cluster across 100 75% bootstraps of the data (see Appendix 6.5). The stability of each k parameter setting was computed as the mean of the k stability scores across the clusters generated at that setting.

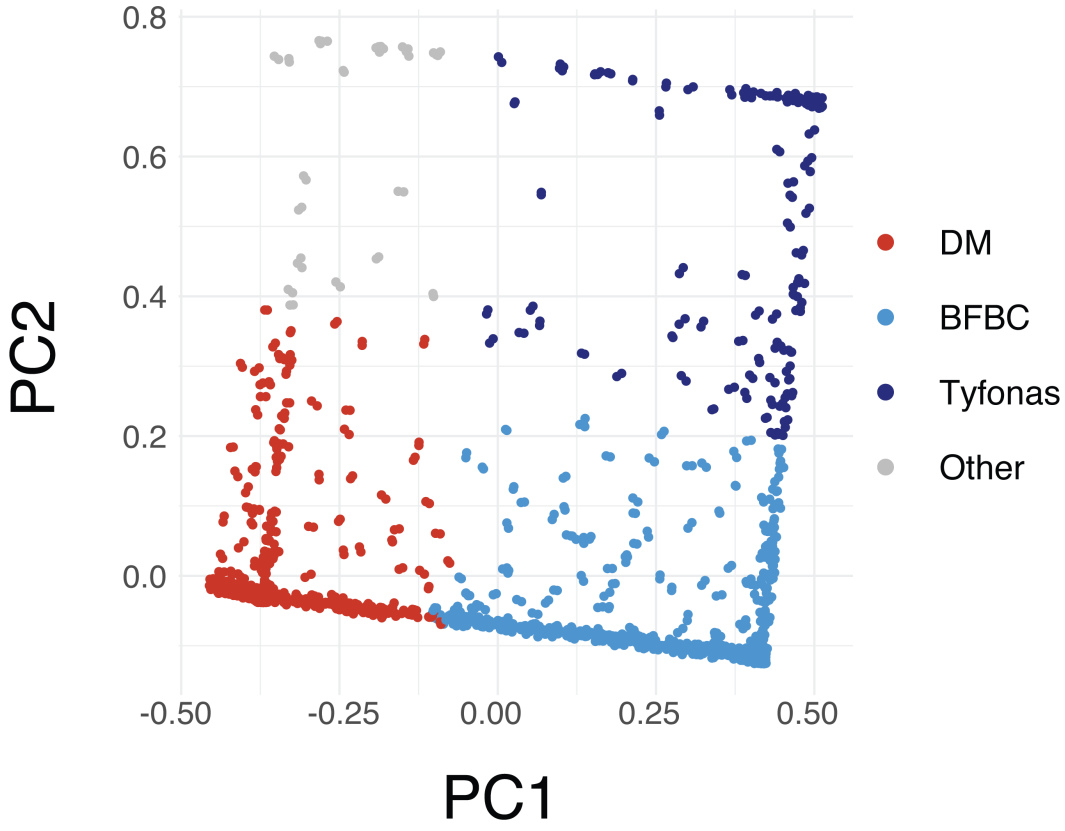

Figure 3.8: **Projection of the amplicons across the first two principal components of normalized amplicon features.**

Clustering and classification of amplicons (Figure 3.2) on the basis of these three features yielded three stable clusters (Figure 3.2,3.2,3.2; see also Appendix 6.5). Upon visual inspection, the first group, harboring low fold-back JCN but high maximal JCN, contained amplicons comprising a single high-JCN junction forming a high CN circular path in the graph (Figure 3.2) as well as more complex cyclic patterns spanning multiple discontiguous loci, consistent with a double minute (or extrachromosomal circular DNAs). The second group, demonstrating high fold-back JCN (> 0.5), a low burden of elevated-JCN junctions (< 26), and a "stair step" pattern of copy gains, was consistent with a BFB cycles (Figure 3.2, [122, 123]). The third group contained both high fold-back JCN (≥0.50) and a significant burden of elevated JCN junctions (≥26). Upon visual inspection, these amplicons comprised dense webs of elevated JCN junctions across subgraphs consisting of > 100 Mbp of genomic material and often reaching CNs higher than 50 (Figure 3.11). We dubbed these extremely large amplicons, which did not fit in previously defined categories, *tyfonas* (τύφωνας, Greek meaning typhoon) (Figure 3.11).

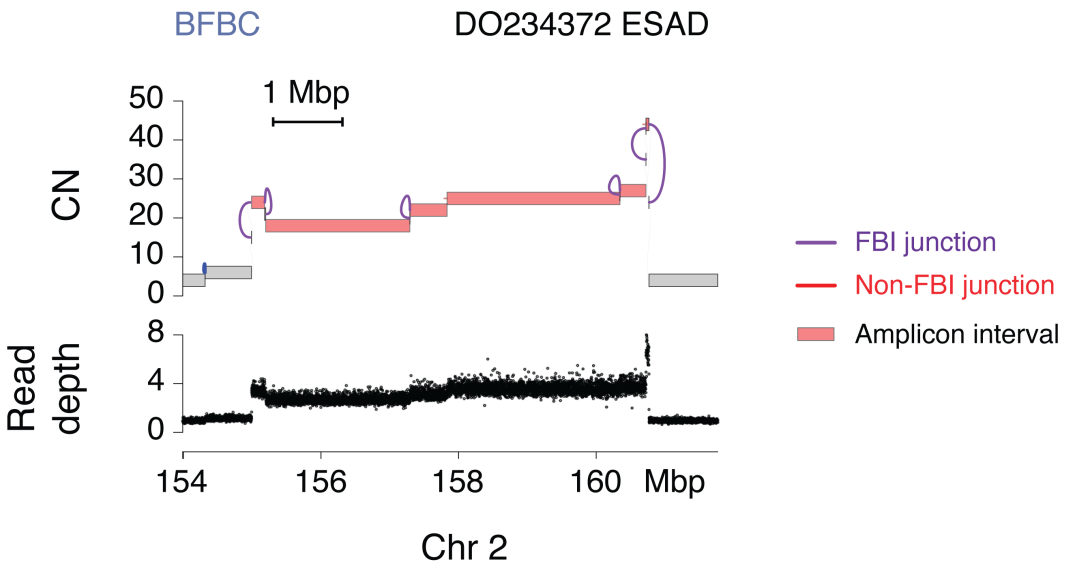

Figure 3.9: **An example of a BFB cycles pattern.**

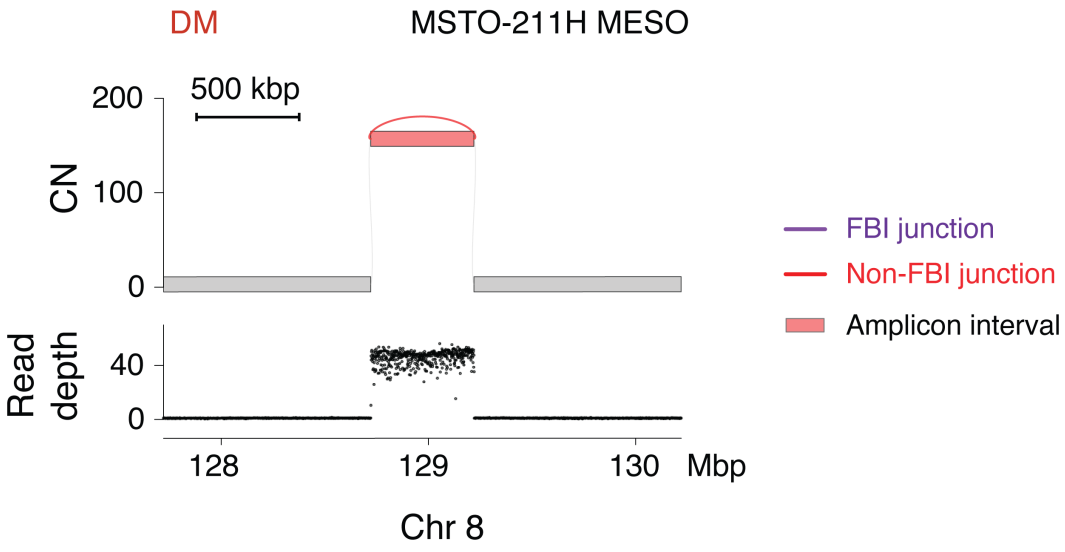

Figure 3.10: **An example of a double minute pattern.**

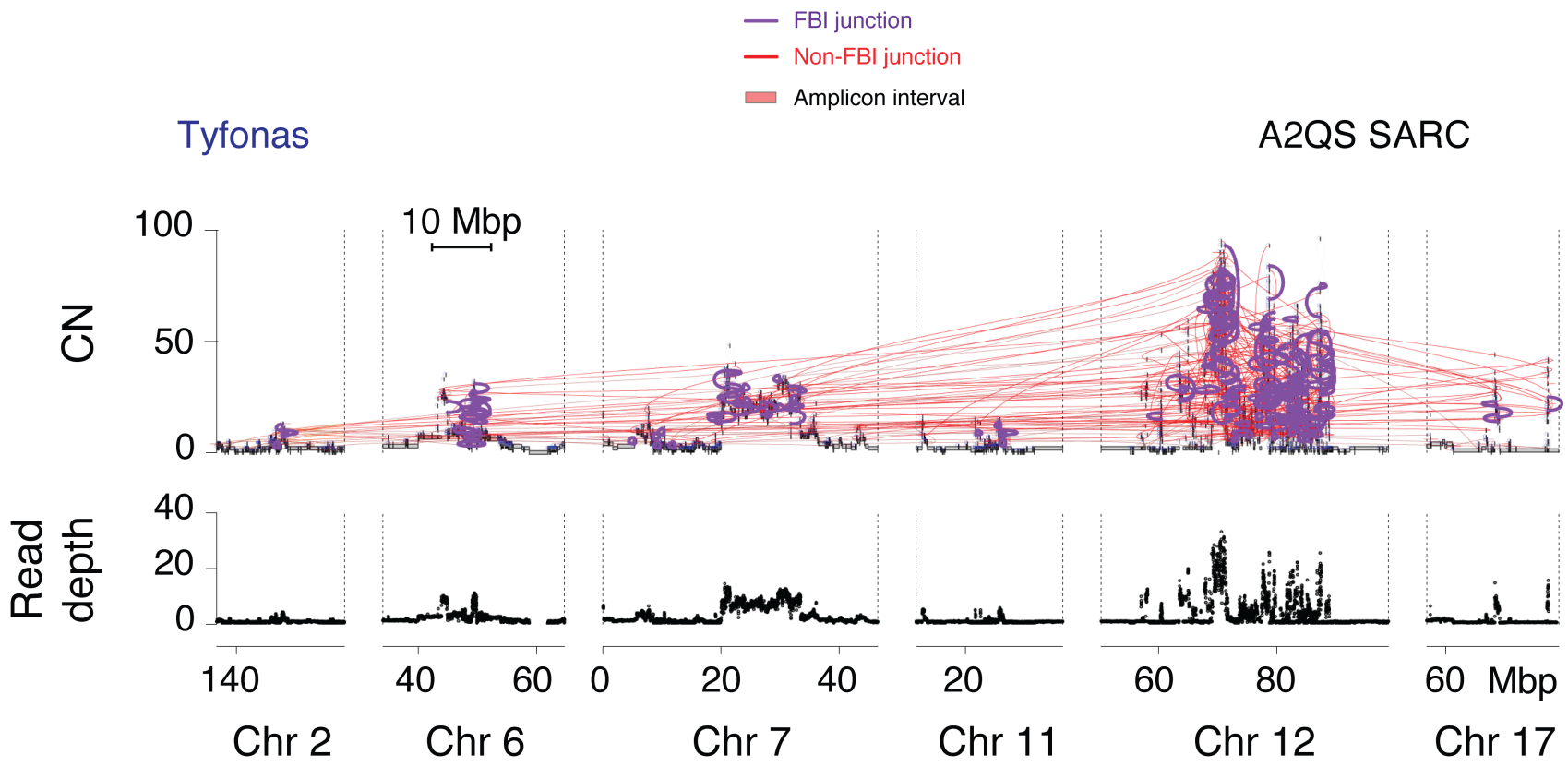

Figure 3.11: **An example of a tyfonas pattern.**

Tyfonas are distinct from BFB cycles and double minutes

Comparing amplicon features, we found that tyfonas had significantly larger genomic mass (summed interval width weighted by CN, Figure 3.2), maximum interval CN (Figure 3.2), junction burden (Figure 3.2), and JCN entropy (based on the distribution of JCN in the amplicon, the higher the more diversity among the JCNs, Figure 3.2) than either double minutes or BFB cycles ($P < 2.2 \times 10^{-16}$, Wilcoxon rank-sum test). All three amplicon classes frequently intersected CGC cancer genes residing within GISTIC pan-cancer amplification peaks (Figure 3.16, recurrent amplification peaks from [124]). *EGFR* was the most frequent target of double minutes. BFB cycles were most frequently implicated in *ERBB2*, *CCND1*, and *CCNE1* amplification. Tyfonas were associated with *MDM2* and *CDK4*. While double minutes were enriched in glioblastoma and small cell lung cancer, BFB cycles were enriched in esophageal squamous cell cancer, lung squamous cell cancer, and head and neck squamous cell cancer (FDR < 0.25, Fisher's exact test, Figure 3.17). In contrast, tyfonas events were enriched in sarcoma, breast cancer, and melanoma (FDR < 0.25).

Analyzing tumor subtypes, we found tyfonas were common in dedifferentiated liposarcomas (80%) and acral melanomas (40%) but rarely seen (< 2%) in fibrosarcomas and cutaneous melanomas (Figure 3.2). 80% of dedifferentiated liposarcomas harbored a tyfonas on chromosome 12q, suggesting that tyfonas represent the genomic footprint of supernumerary ring chromosomes in this tumor type [125]. Given the enrichment of tyfonas events in acral melanoma, an immunotherapy-responsive melanoma with a low SNV burden [126], we hypothesized that tyfonas events may provide an alternate source of neoantigens through the generation of expressed protein-coding fusions. Analyzing the sub-

set of our samples with RNA-seq data within the TCGA dataset, we found that tyfonas were significantly enriched with expressed fusion transcripts relative to double minutes, BFB cycles, and even chromothripsis (CT) (Figure 3.2, left), even after accounting for the larger genomic footprint of tyfonas (Figure 3.2, right).

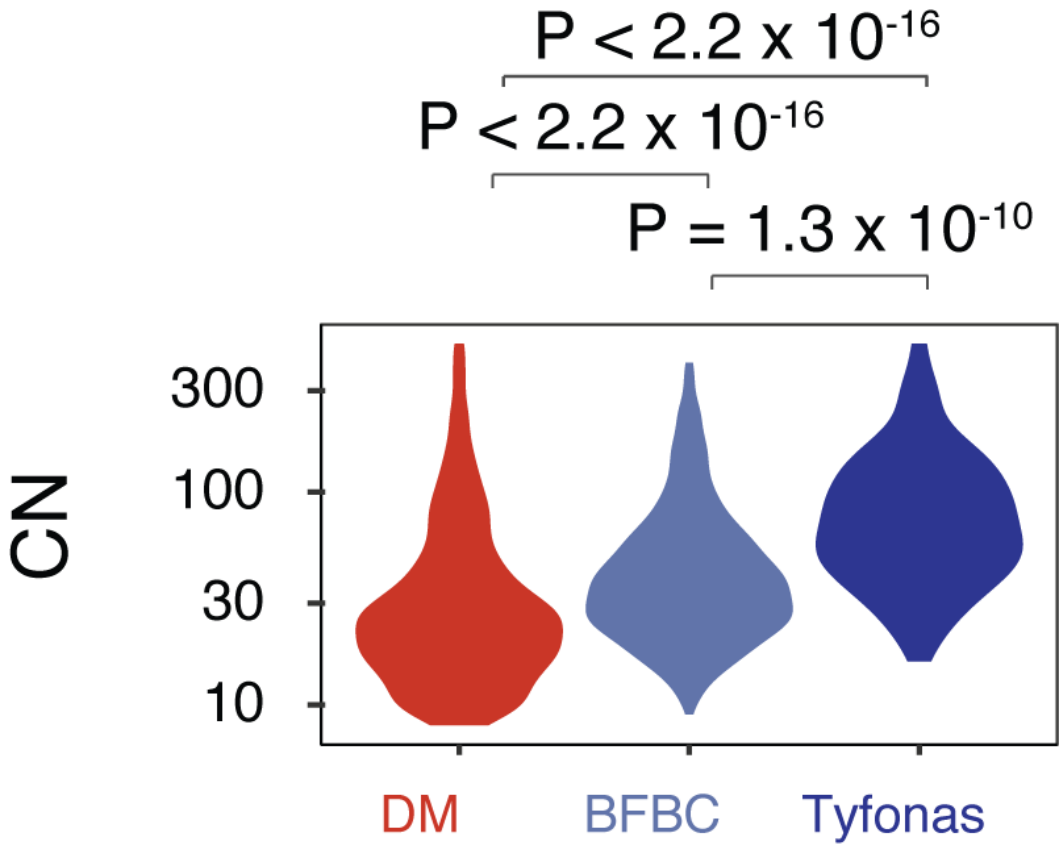

Figure 3.12: **Tyfonas amplify parts of genome to higher CN states than BFB and DM.**

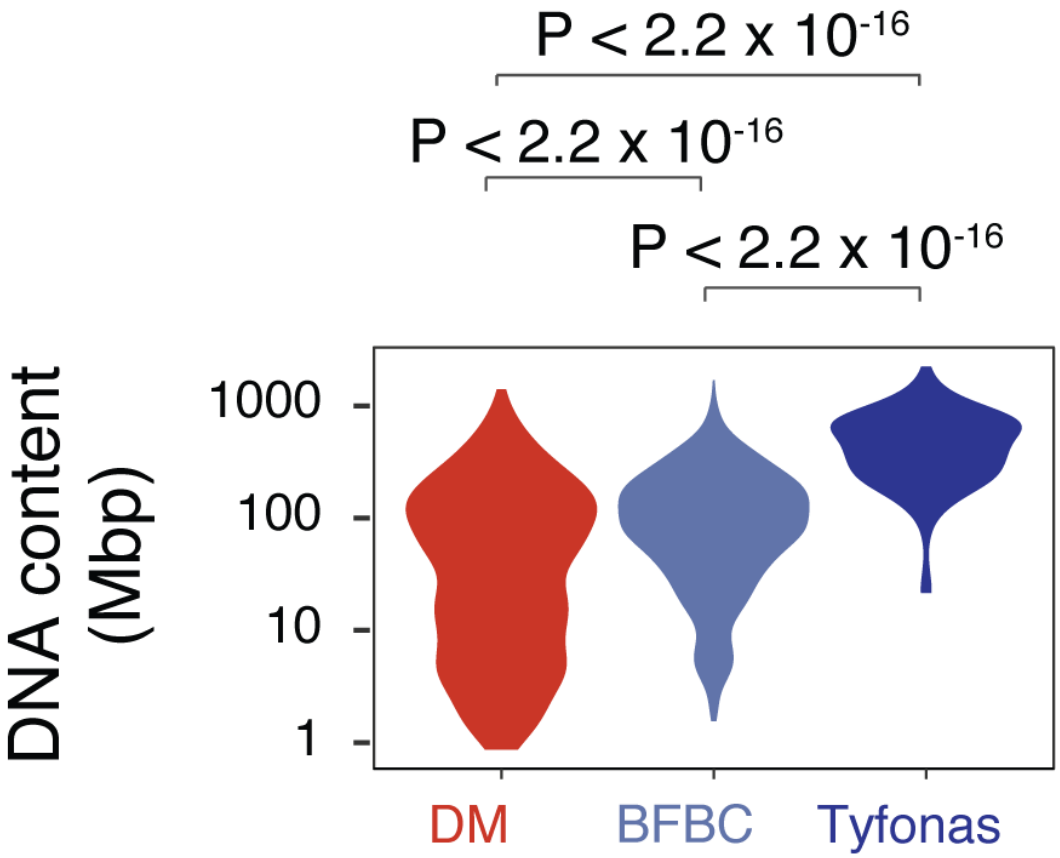

Figure 3.13: **Tyfonas result in larger mass of amplified genomic DNA than BFB and DM.**

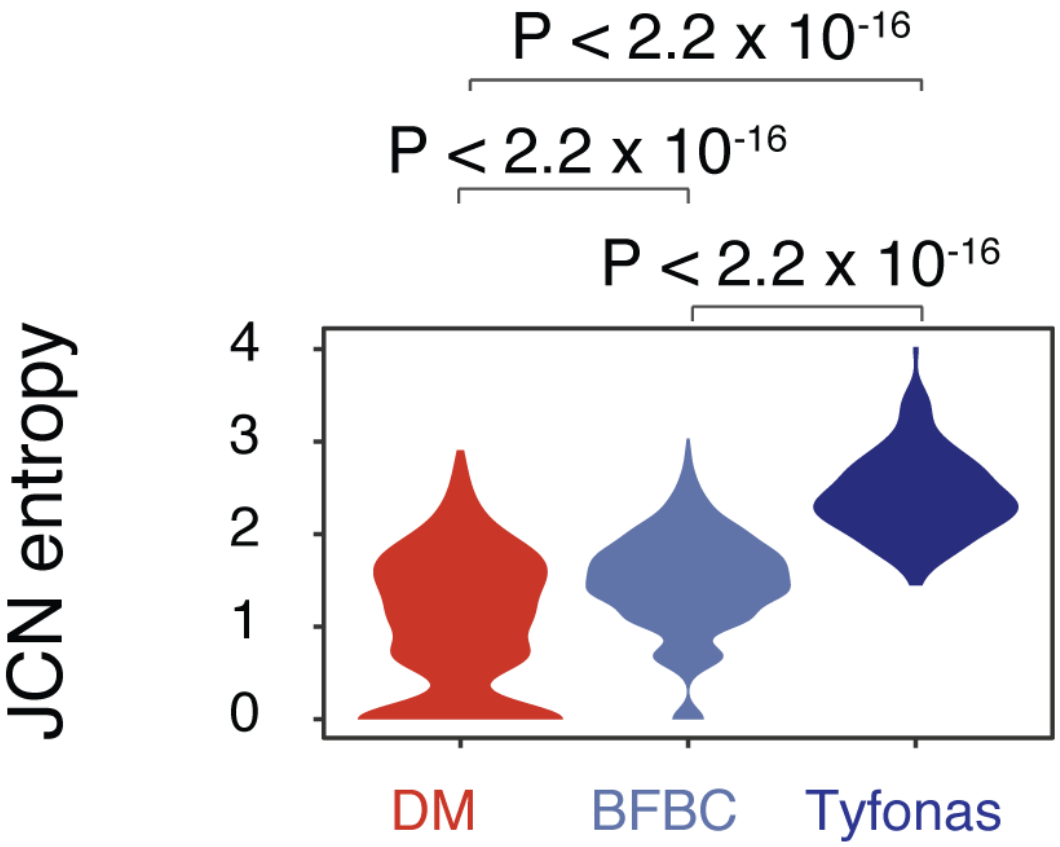

Figure 3.14: **Tyfonas contain more heterogeneous junction copy numbers than BFB and DM.**

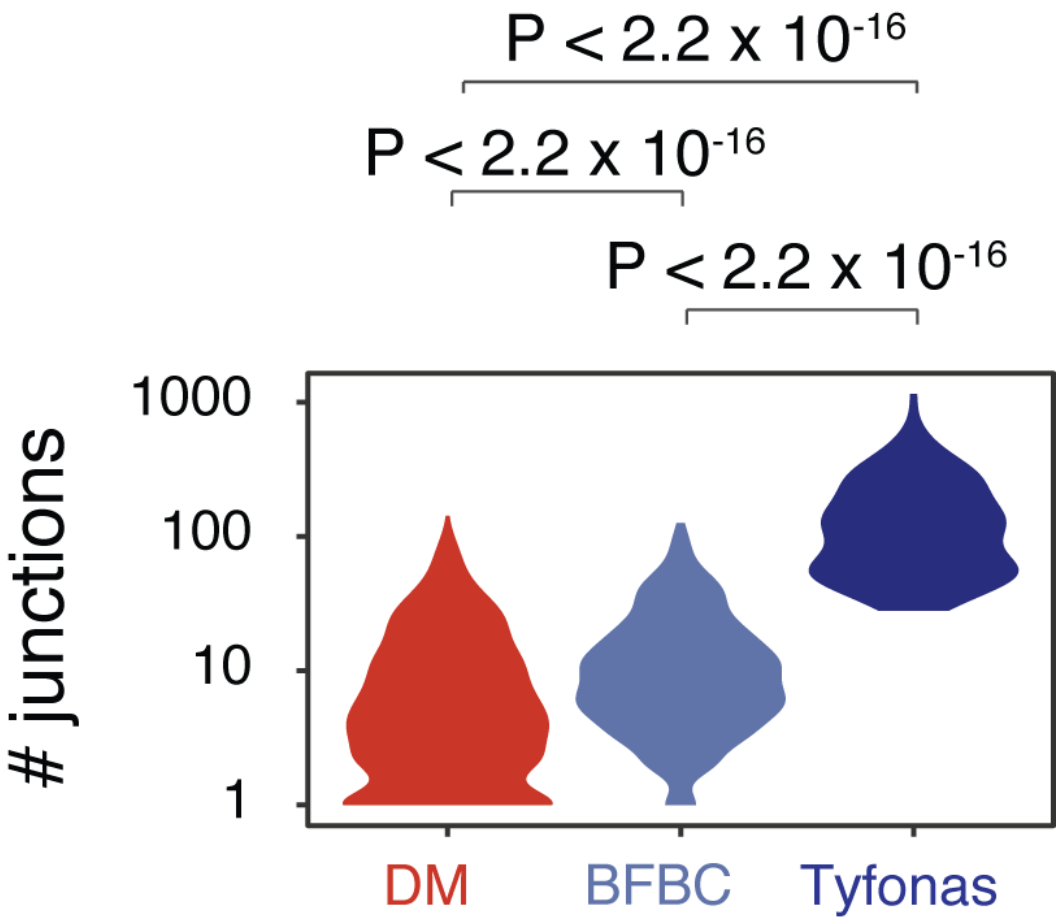

Figure 3.15: **Tyfonas contain more junctions than BFB and DM.**

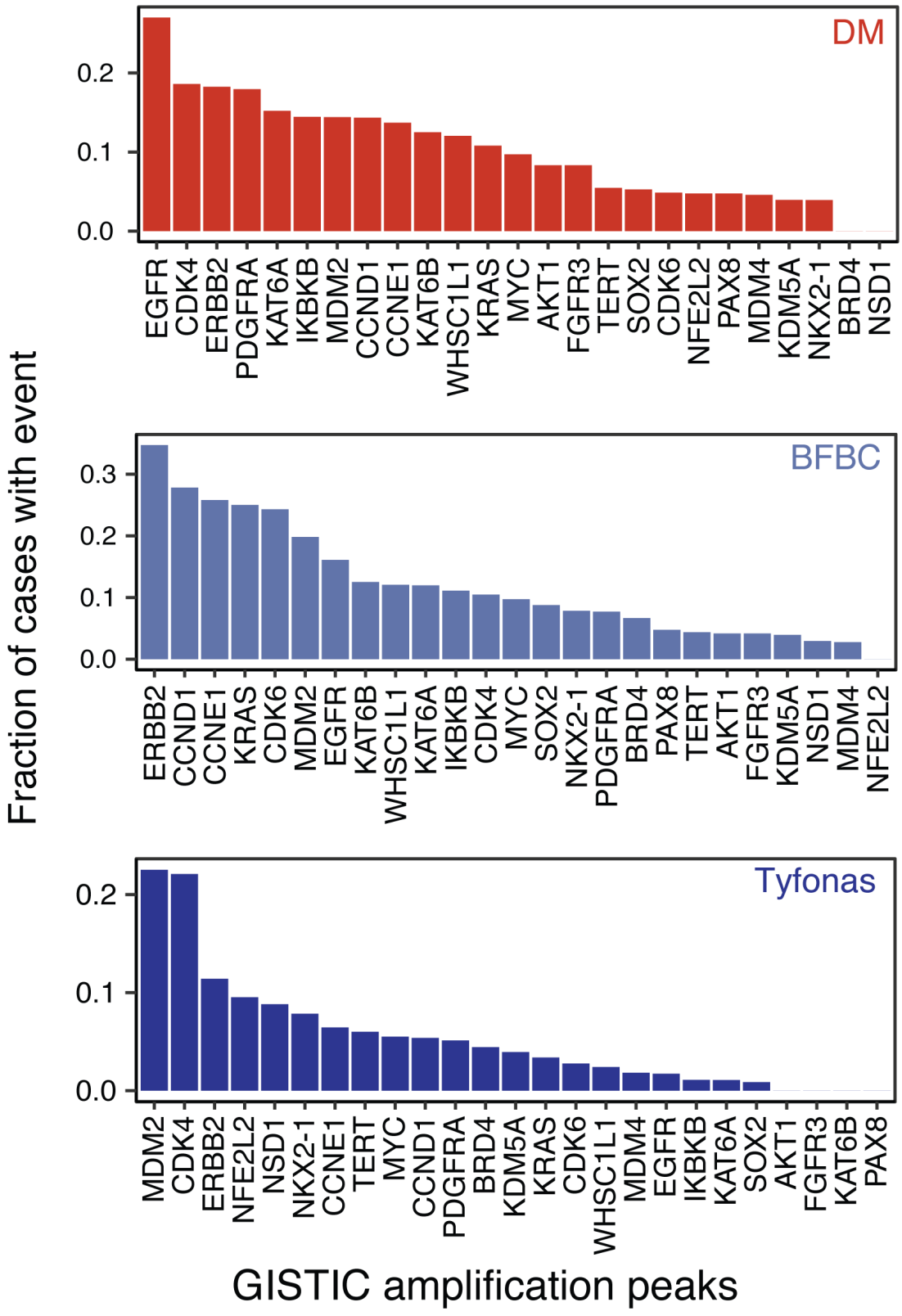

Figure 3.16: **Tyfonas, DM, and BFB recurrently amplify different oncogenes**

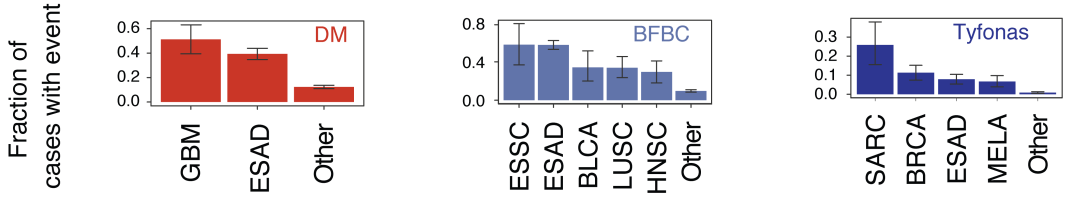

Figure 3.17: **Tyfonas, DM, and BFB enrich in different tumor types.**

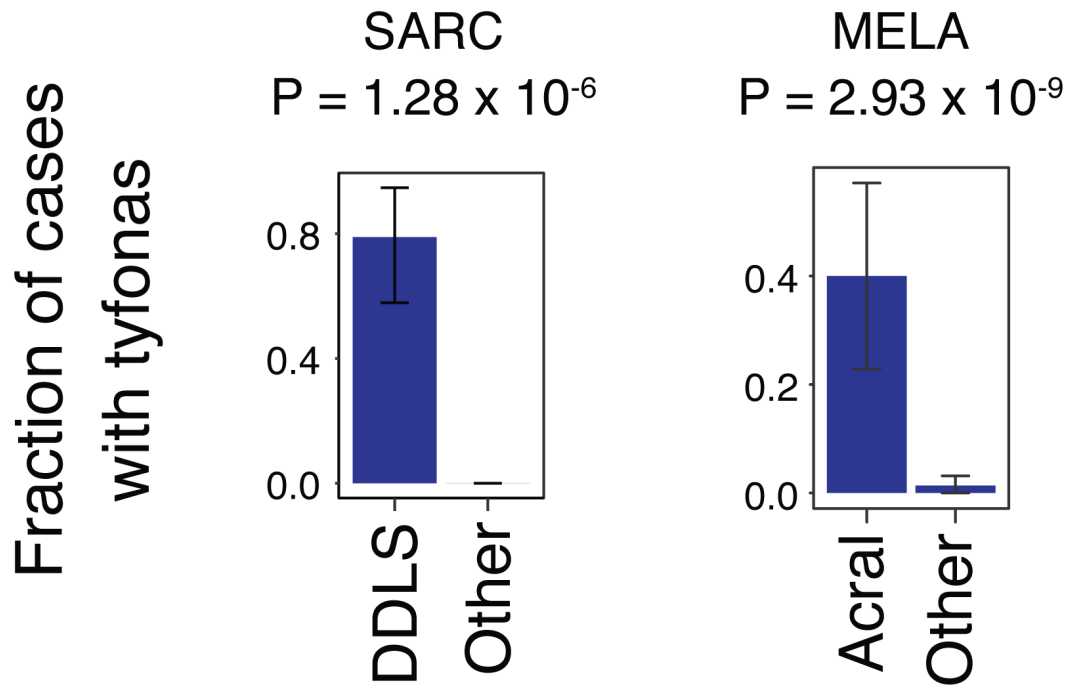

Figure 3.18: **Tyfonas is highly enriched in liposarcoma and acral melanoma.**

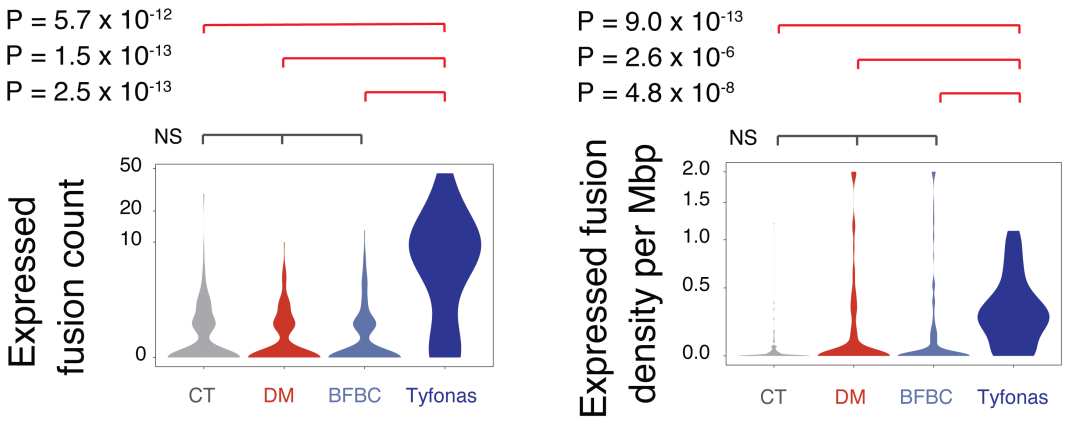

Figure 3.19: **Tyfonas create more fusion genes and highly expressed fusion genes than BFB, DM, and chromothripsis.**

Enrichment of expressed protein-coding fusion transcripts by count (left) and density per Mbp of event territory (right) in tyfonas relative to other amplicon types and CT (Wilcoxon rank-sum test). Error bars on bar plots represent 95% confidence intervals on the Bernoulli trial parameter.

3.3 Clusters of pan-cancer patients based on SV event burdens show differential prognosis and is associated with tumor types and genetic backgrounds

Tallying normalized junction counts across 13 event categories and 2,487 patients, we found 14 stable clusters using standard model selection metrics (Figure 3.22). Most clusters were dominated by 1-3 event types (e.g. CT = chromothripsis, BR = BFB cycles, rigma, DDT = deletion, duplication, TIC) with the exception of two: QUIET (few events) and SPRS (sparse, miscellaneous events).

Consistent with previous reports, the CT cluster was significantly enriched in prostate adenocarcinoma (PRAD, $P = 2.05 \times 10^{-5}$, $OR = 1.99$, single-sided z-test, Bayesian logistic regression, [127]) and glioblastoma multiforme (GBM, $P = 5.00 \times 10^{-8}$, $OR = 2.61$, [128], Figure 3.23, 3.24). Similarly, the CP (chromoplexy) cluster was significantly enriched in PRAD ($P = 2.32 \times 10^{-10}$, $OR = 3.18$, [35]). DDT tumors (defined by high burdens of deletions, duplications, and templated insertion chains) were enriched in triple-negative breast cancer (TNBC) ($P < 2.2 \times 10^{-16}$, $OR = 8.80$), ovarian cancers ($P = 7.03 \times 10^{-16}$, $OR = 6.89$), and more broadly sex-hormone driven tumors ($P = 3.18 \times 10^{-14}$, $OR = 19.0$).

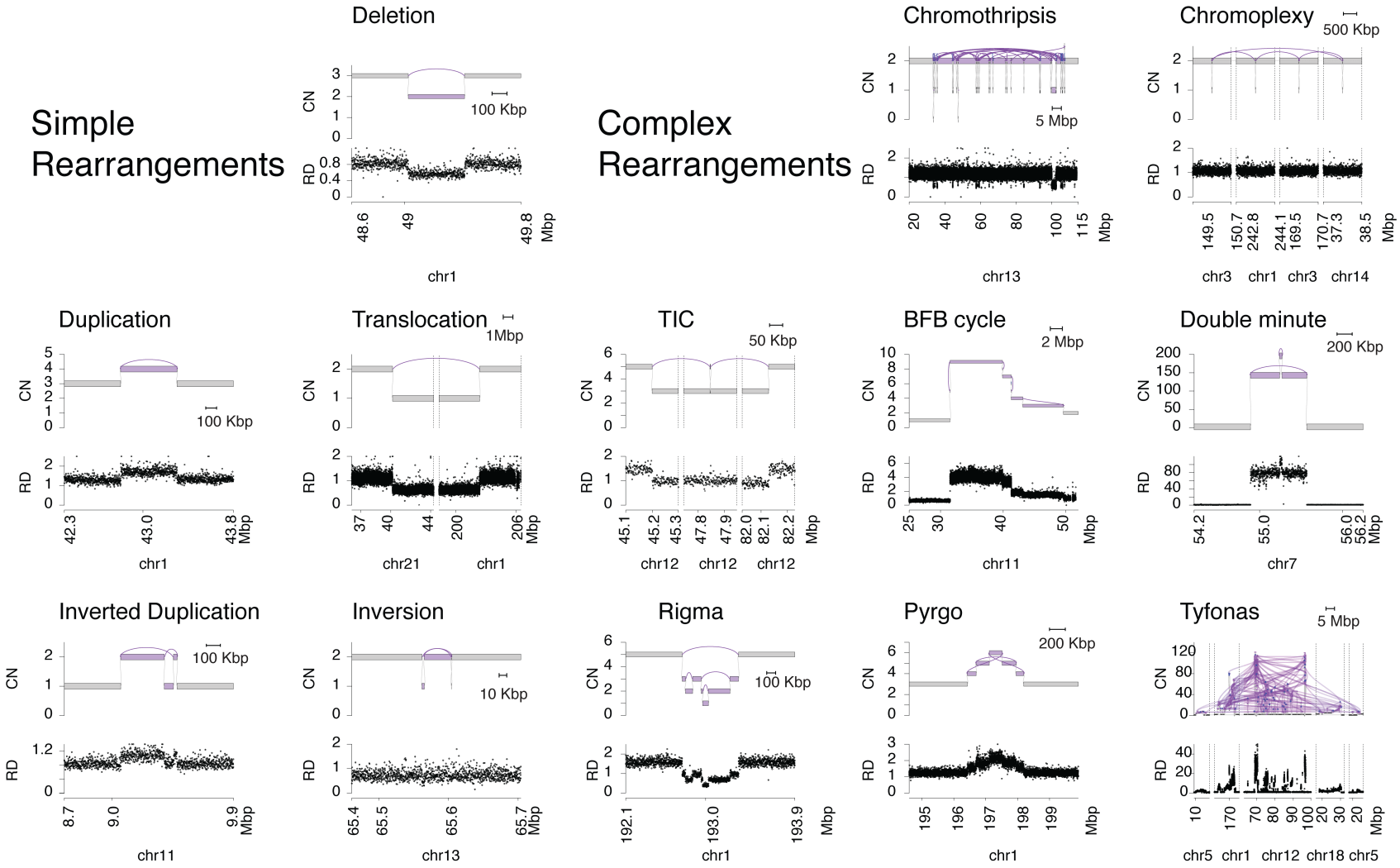

Figure 3.20: **Dictionary of genome graph-derived event patterns**

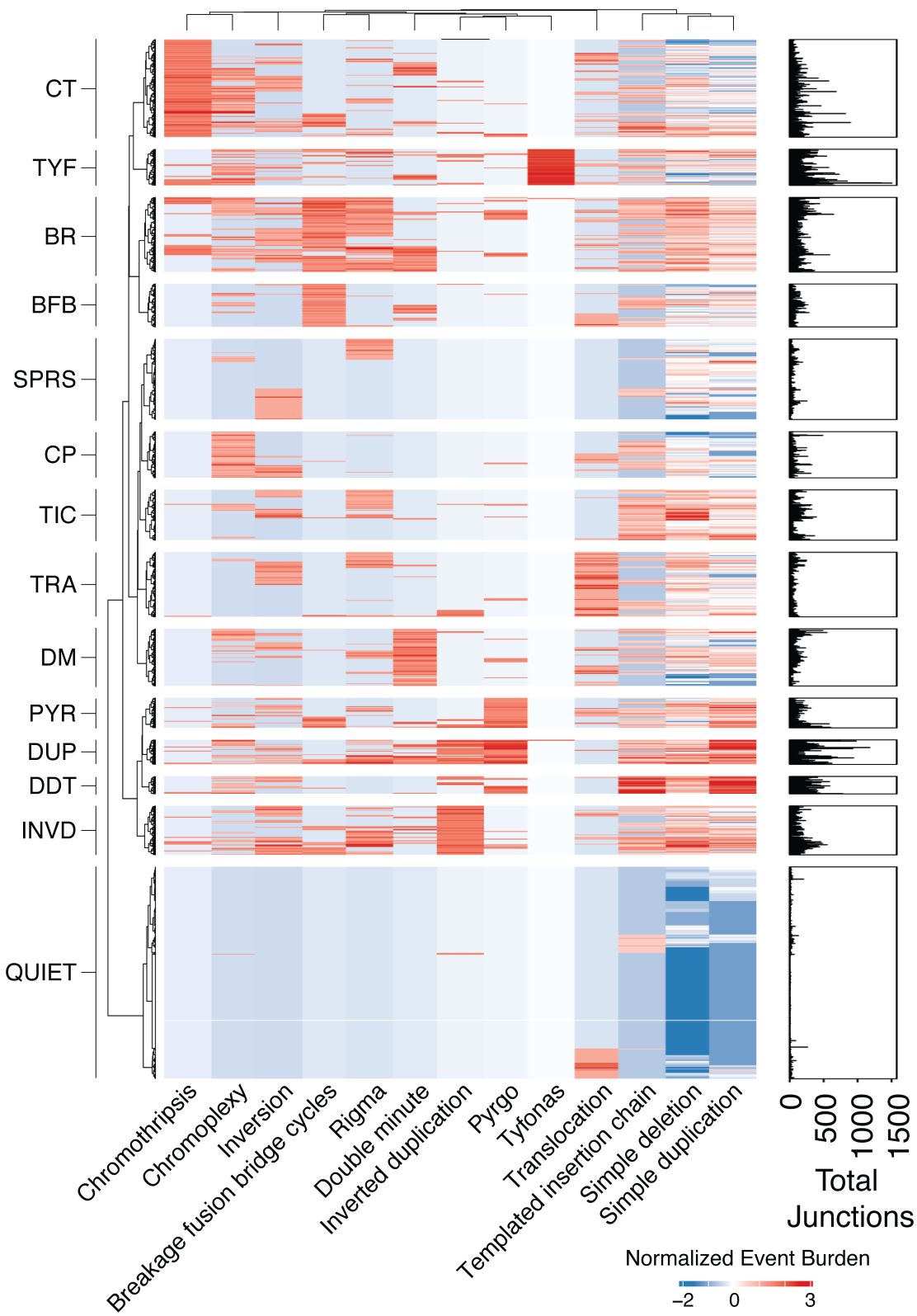

Figure 3.21: **Genome graph-derived features define biologically distinct patient groups**

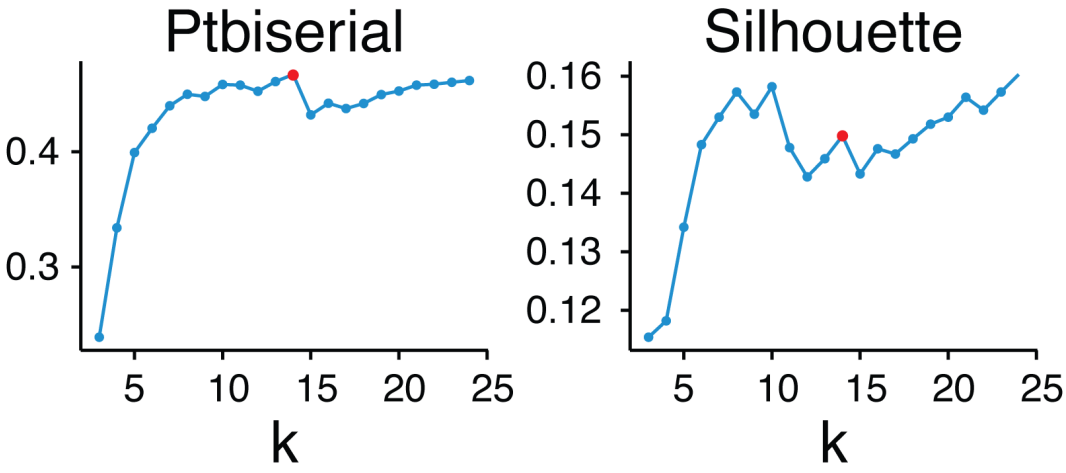

Figure 3.22: **Metrics for determining the optimal number of clusters.**

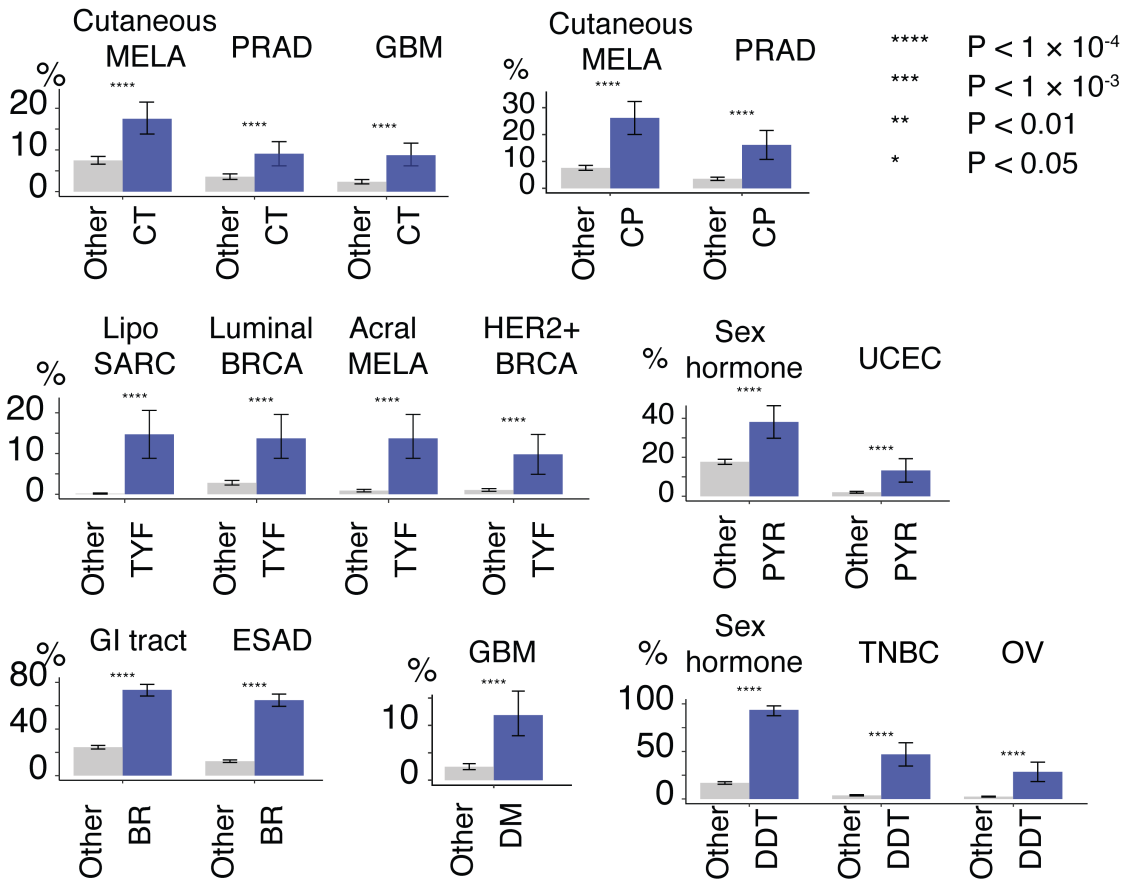

Figure 3.23: **Selected associations between clusters and tumor types.**

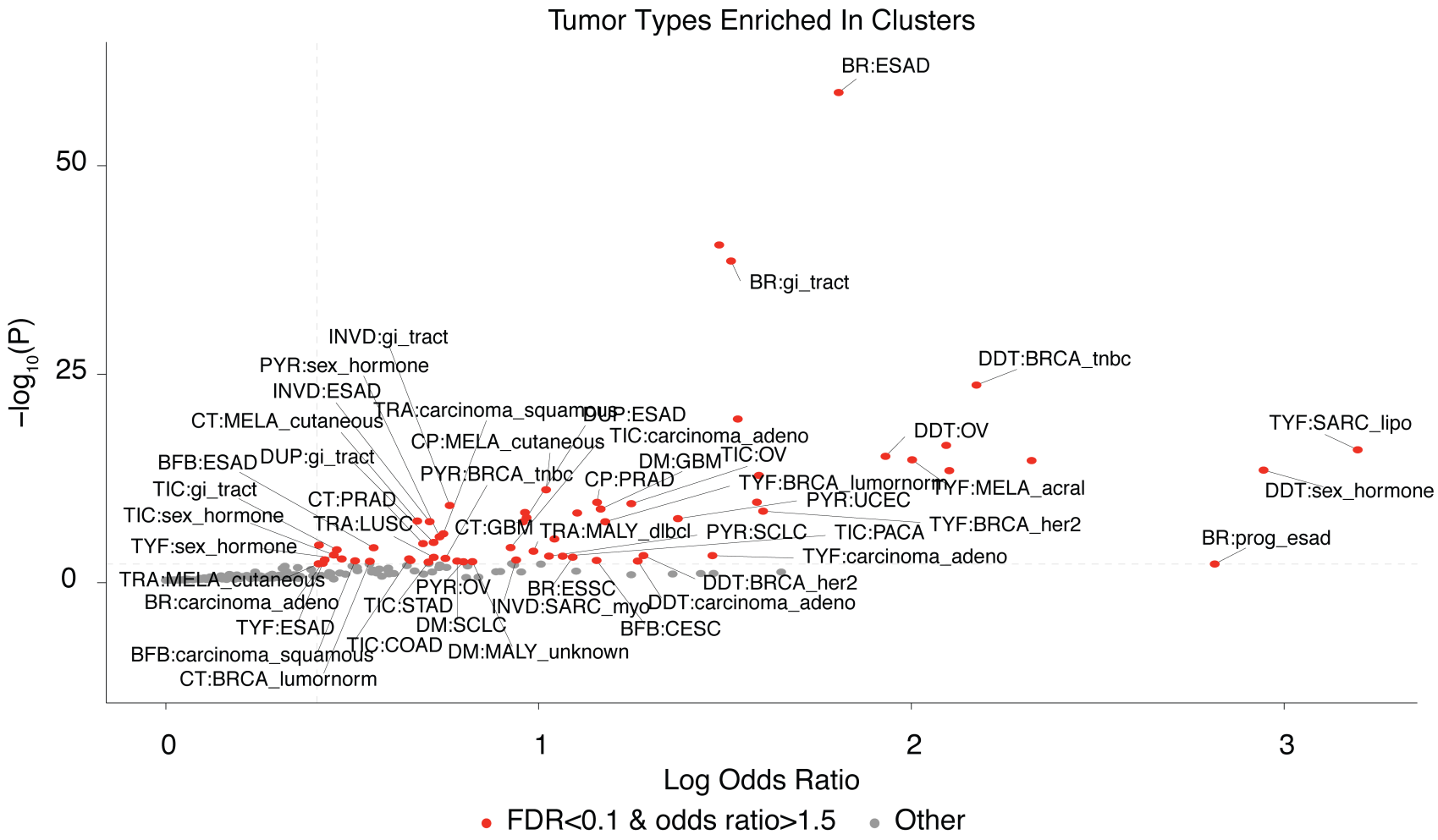

Figure 3.24: **All associations between clusters and tumor types.**

Inspection of the heatmap in Figure 3.21 showed that the classes of complex SV introduced in this study (pyrgo, rigma, tyfonas) largely clustered independently from known complex SV types (double minute, BFB cycles, chromothripsis, chromoplexy). Among these, the BR (BFB cycles and Rigma dominated) cluster was primarily (60%) composed of ESAD cases ($P < 2.2\times10^{-16}$, $OR = 6.08$) and enriched in gastrointestinal tumors (e.g. esophageal, colorectal, and gastric adenocarcinoma) ($P < 2.2\times10^{-16}$, $OR = 4.56$) (Figure 3.21,3.24). The TYF (tyfonas dominated) cluster was enriched in both luminal breast cancer ($P = 4.87 \times 10^{-8}$, $OR = 3.25$), HER2+ breast cancer ($P < 2.64 \times 10^{-9}$, $OR = 4.96$), dedifferentiated liposarcoma ($P < 2.2 \times 10^{-16}$, $OR = 24.5$), and acral melanoma ($P < 1.84 \times 10^{-15}$, $OR = 7.40$). In contrast, cutaneous melanomas were enriched in the CT cluster. Additional associations are shown in Figure 3.24 and [83] Table S3.

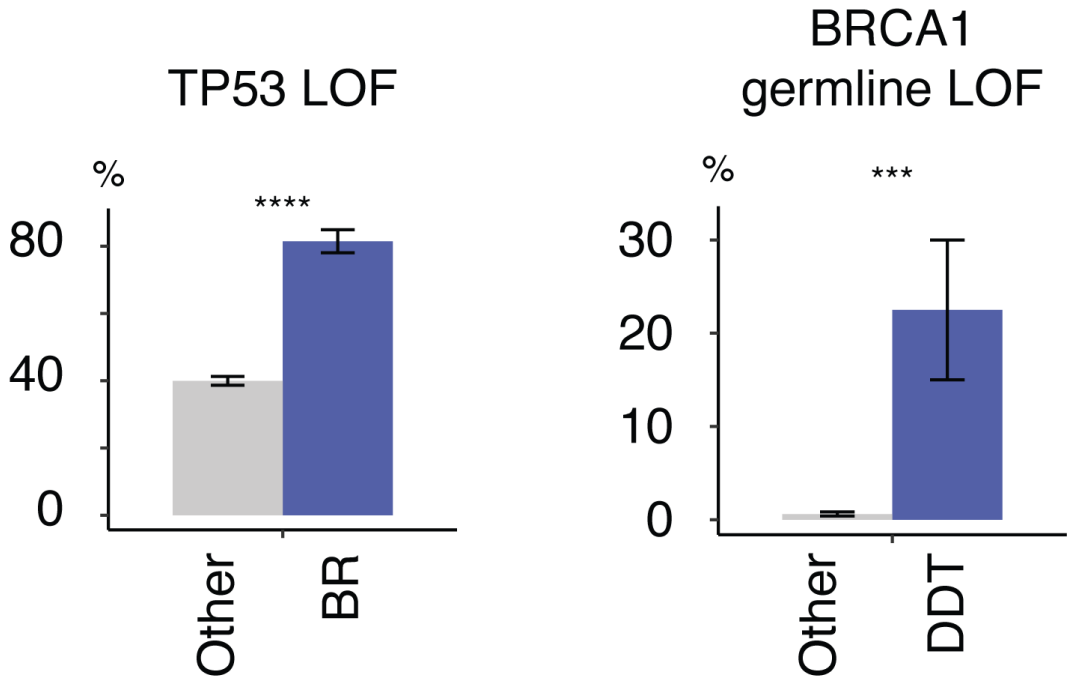

Figure 3.25: **DDT cluster is associated with BRCA1 familial LoF alterations and BR cluster with *TP53* LoF.**

We associated somatic or constitutional genotypes in COSMIC Cancer Gene

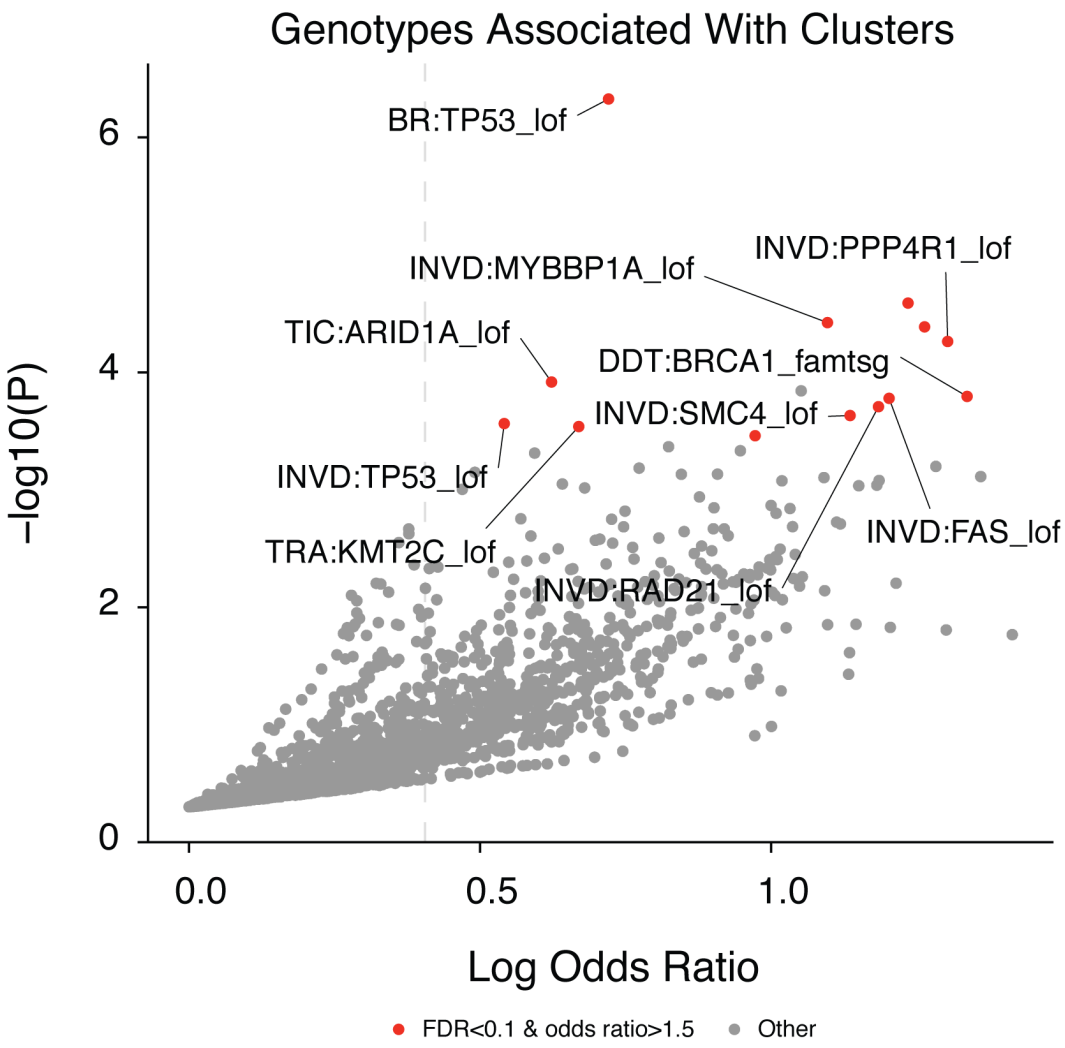

Figure 3.26: **All associations between clusters and LoF alterations of tumor suppressors or DNA damage response genes.**

Census genes with cluster membership, after correcting for tumor subtype as a covariate. More than 20% of cases in the DDT cluster harbored constitutional ($P = 1.60 \times 10^{-4}$, $OR = 3.81$) loss of function lesions in *BRCA1* (Figure 3.25). BR-cluster tumors were also significantly enriched in somatic *TP53* mutations ($P < 4.71 \times 10^{-7}$, $OR = 2.06$). Additional somatic genotype associations (Figure 3.26), including an enrichment of *SMC4* ($P = 2.33 \times 10^{-3}$, $OR = 3.11$) and *RAD21* ($P = 1.96 \times 10^{-3}$, $OR = 3.27$) mutations in the INVD (inverted duplication dominant)

cluster, *ARID1A* ($P = 1.21 \times 10^{-3}$, $OR = 1.86$) mutations in the TIC (templated insertion chain dominant) cluster, and *KMT2C* ($P = 2.89 \times 10^{-3}$, $OR = 1.95$) mutations in the TRA (translocation-dominated) cluster.

Compared to QUIET cluster, Kaplan-Meier analysis revealed poorer survival among novel SV-class dominated clusters (BR, PYR, and TYF; FDR < 0.1, log rank test) as well as several clusters dominated by previously-described SV classes (CP, CT, and INVD; Figure 3.27). These effects persisted after correcting for clinical and molecular covariates in a Cox regression analysis, with BR ($P = 1.17 \times 10^{-2}$, $HR = 1.72$, likelihood ratio test, Cox regression), PYR ($P = 6.13 \times 10^{-3}$, $HR = 2.01$), TYF ($P = 5.37 \times 10^{-3}$, $HR = 2.12$, CP ($P = 1.76 \times 10^{-3}$; $HR = 1.91$), CT ($P = 6.69 \times 10^{-4}$; $HR = 1.83$), and INVD ($P = 8.79 \times 10^{-4}$; $HR = 2.06$) clusters each demonstrating reduced survival relative to the QUIET cluster (Figure 3.28).

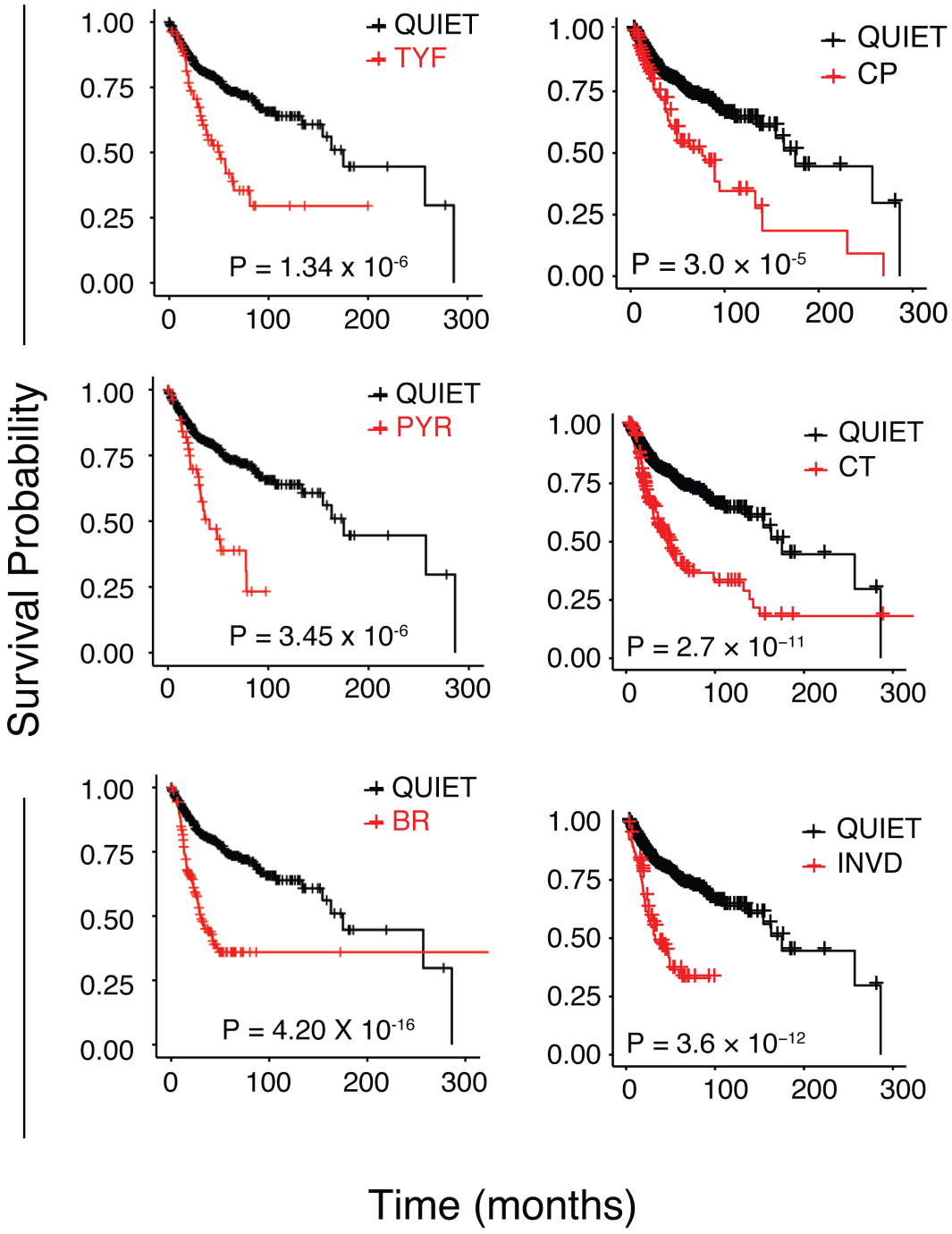

Figure 3.27: **Overall survival worse in 6 clusters than the QUIET.**

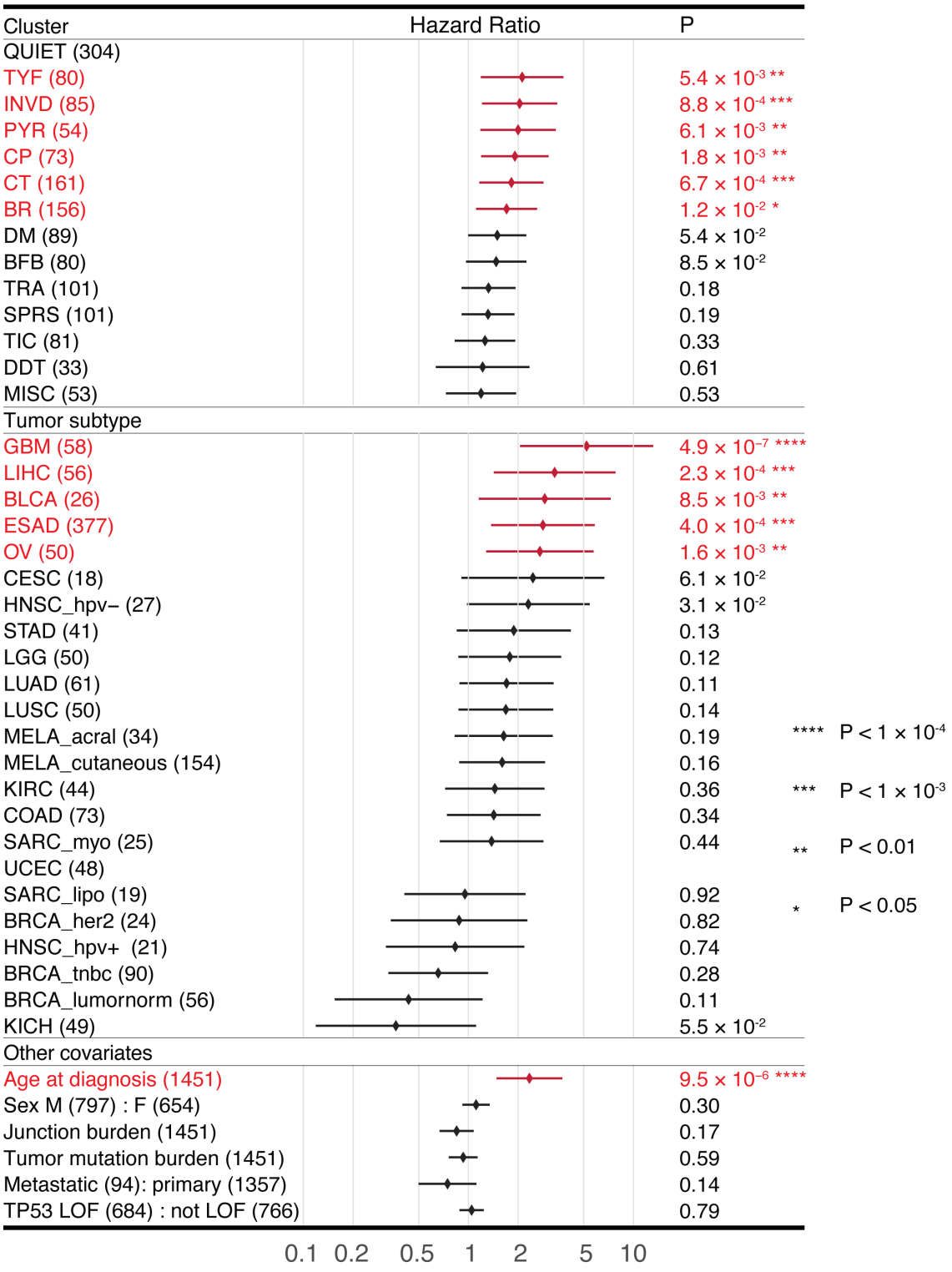

Figure 3.28: **Cox model correcting for covariates show the associations between TYF, INVD, PYR, CP, CT, BR and worse overall survival after correction of related covariates.**

3.4 Discussion

Leveraging JaBbA reconstructed genome graphs from 2778 cancer WGS datasets, we define 13 types of simple and complex classes of structural variants including pyrgo, rigma [83], and tyfonas (see Chapter 3) that have not been described rigorously classified before. These three new event classes all have distinct associations with regional, genotypic, and tumor types from previously known complex patterns like chromothripsis or BFB cycles. They also form clusters differently within the patient cohorts. All of these suggest that they result from novel mutational processes and/or somatic selection. In this chapter, I present two important parts of our discoveries, the taxonomy of complex amplicon and patient clustering based on SV event burdens.

Among amplified subgraphs with high-JCN junctions, tyfonas spontaneously emerge distinct from previously known complex amplification patterns double minutes and BFB cycles cycles by higher total fold-back inversion JCNs and higher elevated JCN counts. Tyfonas further differentiated from DM and BFB cycles in its massive scale, excessive junctions, enrichment in liposarcoma and acral melanoma, and recurrent amplification of *MDM2-CDK4* locus.

What is the mechanism that created tyfonas? We observed tyfonas are present in about 80% of liposarcoma, which coincides with the high frequency of supernumerary rings in that tumor type, indicating they may be orthogonal readouts of the same structure. Furthermore, Garsed et al. [95] proposed *neo-chromosome* structures in five liposarcoma cell lines, which was suggested to arise from circular BFB with centromere eventually regaining telomeres to be stabilized linearly. Based on the similarity of amplification patterns presented in that study,

extensive complex junctions, and the recurrent localization at *MDM2-CDK4* locus same as tyfonas, we conjecture that tyfonas is reflecting the same process. In fact, using fluorescent *in situ* hybridization (FISH) with NCI-H526 small cell lung cancer cell line which possesses a tyfonas event covering both *MYCN* and *MYCL* genes, we found this particular tyfonas is integrated into two chromosomes near termini. This observation confirms the localization of the amplicon in linear form, that is likely neo-chromosomes.

However, our definition of tyfonas differ from the model proposed in [95], partially due to the abundance of fold-back inversions captured in tyfonas. Garsed et al. were propelled to construct the circular BFB theory due to the lack of "inverted duplication" signature of canonical linear form BFB. Yet, we observed many fold-back inversions, and it is very likely that these transient structures gain fold-back inversions in linear forms where the blunt ends can be fused to its newly synthesized 'sister chromatid'. Furthermore, our definition of tyfonas arise spontaneously from analyzing thousands of pan-cancer genome graphs, which indicates such events are not limited to liposarcoma alone, but widely observed in other tumor types like acral melanoma, breast carcinoma, small cell and non-small cell lung cancers.

In theory, there are at least two frequent mechanisms that allow genomic DNA to amplify exponentially, asymmetric inheritance and fold-back on sister chromatid. Double minute and canonical BFB cycles are the prototypical events, respectively. In the former, doubling is conferred when DNA unbounded by centromeres inequally distribute into two daughter cells. Theoretical models of double minute CN dynamics verifies a >1 selection coefficient can quickly result in high mean CN among the cell population [129]. On the other hand, in a

canonical BFB cycle, unprotected chromosome ends fuse and are broken asymmetrically into daughter cells. If new ends stay unprotected, the process can repeat, resulting in focal high-level amplification in the lineage that elongated the arm. We postulate that tyfonas might have taken a combination form of the two.

Different from chromothripsis, initial shattering does not get repaired quickly enough to form scrambled new sequence, rather, the linear fragments are preserved in the nucleus through an extended period of time, for example, in structures like micronuclei. They exist in either linear or circular forms and experience continuous instability which can further 'hijack' distal loci in the genome that was not in the initially shatttered regions, resulting in numerous amplified loci.

Interestingly, tyfonas also have more heterogeneous JCN composition, supporting the asynchronous genesis of junctions during the ongoing amplification process. This is fundamentally different from an 'amplified chromothripsis', where the junctions are supposedly created almost at the same time and amplified together later, which would lead to rather homogeneous elevated JCN within the final amplicon. The events arising from PCAWG's analysis of clustered variants termed 'amplified chromothripsis' [2] actually map to multiple classes under our taxonomy (see [83], Figure S4G-K). This improvement of specificity is also visible in the tumor type enrichments for tyfonas (see [83], Figure S4F).

Despite the proposition of a new and likely impactful amplicon pattern, considerable amplicon candidates still remain unclassified. Some of them are excluded from our current analysis because they do not contain high-JCN junc-

tions (JCN>7). While certain pyrgo instances can achieve this through only 'stacking up' low-JCN tandem duplications, more candidates are weaved by various types of junctions of low to moderate JCN (JCN $\leq$ 7), both with and without fold-back inversions (see Chapter 4, Figure 4.5). We also notice certain double minutes with relatively lower CN than classic examples like *EGFR* amplification in glioblastoma, whose maximum JCN barely passed the current threshold (see Chapter 5, Figure 5.24), yet the homogeneous JCN composition indicates extrachromosomally amplified more than other hypotheses. It is not hard to conjecture then, that there could be other eccDNAs created by the same mechanism as high-level double minutes, yet for various reasons do not get amplified to such extraordinary levels [129].

Even within the high-JCN candidates in our amplicon clustering analysis, there is a fourth cluster with high maximum JCN like double minutes, low fold-back inversion JCN contrary to BFB cycles and tyfonas, yet extensively rearranged with high elevated JCN junction count. The cluster is small and not as stable hence named other and left out of the discussion of this paper [83]. Because of the simultaneous presence of high-JCN DUP-like junctions and abundant lower JCN junctions within the amplified regions, we hypothesize this subset may be reflecting unstable double minutes that keep on accumulating new junctions extrachromosomally. Recently, Shoshani et al. published a study showing cells with eccDNA carrying the further rearranged their amplicon structures in response to continuous selective pressure. The structure and mechanism shown in the study resembles what we observed and is of particular interest to follow up, especially in post-treatment metastatic samples [32].

Looking at the big picture, the burden of our 13 simple to complex SV classes

stratify pan-cancer patients into clear clusters, indicating once again the links to different mutagenesis pathways active during the formation of different tumor samples. It also shows the potential of a genome graph-based SV taxonomy to derive variant count matrix and extract SV signatures eventually. Poorer prognosis in several complex SV infested clusters compared to the one largely free of SVs, reaffirms the understanding that complex SVs can create more aggressive phenotypes. As more studies start to focus on chromosomal instability as a therapeutic target [130], we believe characterization of SV patterns, signatures, and mutagenesis will be key to the clinical application of WGS.

CHAPTER 4

STRUCTURAL VARIANT EVOLUTION AFTER TELOMERE CRISIS *

In this chapter I set out to capture the SV events in human cell lines directly resulting from natural telomere crisis. This chapter is based on our published article [111]*.

Individual contributions: *Sally Dewhurst, Titia de Lange, Marcin Imielinski conceptualized and initiated the project. Sally Dewhurst collected the previous published cell lines spontaneously escaped telomere crisis. Sally Dewhurst designed and created the in vitro telomere crisis model, executed all the experiments, and analyzed the data except for the WGS parts. I carried out all the analysis involving whole genome sequencing data. Marcin Imielinski and I designed multi-sample algorithms for inferring evolving karyotypes. Marcin Imielinski and I led the effort in reconstructing the evolutionary trajectory of the clones.*

4.1 Introduction

Telomeres are the protective sequences at human chromosome ends made up with hexanucleotide repeats 5′-(TTAGGG)-3′ [131]. In healthy cells, telomeres are bound by shelterin complex, and prevent the natural chromosome ends from being recognized as double strand breaks by various DNA damage response (DDR) pathways [132]. Shelterin binds t-loop structures formed by single stranded 3′ overhang at the terminus invading back into upstream double

*Dewhurst, S. M., Yao, X., et al. Structural variant evolution after telomere crisis. Nat. Commun. 12, 2093 (2021)

stranded telomere repeats. At every mitosis, the 3' overhang needs to be recreated, and at the same time DNA polymerase is unable to fully replicate the telomeres, hence telomeres shortens naturally by about 50bps per cell division [132].

In healthy cells, this replicative telomere attrition creates a barrier to their proliferation. When the telomeres are below a certain critical lengths that shelterin complex cannot no longer form a proper protective structure, the cells will undergo a period of time termed *telomere crisis*, during which chromosome ends erroneously trigger DNA repair pathways, fusing telomere to telomere, create dicentric chromosomes [133, 134], and eventually lead to senescence or apoptosis. Cancer cells on the other hand, need to overcome telomere crisis to become immortalized, and this is thought to occur at an early stage of cancer development [135] most commonly through reactivating telomerase, the reverse transcriptase responsible for synthesizing telomere sequences. Thus, the genomic sequences post-crisis cell reflect the alterations introduced during this period and it gives us a chance to deduce what has happened.

What exactly are the consequences of dividing cells when dicentric chromosomes are present? Many events are known with the earliest dated back decades, when McClintock [123] defined breakage-fusion-bridge (BFB) cycles for the broken chromosome fused to sister chromatid during maize meiosis. Gisselsson et al. [136] used cytogenetics and constructed breakpoint profiles for 102 pancreatic carcinoma an 140 osteosarcoma samples, to find some BFBs among extensively variable breakpoint patterns. In other cancer types, the iconic fold-back inversion junctions are found among complex chromothripsis-like patterns in acute lymphocytic leukemia (ALL) [137], triggering the hypoth-

esis these could start with BFB by dicentric chromosomes and are followed by chromothripsis before finally stabilized. Modeling of telomere crisis in late generation telomerase-deficient mice lacking p53 showed that telomere dysfunction engenders cancers with non-reciprocal translocations, as well as focal amplifications and deletions in regions relevant to human cancers [133, 138]. Furthermore, mouse models of telomerase reactivation after a period of telomere dysfunction showed that acquisition of specific copy number aberrations and aneuploidy could drive malignant phenotypes [139].

Lately, high-throughput genome sequencing of cultured human cells started to paint a more detailed picture. Liddiard et al. 2016 [140] examined intra- and inter-chromosomal telomere-to-telomere fusions, and found that a single artificially deprotected telomere can fuse with multiple intra- and inter-chromosomal loci leading to complex fusion products. By expressing a dominant negative allele of *TRF2*, a critical component of shelterin, multiple studies have shown that the resolution of the resulting dicentric chromosomes can lead to dramatic chromosome shattering known as chromothripsis [141, 28, 142]. However, all the studies mentioned above either used systems lacking crucial genome maintenance components (e.g. *TRF2*, *LIG3/4*), or are confounded by the selection process happened naturally (in cancer) or during culture.

The only direct observation of SV consequences in relatively healthy cells after a sustained period of telomere crisis did not capture major complex events implicated before [143]. This is probably due to the fact that the HCT116 colon carcinoma cells used in this experiment have already expressed *TERT* and are able to readily escape from telomere crisis. Again, in the same study, there are various complex SV events reported for engineered HCT116 cells deficient

in non-homologous end joining (NHEJ), but the results cannot be simply attributed to telomere crisis.

Given the expanding repertoire of structural variant present in so many cancer types, and the potential contribution of telomere dysfunction to some of these aberrations, we set out to characterize the extent and type of structural variant that can be unleashed by telomere crisis and subsequent genome stabilization by telomerase expression. We tackle this problem from two aspects. First, we conducted WGS on nine isolated cell lines previously found to survive telomere crisis by spontaneous expression of telomerase, both in the post-crisis cells and in their pre-crisis ancestors as control. If we observe SV events specific to the post-crisis cell lines and not pre-crisis, those can be largely attributed to telomere crisis.

Second, considering the complications during spontaneous *TERT* activation, we set out to design a system known to experience natural replicative telomere attrition, activates *TERT* on cue, and allow us to harvest cells precisely after they escaped telomere crisis. We use low-pass (low read depth) WGS to screen for CNAs and further subject representative clones taking different CNA configurations to high-pass WGS to reconstruct the exact rearranged haplotypes after telomere crisis by genome graph analysis. This will enable us to infer phylogeny among diverging clones, deduce the sequence of events triggered by crisis, and gain much deeper insight into how they start, progress, and conclude. In short, through both approaches, we observe a wide spectrum of SV events, from totally quiet genomes to simple arm-level gain and loss to complex chromothripsis or BFB patterns. Hence, telomere crisis can instigate varying SV events, and

using BFB and chromothripsis alone cannot accurately predict the presence of telomere crisis in the life history of a cancer.

4.2 Genomic complexity after spontaneous telomerase activation

In order to determine the SVs in post-telomere crisis genomes, we examined nine SV40 large T-transformed cell lines that had undergone spontaneous telomerase activation after passage into telomere crisis (Figure 4.1). In Figure 4.1, binned purity- and ploidy-transformed read depth is shown in the periphery, with colored links in the center representing variant (rearrangement) junctions. A series of red colors is used to show junctions and read-depth bins belonging to distinct clusters of complex gains in each cell line. Additional colors describe junctions and bins, including those belonging to simple losses and gains.

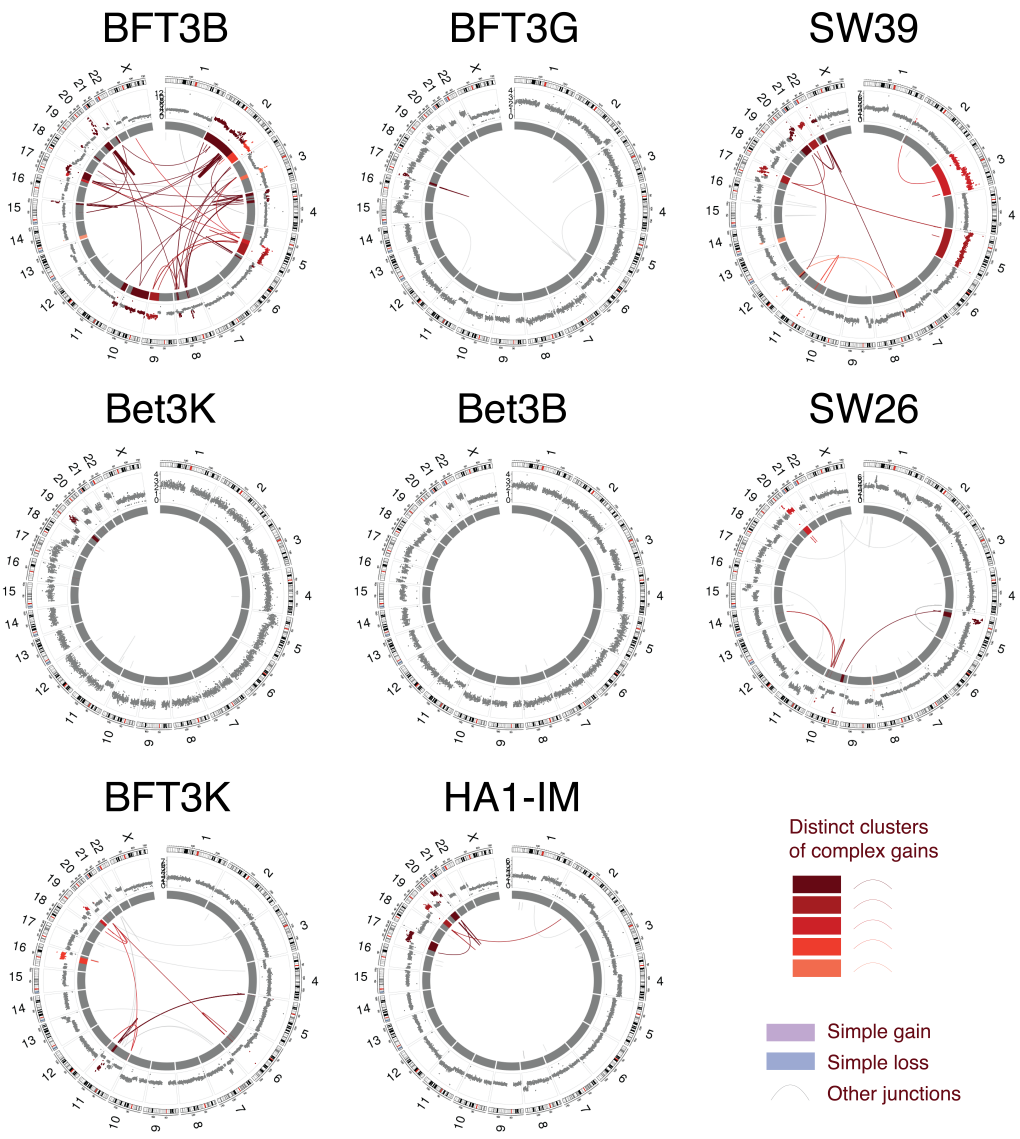

Figure 4.1: **Post-crisis SV landscape in 8 spontaneously derived cell lines.**

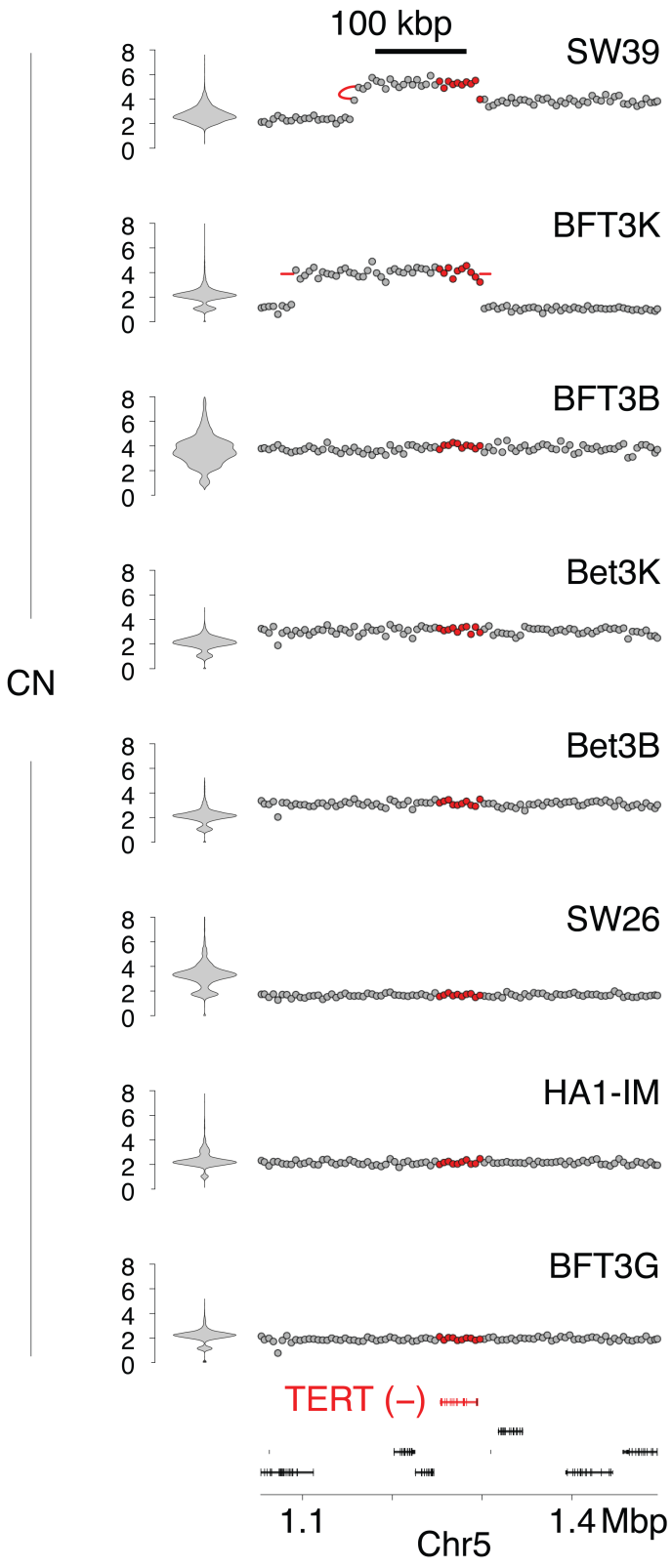

Figure 4.2: **SVs around *TERT* locus in spontaneous post-crisis cell lines.**

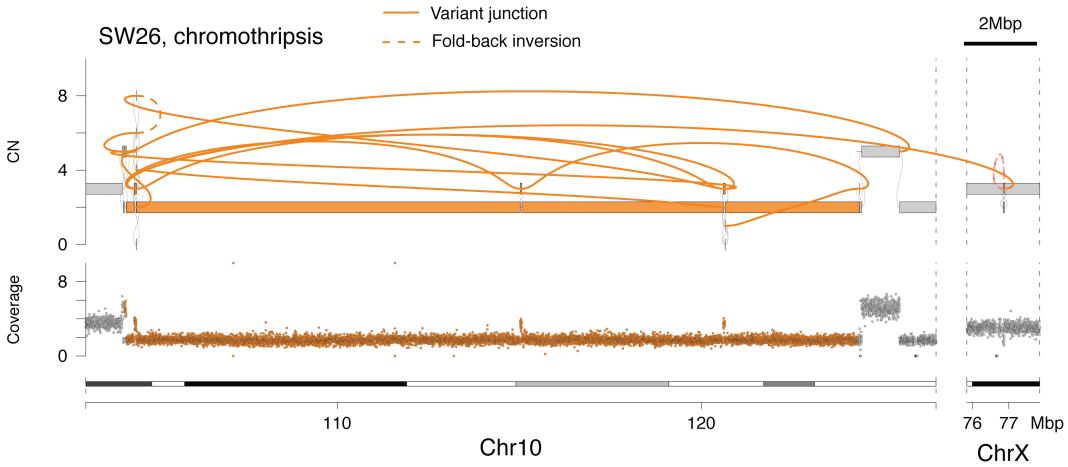

Figure 4.3: **Chromothripsis event in SW26 cell line.**

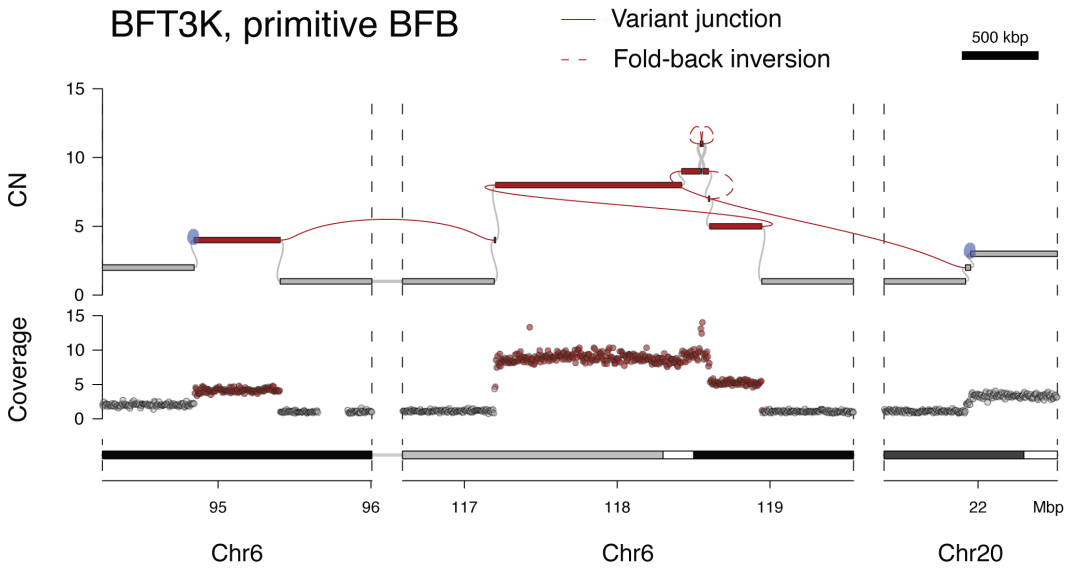

Figure 4.4: **Primitive BFB event in BFT3K cell line.**

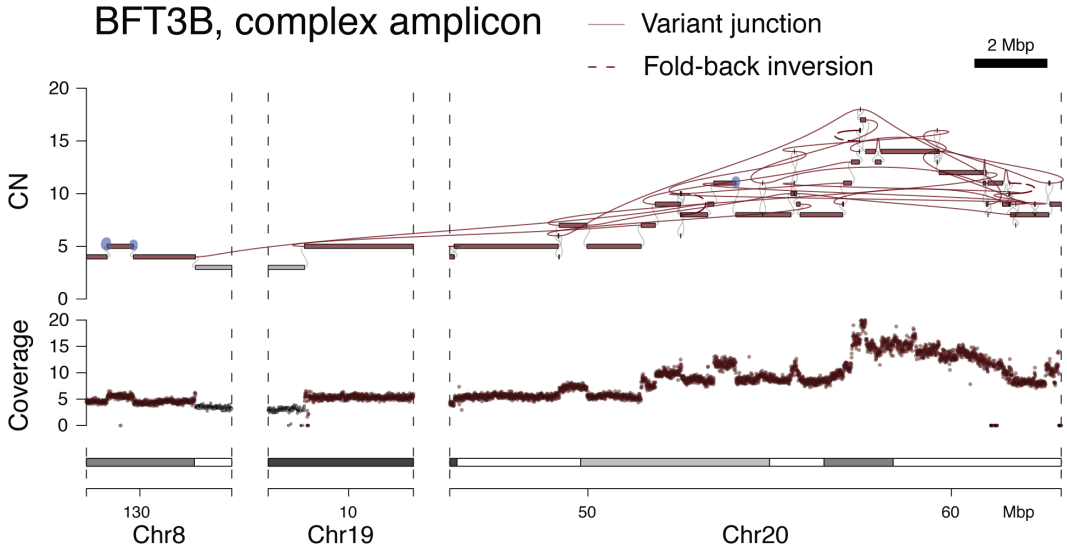

Figure 4.5: **Complex amplification event in BFT3K cell line.**

Table 4.1: **SV40T immortalized pre-crisis and post-crisis telomerase positive cell lines.**

Name	Cell Line/Tissue Source	Telomerase Status	Reference	Kind gift of
HA1 p15	Human Embryonic Kidney cells	-	[144]	Silvia Bacchetti/AdVec
SW13 PD73	IMR90 lung fibroblasts	-	[145]	Jerry Shay, UTSW
SW26 PD68.5	IMR90 lung fibroblasts	-	[145]	Jerry Shay, UTSW
SW39 PD72.5	IMR90 lung fibroblasts	-	[145]	Jerry Shay, UTSW
Bet3B p3	NHBE-10 bronchial epithelial cells	-	[146]	Roger Reddel, CMRI
Bet3K p7	NHBE-10 bronchial epithelial cells	-	[146]	Roger Reddel, CMRI
BFT3B p10	BF-10 bronchial fibroblasts	-	[146]	Roger Reddel, CMRI
BFT3G p7	BF-10 bronchial fibroblasts	-	[146]	Roger Reddel, CMRI
BFT3K p9	BF-10 bronchial fibroblasts	-	[146]	Roger Reddel, CMRI
HA1-IM PD216	Human Embryonic Kidney cells	+	[144]	Silvia Bacchetti/AdVec
SW13 PD184	IMR90 lung fibroblasts	+	[145]	Jerry Shay, UTSW
SW26 PD130+	IMR90 lung fibroblasts	+	[145]	Jerry Shay, UTSW
SW39 PD130+	IMR90 lung fibroblasts	+	[145]	Jerry Shay, UTSW
Bet3B p25 post-crisis	NHBE-10 bronchial epithelial cells	+	[146]	Roger Reddel, CMRI
Bet3K p25 post-crisis	NHBE-10 bronchial epithelial cells	+	[146]	Roger Reddel, CMRI
BFT3B p28 post-crisis	BF-10 bronchial fibroblasts	+	[146]	Roger Reddel, CMRI
BFT3G p28 post-crisis	BF-10 bronchial fibroblasts	+	[146]	Roger Reddel, CMRI
BFT3K p34 post-crisis	BF-10 bronchial fibroblasts	+	[146]	Roger Reddel, CMRI

These cell lines represent independent immortalization events in a variety of cell lineages [146, 145, 144]. We carried out whole genome sequencing of these nine post-crisis cell lines and their pre-crisis counterparts to a median depth of 40X (range: 15-51) and generated junction-balanced genome graphs[83] via JaBbA from SvABA[97] and GRIDSS[147] junction calls (see Appendix 6.5). We then apply JaBbA algorithm (Chapter 2) to reconstruct an optimal genome graph for each cell line and identify the patterns of structural variants (Chpater 3).

Comparison of ancestral (pre-crisis) and derived (post-crisis) genome graphs showed that eight of nine post-crisis cell lines acquired virtually all (61.9% - 100%, median 96.6%) of their observed Structural variant during or after crisis (Figure 4.1 and Supplementary Figure 1b in [111]). One cell line (SW13) had acquired significant aneuploidy and genome rearrangement prior to crisis and was therefore difficult to interpret (Supplementary Figure 1b in [111]). The other eight post-crisis genome graphs demonstrated varying levels of aneuploidy (ploidy ranges: 1.9-3.4) with variable numbers of clonal junctions per genome (range: 5-115, median 25). Analysis of junction-balanced genome graphs [83] revealed complex multi-chromosomal gains in six samples, with the other two lines harboring only broad arm level losses or gains (Figure 4.1).

Strikingly, besides one instance of chromothripsis (Figure 4.3), genome graph-based categorization of complex SVs [83] identified few classic footprints of chromothripsis or BFB cycles in these genomes. However, several amplified subgraphs were associated with step-wise copy number gains reminiscent of BFB cycles (Figure 4.4). The majority of copy changes in these subgraphs could not be attributed to fold-back inversion junctions (a hallmark of BFB cycles) but were instead driven by a spectrum of duplication and translocation-like junctions and templated insertion chains. These patterns are exemplified in a 10 Mbp region of 20q of post-crisis cell line BFT3B that is amplified to 10-15 copies, incorporating Mbp scale fragments from 11 other chromosomes at lower copy number including chromosome 8 and 19 (Figure 4.5). Of note, five of eight cell lines showed modest increases in *TERT* copy number, providing a possible genomic basis for escape from telomere crisis (Figure 4.2). In Figure 4.2, each track shows binned purity- and ploidy-transformed read depth in units of CN, with variant (rearrangement) junctions and loose ends plotted as red arcs. The bottom track highlights the TERT gene among genes on chromosome 5p15. Violin plots to the left of tracks show the genome-wide distribution of read depth in units of CN, demonstrating that 6 of 7 clones have elevated CN at the TERT locus.

In summary, across the eight post-crisis cell lines, spontaneous escape from crisis was associated with a highly variable spectrum of SV patterns, ranging from relatively unaltered genomes to complex non-canonical patterns of amplification as well as numerical gains and losses. Importantly, BFB-like patterns and chromothripsis were not a general feature of the post-crisis genomes.

4.3 An in vitro system for telomerase-mediated escape from natural telomere crisis

Despite the conclusions that both simple and complex SVs can appear in genomes after spontaneous escapes from telomere crisis, it still remains unclear what SVs are the direct result of telomere crisis. These existing cell lines spent various time in crisis, represent limited lineages that emerged from crisis, probably had undergone selection 4.2. To that end, Sally Dewhurst designed a new *in vitro* model of telomere crisis using MRC5 human lung fibroblasts, which is known to reach mitotic barrier after about 50 generations. The Rb and p21 pathways were silenced with shRNAs to bypass senescence (data not included, refer to [111] Supplementary Figure 2A). Meanwhile, an inducible CRISPR activation system (iCRISPRa) is introduced to activate *TERT* transcription (data not included, refer to [111] Supplementary Figure 2B). This system couples nuclease-dead Cas9 to a tripartite transcriptional activator and is guided by four gRNAs binding to the *TERT* promoter region. Upon addition of doxycycline (dox), these MRC5/Rbsh/p21sh/iCRISPRa-TERT cells acutely expressed *TERT* mRNA within 96 hours, in contrast to untreated cells in which *TERT* transcripts were undetectable (data not included, refer to [111] Figure 2B). Sally further verified that the telomerase activity (TRAP assay, telomerase repeated amplification protocol; data not included, refer to [111] Figure 2C) are detectable, but both the expression and activity are lower than the telomerase-positive control cell line HCT116. This is a desired feature of this system as it has been shown that common *TERT* promoter mutations found in tumor cells only sustain a low telomerase level and are not able to restore bulk telomere lengths [148].

In about 120 days after the initial cultutre (55 days with dox treatment), the MRC5/Rbsh/p21sh/iCRISPRa-TERT cells emerged more proliferative than the untreated group (Figure 4.6). Using Single Telomere Length Analysis (STELA, [149]) at 150 days, it is observed that the activated telomerase expression can sustain proliferation yet is unable to restore bulk (average) telomere length (data not included, refer to [111] Figure 2E). However, fewer critically short telomeres can be seen in the telomerase-activated group (data not included, refer to [111] Figure 2E-F). This phenomenon is consistent with the hypothesis that in the cell population without telomerase, cells with critically short telomeres shift towards decease, resulting in the generally longer telomeres in the remaining cells. On the contrary, with telomerase, cells gain the ability to salvage the critically short telomeres and survive longer with on average shorter telomeres.

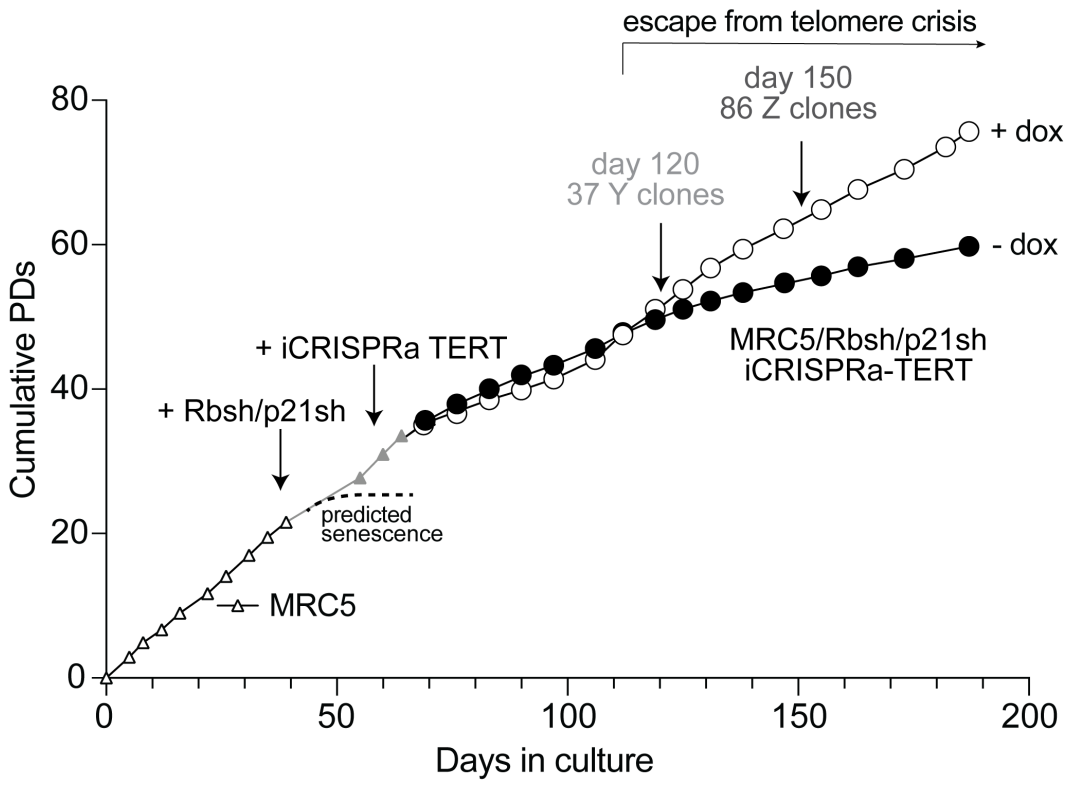

Figure 4.6: **Clones are isolated at day 120 and day 150 for whole genome sequencing.**

Arrows indicate when each construct was introduced. Days in culture represent total time in culture from parental MRC5 cells to late passage MRC5/Rbsh/p21sh/iCRISPRa-TERT cells. Time points for telomere analysis and the approximate onset of senescence in the parental MRC5 cells are indicated.

4.4 Genomic screening of post-crisis clones

To assess the genome structure of proliferating post-crisis cells, single cell clones were isolated from induced MRC5/Rbsh/p21sh/iCRISPRa-TERT cells at day 120 ('Y clones') and day 150 ('Z clones') (Figure 4.6). The clonal yield at day 150 was greater than at day 120 in induced cells but no clones could be isolated from the uninduced population at either time point. The lower clonal yield at day 120

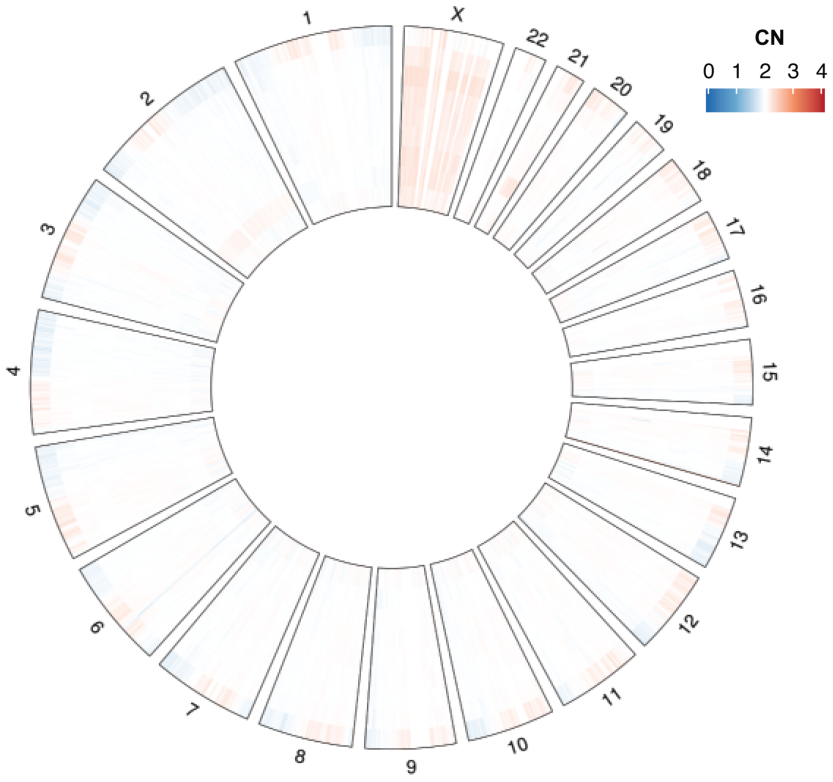

Figure 4.7: **Control group with exogenous *TERT* show no CNA.**

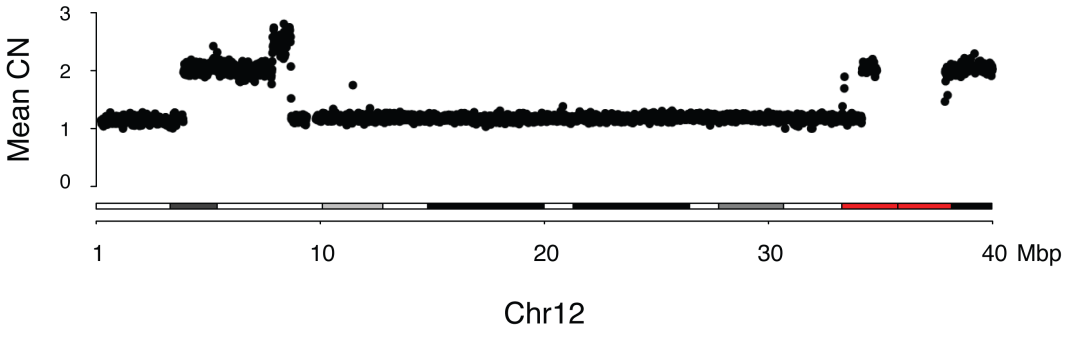

Figure 4.8: **Aggregated coverage of the putative BFB clones.**

Table 4.2: **Number of clones analyzed by high and low-pass WGS.**

	Number Sequenced	
Name	Low Pass	High Pass
Parental cell lines	3	1
Day 120 (Y) clones	37	5
Day 150 (Z) clones	83	8
Control (CT) clones	8	0

may be due to incomplete stabilization of the telomeres since clones from this time-point showed a higher burden of fused telomeres than those derived from day 150 (data not included, refer to [111] Supplementary Figure 4A). Post-crisis clones from both time points showed evidence of ultra-short telomeres and reduced telomere length (data not included, refer to [111] Supplementary Figure 4B-C). Telomerase activity in post-crisis clones was comparable to the parental induced population, indicating that clone viability was not due to selection for increased telomerase activity (data not included, refer to [111] Supplementary Figure 4D). To generate control clones which had not passed through a period of telomere crisis, early passage MRC5 cells were infected with a retrovirus expressing hTERT and single cell clones were isolated (data not included, refer to [111] Supplementary Figure 4E). Genome profiling with low pass (5X) WGS was performed on eight hTERT-expressing control clones (CT clones), 36 Y clones from day 120, and 82 Z clones from day 150 (Table 4.3).

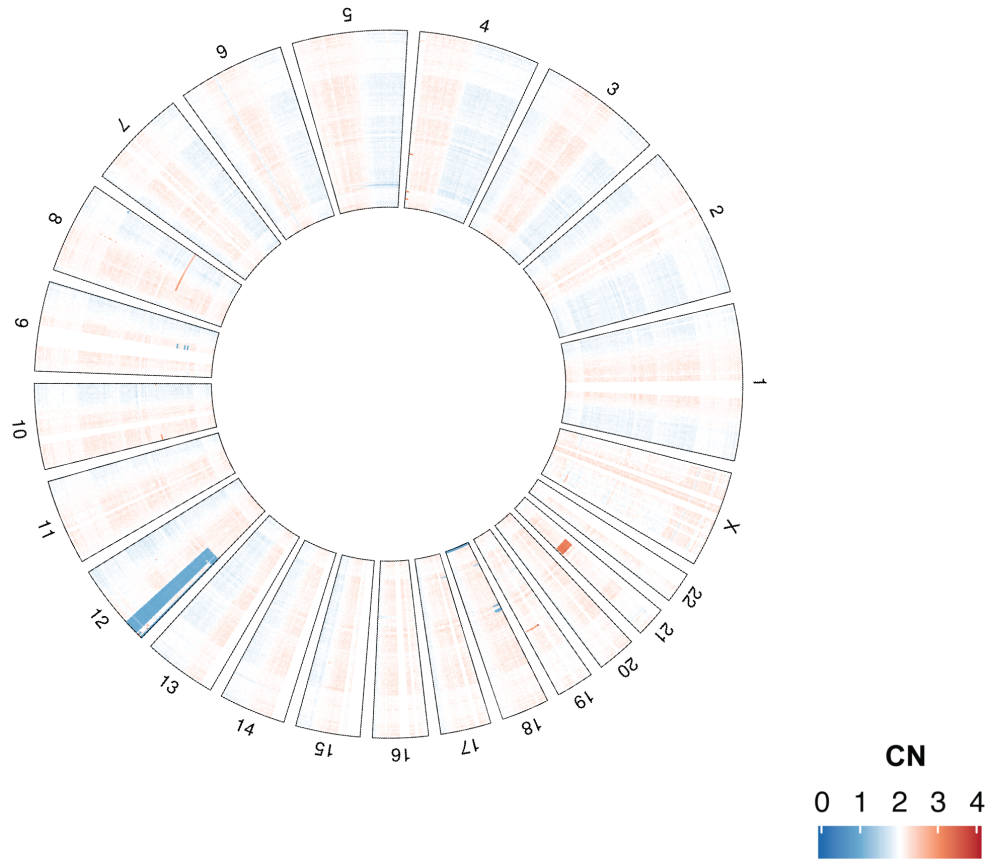

Figure 4.9: **Circular heatmap showing genome-wide binned purity- and ploidy-transformed read depth.**

Units are CN across 118 low-pass WGS-profiled clones. Heatmap rows correspond to concentric rings in the heatmap. Clones are clustered with respect to genome-wide copy number profile similarity (see "Methods").

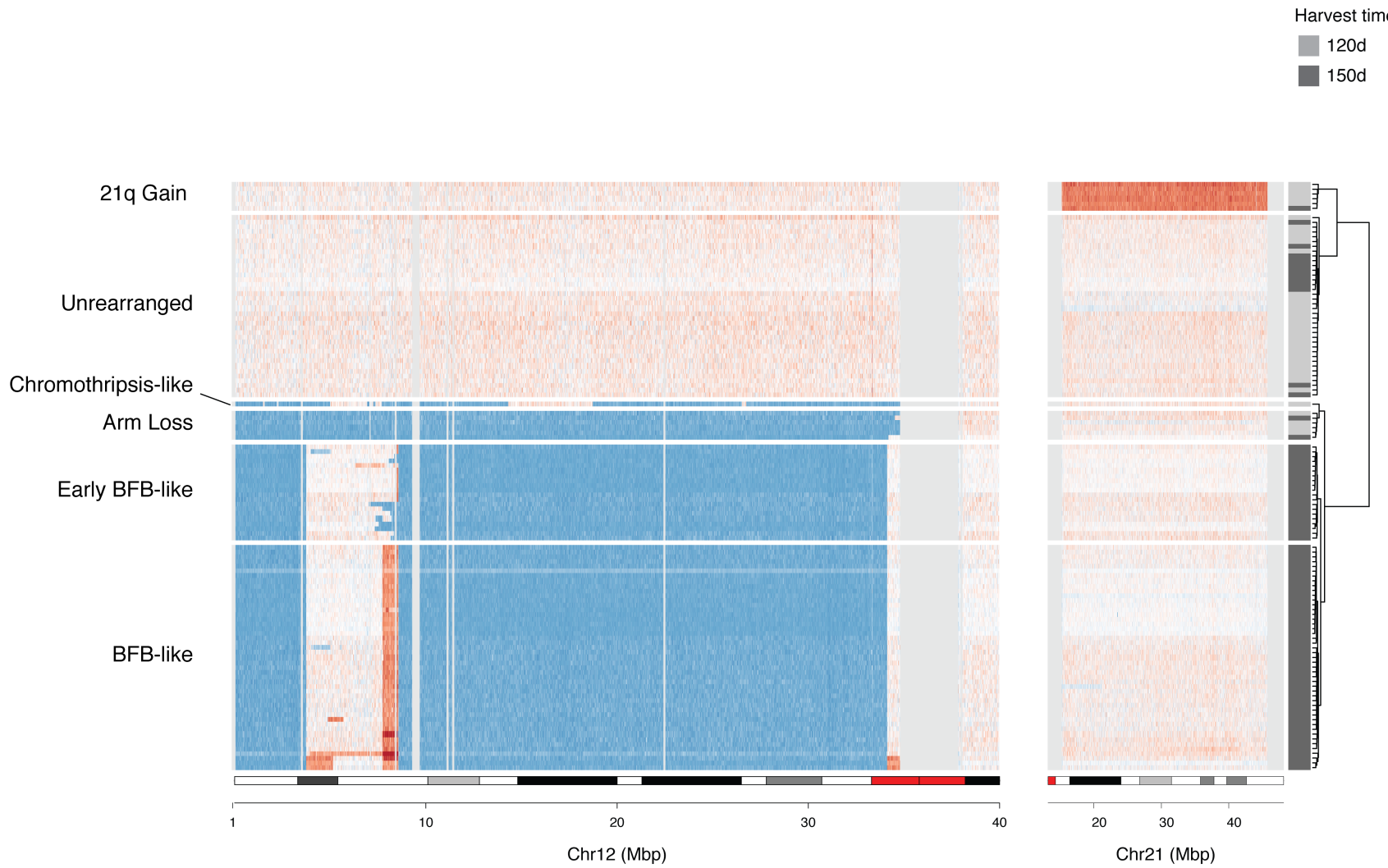

Figure 4.10: **Post-crisis clones are clustered into six groups based on CN profile of chromosome 12 and 21.**

Analysis of genome-wide read depth across 118 clones from both day 120 (Y clones) and day 150 (Z clones) demonstrated predominantly diploid genomes with a striking enrichment of clones with DNA loss on most of chromosome 12p (63%, 74/118, Figure 4.9). Within the other 44 samples, we observed a subset of clones (5%, 6/118, Figure 4.9) with gains of chromosome 21q. As expected, control CT clones showed no evidence of SVs or copy number variants (Figure 4.7). Hierarchical clustering of all clones by their coverage on chromosomes 12p and 21q revealed six distinct clusters (Figure 4.10). A minority of clones were diploid on chromosomes 12 and 21 and elsewhere in the genome and are therefore designated as 'unrearranged' (32% of clones, 38/118). Of note, the unrearranged group was enriched in day 120 (Y) samples compared to day 150 (Z) samples ($p = 1.79 \times 10^{-9}$, odds ratio 14.7, Fisher's exact test; Figure 4.10), suggesting that these clones may have largely avoided crisis prior to telomerase induction. The cluster of clones with 21q gain were diploid on 12p.

The remaining 74 clones (63%) all showed a heterogeneous pattern of copy number alterations targeting 12p (Figure 4.10). One out of the 118 clones (0.8%) displayed the singular pattern of distinct interspersed losses that resembled chromothripsis. Complete loss of one copy of 12p ('arm loss') was found in a cluster of 6 clones (6/118, 5%). A second cluster of 67 clones all shared a breakpoint near the distal end of 12p and a large deletion starting 9 Mbp from the centromere. These clones were differentiated into two clusters by the presence or absence of an amplification around 8-9 Mbp from the 12p telomere. In the 47 clones that contained this amplification, aggregated consensus read depth profiles revealed step-wise gains at the distal end of 12p, a pattern reminiscent of BFB cycles (Supplementary Figure 4.8). This cluster was therefore labeled 'BFB-like', a designation which is further supported by data presented below.

The 20 clones (17%) that lack the amplicon around 8-9 Mbp harbored varying boundaries of the shared larger deletion; based on the analysis described below we designate these as 'early BFB-like'. In summary, these low-pass WGS copy number profiles indicated a limited set of distinct lineages surviving telomere crisis, with at least two lineages independently converging on 12p.

4.5 Joint inference of junction balance in MRC5

To chart structural variant evolution across sub-clades of MRC5 clones, a procedure was developed to jointly infer junction balanced genome graphs in a lineage (e.g. BFB lineage in Figure 4.11). This co-calling algorithm augmented the existing JaBbA model, described in detail in [83], enabling application to a compendium of genome graphs by minimizing the total number of unique loose ends assigned a nonzero copy number across the graph compendium.

Formally, we define a collection $G^i = (V^i, E^i), i \in 1, \ldots, n$ of identical genome graphs across n clones, each a replica of a *prototype* genome graph $G^0 = (V^0, E^0)$. The mapping $p : V \cup E \rightarrow V^0 \cup E^0$ maps each vertex $v \in V^i$ and edge $e \in E^i, i \in 1, \ldots, n$ to its corresponding vertex $p(v) \in V^0$ and edge $p(e) \in E^0$ in the prototype graph. We then jointly infer unique copy number assignments κ^i to the vertices and edges of each genome graph G^i by solving the mixed integer program:

$$\underset{\kappa^i : V_I(G^i) \cup E(G^i) \to \mathbb{N}, i \in 1,\ldots,n}{\text{minimize}} \lambda \mathcal{R}(\{G^i\}_{i\in 0,\ldots,n}, \{\kappa^i\}_{i\in 1,\ldots,n}, p) + \sum_{i\in 1,\ldots,n} \mathcal{V}(G^i, \kappa^i, x^i, J^i)$$

$$\begin{aligned} \text{subject to: } & \kappa^i(v) = \kappa^i(\bar{v}), \forall v \in V_I^i \\ & \kappa^i(e) = \kappa^i(\bar{e}), \forall e \in E^i \\ & \kappa^i(v) = \sum_{e\in E^+(v,G^i)} \kappa^i(e) = \sum_{e\in E^-(v,G^i)} \kappa^i(e) \\ & \kappa^i(e) \leq u^i(e) \end{aligned} \tag{4.1}$$

where x^i and J^i represent the binned read depth data and bin-node mappings for clone i and $\mathcal{V}(G^i, k^i, x^i, J^i)$ is the read depth residual for genome graph i, analogous to 2.11. A new term in this joint formulation is $u^i : E(G^i) \to \{0, \inf\}$, which is a constraint on the upper bound of each edge CN $\kappa(e) \in E(G^i)$, e.g. zero if there have been no read support for the junction in clone i.

The coupling of multiple graphs happen through a joint complexity penalty $\mathcal{R}$, which only penalize a loose end once for all the clones. Formally,

$$\mathcal{R}(\{G^i\}_{i\in 0,\ldots,n}, \{\kappa^i\}_{i\in 1,\ldots,n}, p) = \sum_{e_L^0 \in E_L(G^0)} [\![\sum_{\hat{e_L} | \hat{e_L} \in E_L^i, p(\hat{e_L}) = e_L^0, i\in 1,\ldots,n} \kappa^i(\hat{e_L})]\!] \tag{4.2}$$

As in 2.11, the hyperparameter λ in Equation 4.1 controls the relative contribution of the read-depth residual and complexity penalty to the objective function.

It is important to note that while each of the graphs G^i have an identical structure, the constraints imposed by the upper bounds u^i and bin profiles x^i confine each graph to its junctions and read depth data, and hence lead to a unique fit k^i on the basis of this data. The L_0 penalty (defined using the Iverson bracket $[\![]\!]$ operator) in Equation 4.2 couples the solutions k^i by adding an exponential prior on the number of unique loose ends across the entire graph compendium, where uniqueness is defined by the mapping p to the prototype graph G^0.

This joint mixed-integer programming model in Equation 4.1 is implemented in the "balance" function of gGnome. The model was applied to a collection of genome graphs representing the structure of chromosome 12 across 13 clones. The prototype graph for this genome graph collection was built from the disjoint union of intervals of the 13 preliminary graphs (via the GenomicRanges "disjoin" function) and the union of junction calls fit across those graphs (via gGnome "merge.Junction" function). Each graph was associated using the read depth data and bin-to-node mappings as per [83]. The mapping u^i for each reference edge was set to inf while variant edges were assigned inf if bwa mem realignments of read pairs in each clone .bam file to the corresponding junction contig via rSeqLib (https://github.com/mskilab/rSeqLib, [150]), otherwise they were assigned 0.

Equation 4.1 was then solved using the IBM CPLEX (v12.6.2) MIQP optimizer within the gGnome package after setting the hyperparameter λ to 100. This value was chosen after a parameter sweep observing for the visual concordance of genome graphs, loose ends, and read depth profiles in the region.

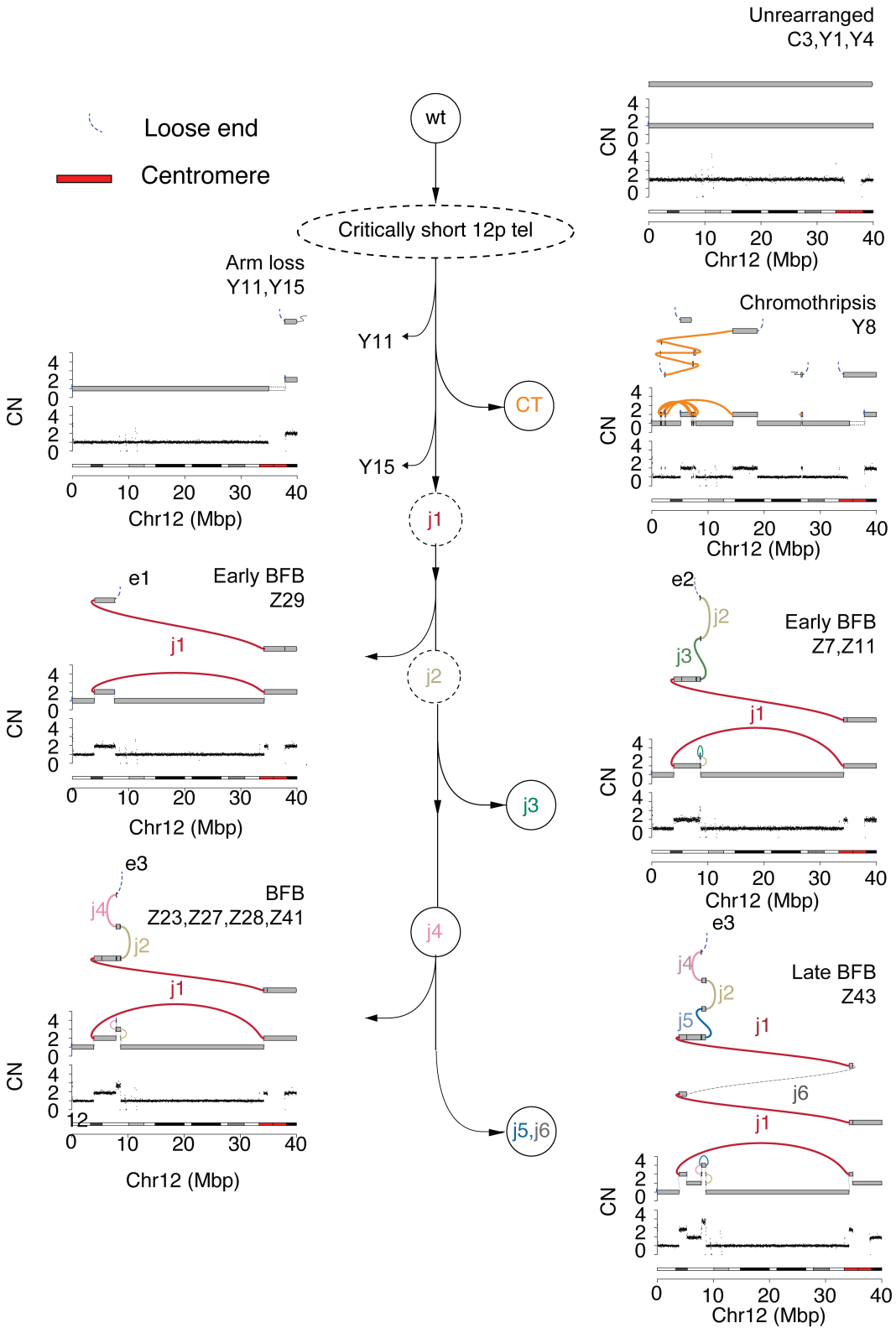

Figure 4.11: **Evolutionary trajectory of SVs in clones with high-pass WGS.**

4.6 High-resolution reconstruction and lineage of post-crisis genomes

To gain further insight into structural variant evolution along these lineages, we chose 15 representative clones spanning the 5 clusters with rearrangements involving 12p for high-depth WGS to a median read depth of 50X (range: 30-88). Phylogeny derived from genome-wide SNV patterns demonstrated a median branch length of 551 SNVs (range: 9-2,409), a low mutation density (<1 SNV/Mbp) that is consistent with previous WGS studies of clones in cell culture [151]. This analysis revealed four major clades (Figure 4.12). These clades had good concordance with copy number alteration and rearrangement junction patterns in the same 12p region, suggesting these clones represent distinct post-crisis evolutionary lineages (Figure 4.12).

In order to further reconcile the shared and distinct rearrangement junctions present in the evolution of these clones, we carried out local assembly of rearrangement junctions and junction balance analysis (see 4.7) which revealed 7 distinct junction-balanced genome graphs spanning 12p (Figure 4.11). With the exception of the chromothriptic lineage (see below), each of these distinct lineages was represented by more than one post-crisis clone.

To reconstruct a set of linear alleles that parsimoniously explain these different genome graph patterns3 (Figure 4.11), we applied gGnome to the data (see Section 4.6). We constrained our model to contain one intact allele of chromosome 12 for the following reasons: 1) karyotypes and chromosome painting showed a single copy of chromosome 12 was altered in the post-crisis clones (see Figure 6b-c of [111]); and 2) rearrangement of one allele is more likely than

rearrangement across two alleles. Application of this constraint to the full set of MRC5 clones in a joint inference revealed a parsimonious set of rearranged alleles that explained the observed collection of clonally-related junction-balanced graphs (Figure 4.11).

4.7 Joint reconstruction of allelic evolution in MRC5

Evolving 12p alleles were jointly reconstructed across 13 MRC5 clones through the decomposition of junction balanced genome graphs (G^i, k^i) (see Section 4.5 section above). The procedure for joint allelic phasing described in [83] and Eq. 2.4 was extended to identify the most parsimonious (least unique walks needed) collection of paths and/or cycles H and associated walk multiplicity $\kappa(H)$ that summed to the vertex and edge copy numbers in the compendium (G^i, k^i).

Formally, the subgraph of vertices and edges with a nonzero copy number in each (G^i, k^i) were exhaustively traversed to derive all minimal paths and cycles H^i, where for each walk $h \in H^i$ maps to incident sequences of $V(h) \subseteq V^i$ and $E(h) \subseteq E^i$ of vertices and edges in the graph G^i. The nodes and vertices of these walks were then projected to via the mapping p to define a unique set of walks H^0 in the prototype graph G^0. We extend our notation p (see previous section) so that for a walk $h \in H^i$ the mapping $p(h) \in H^0$ denotes the walk formed by projecting the vertices and edges of h via p to H^0. With these definitions, the single graph haplotype inference defined in 2.4 was extended to a joint inference by solving the following mixed integer linear program to assign a copy number $\kappa^i(h) \in \mathbb{N}$ to each walk $h \in H^i$:

$$\underset{\kappa^i(h):H^i\to\mathbb{N},i\in 1,\ldots,n}{\text{minimize}} \sum_{h\in H^0} [\![\sum_{\hat{h}\in H^i | p(\hat{h}=h)} \kappa^i(\hat{h})]\!]$$
$$\text{s.t. } \kappa^i(v) = \sum_{\hat{h}\in H^i} \kappa(\hat{h})\delta(\hat{h}, v), \forall v \in V^i, i \in 1,\ldots,n \tag{4.3}$$
$$\kappa^i(e) = \sum_{\hat{h}\in H^i} \kappa(\hat{h})\delta(\hat{h}, e), \forall e \in E^i, i \in 1,\ldots,n$$

where the function $\delta(v, h)$ and $\delta(e, h)$ is 1 if vertex v and edge e belong to walk h and 0 otherwise, just like in 2.4. The Iverson bracket ($[\![]\!]$) operator in the objective function Equation 4.3 minimizes the total number of unique walks used across the compendium, hence identifying a jointly parsimonious assignment of copy number to walks across the compendium of graphs. Equation 4.3 was solved using the IBM CPLEX (v12.6.2) MIQP optimizer within the gGnome package. Variant cycles and paths from the resulting solution were manually combined to yield a set of consistent linear paths, i.e. somatic haplotypes, to yield allelic reconstructions in Figure 4.11.

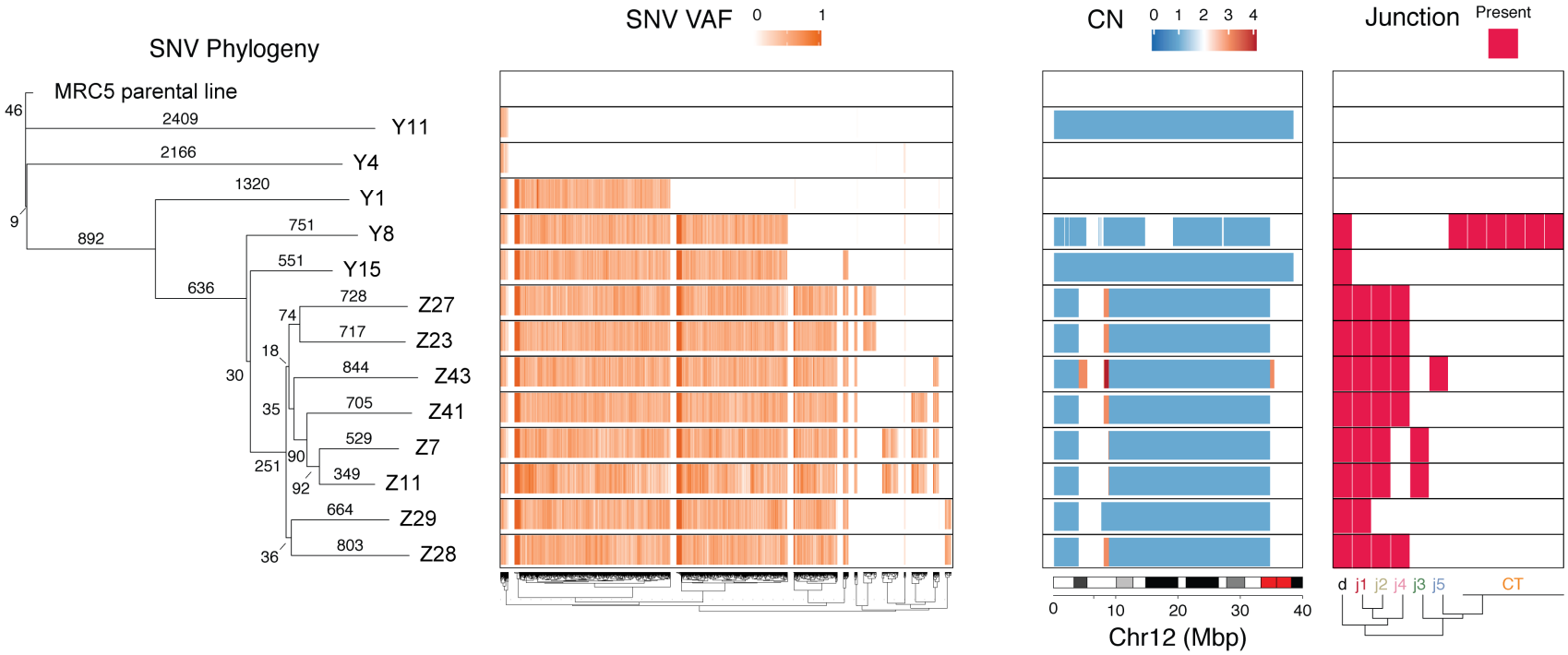

Figure 4.12: **Phylogeny reconstructed based on SNV VAFs, CNA segmentation, and the presence of junctions.**

4.8 Evolutionary trajectory of a post-crisis chromosome 12

Analyzing the clonal evolution of these rearranged 12p alleles, we identified 8 clones demonstrating progressive stages of a BFB cycle. This complex variant evolved after a long-range inversion junction (j1) joined a distal end of 12p to its peri-centromere. This junction was followed by subsequent fold-back inversion junctions (j2, j3, j4), clustered at the 8-9 Mbp focus on 12p, which are present in two different sets of post-crisis clones (Early BFB, BFB, Figure 4.11). The earliest of the fold-back inversion junctions (j2) in the BFB lineage was associated with a cluster of 3 G or C mutations within 2 kbp of each other, consistent with APOBEC-mediated mutagenesis [142] (Figure 4.13). The most complex locus in the BFB lineage (Z43, Late BFB, Figure 4.11), contained six variant junctions in *cis*, including two late DUP-like junctions (j5, j6). Although j6, which connects the distal portion of 12p to the 12p centromere, was not directly observed in the short read WGS data, it was imputed (dashed line, j6, Figure 4.11) to resolve the duplication of j1 in clone Z43 (JCN=2), as well as two loose ends in the genome graph. Remarkably, the vast majority (97%) of SNVs detected in this BFB lineage (Figure 4.12) were either shared by all clones or private to a single clone, indicating that these stages of BFB evolution occurred rapidly in the history of the experiment.

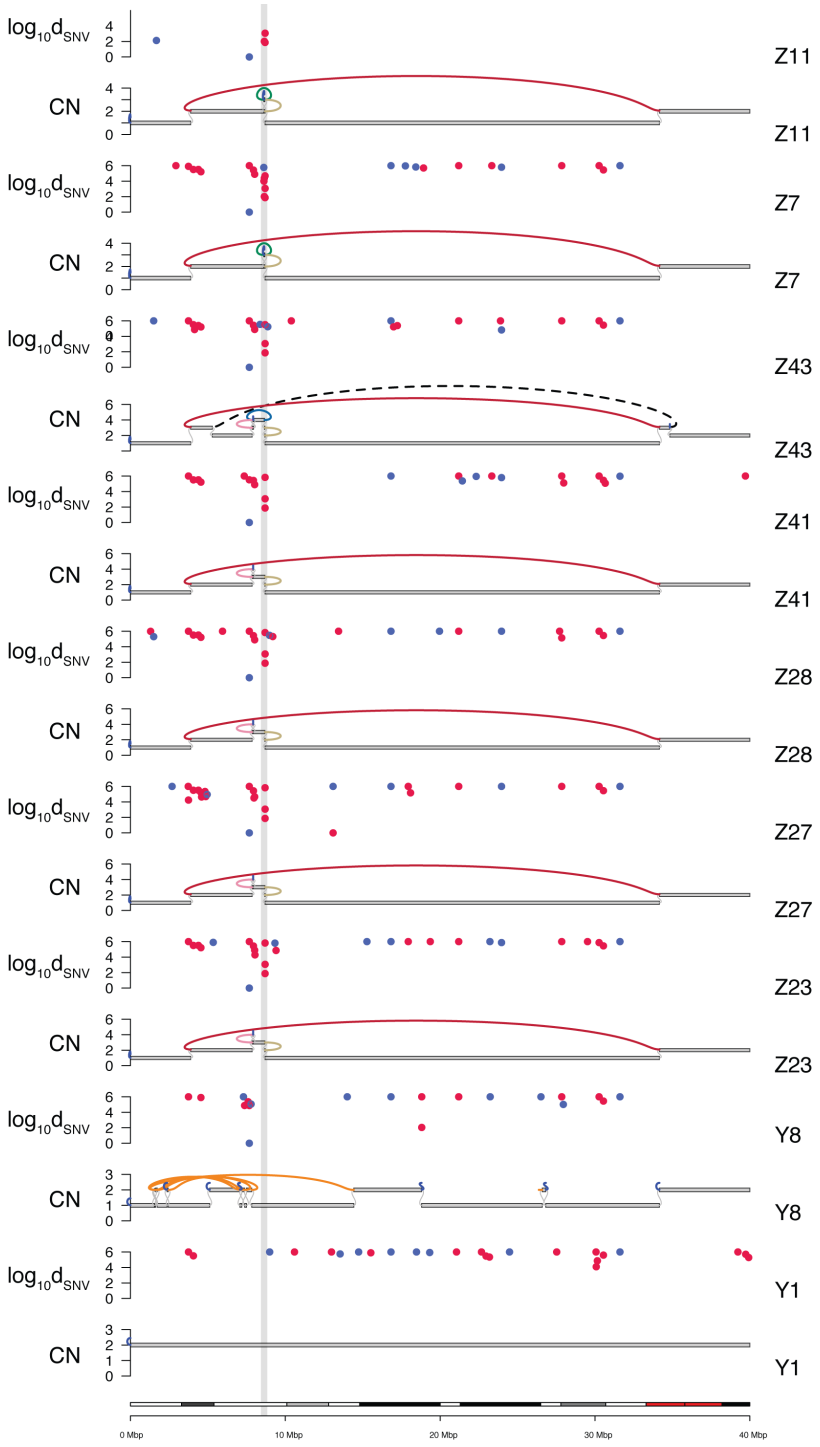

Figure 4.13: **Clustered GC mutations around junction breakends.**

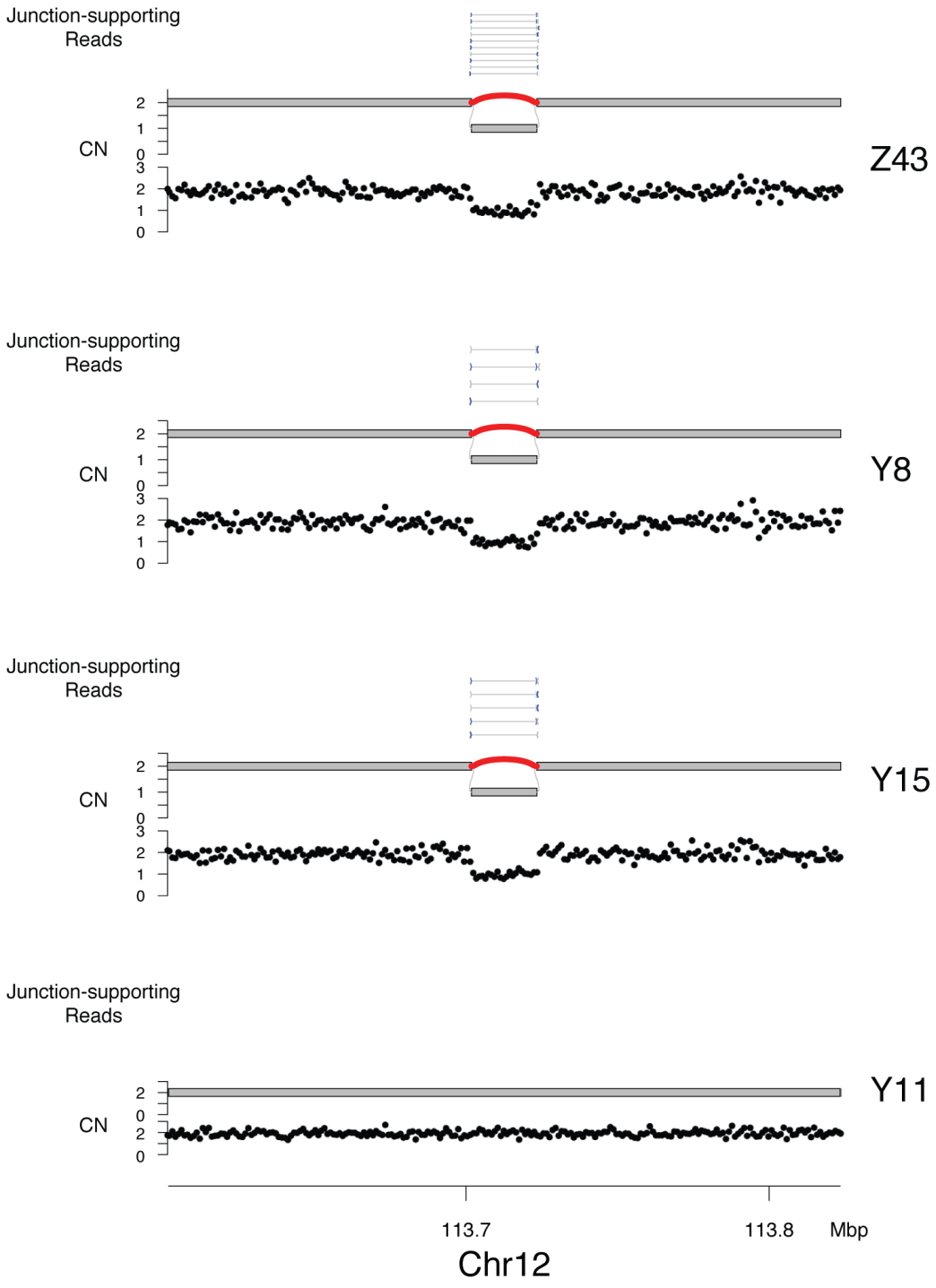

Figure 4.14: **Simple deletion on chromosome 12q present in Y15, Y8, Z43, but not Y11.**

Junction supporting read pairs and corresponding drop in coverage at the small deletion on chromosome 12q in Y15, Y8, Z43, and lack of such evidence in Y11.

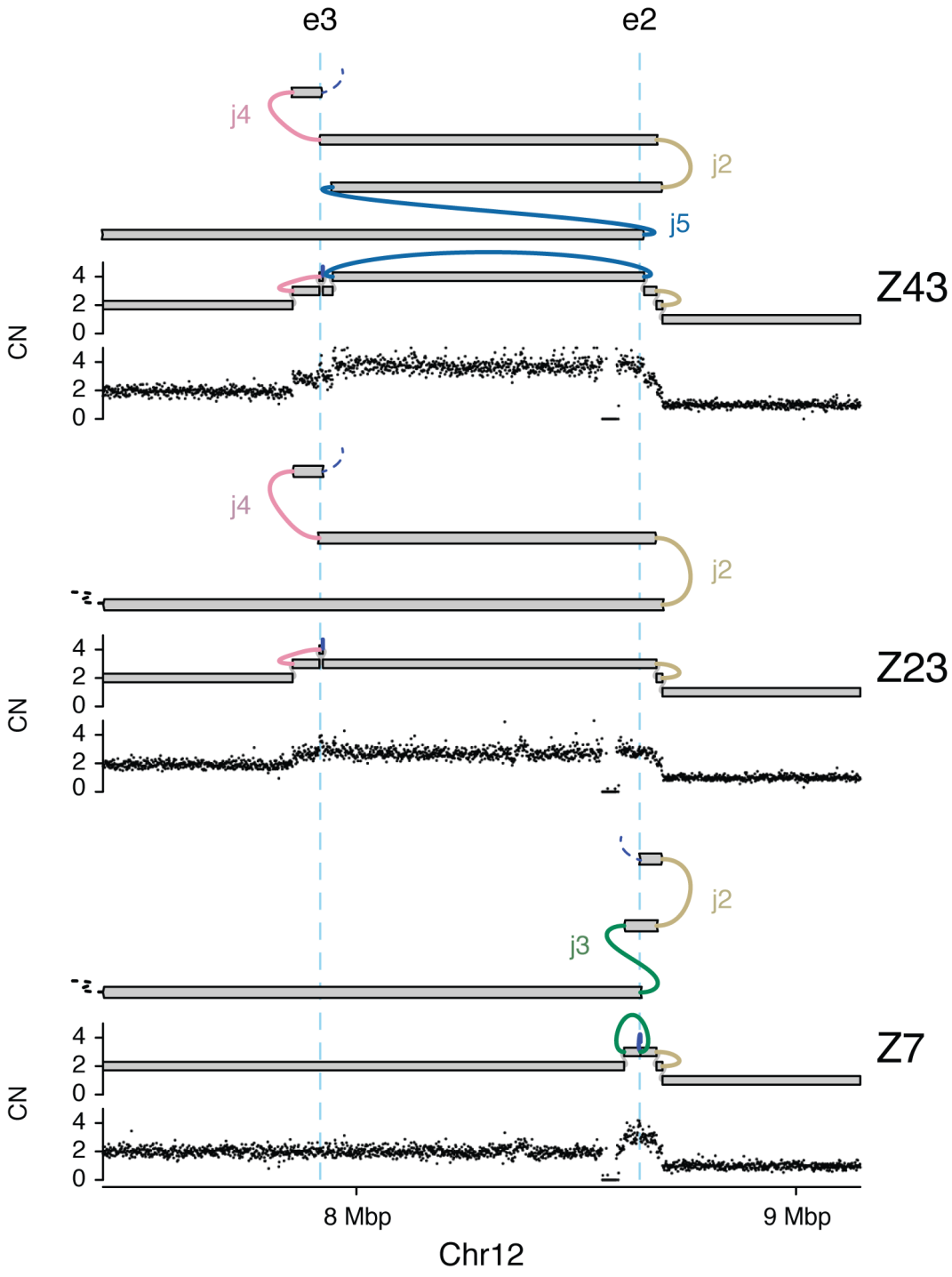

Figure 4.15: **Distinct loose ends among the diverging BFB clones.**

We confirmed a chromothripsis event in an independent lineage (Y8), which lacked j1 and all subsequent junctions of the BFB lineage, further supporting the idea that this is an independent lineage (Figure 4.11). Integration of copy number data with the SNV phylogeny showed clones from the unrearranged lineage (Y1 and Y4) and one of the 12p arm loss clones (Y11) to be mutationally distant (>2,000 SNVs) from the chromothripsis (Y8) and BFB lineages, which shared over 1,583 SNVs (Figure 4.12). Supporting this, a small (21.5 kbp) simple deletion junction was shared across Y8, Y15, and all the BFB lineage samples, yet was absent in Y11 (Figure 4.14).

This comparison established that the 12p loss in Y11 could not have occurred after j1 and indicates that a second independent arm loss must have given rise to Y15. Interestingly, the Y15 arm loss clone was clustered in the BFB/Y8 clade in the SNV phylogeny, sharing 30 SNVs with the BFB lineage which it did not share with Y8 (Figure 4.12). This indicates that the 12p arm loss in Y15 may have arisen either before or after j1. Although the breakpoints of the Y11 and Y15 arm losses could not be mapped due to their location in the 12 centromeric region, based on the SNV phylogeny, they likely represent distinct events. Taken together, these results support a model whereby at least three lineages independently rearranged a previously wild type 12p during telomere crisis (Figure 4.11). Our data appear to have captured sequential steps in the formation of an increasingly complex BFB-like event. Each of these stages must represent a stabilized allele since the post-crisis lines are clonal, and multiple clones share the same rearrangement junctions (Figure 4.12). This necessarily raises the question as to what caused the on-going instability, and how and where these complex alleles are terminated.

4.9 Resolution of BFB cycles in telomere crisis

Analysis of junction-balanced genome graphs allows for the nomination of 'loose ends' (or allelic ends), representing copy number changes that cannot be resolved through assembly or mapping of short reads. We identified three distinct loose ends across the 4 variant graphs spanning the 8 clones in the BFB lineage (Figure 4.15). Each of these loose ends were placed at the terminus of their respective reconstructed allele, and we posit they represent the new 'ends' of the derivative alleles of the BFB lineage. Distinct ends for each of these rearranged lineages suggests the derivative 12p allele could have been stabilized independently. We did not observe telomere repeat-containing reads mated to these loose ends (Appendix 6.5), arguing against neo-telomere formation at these loci. Instead, loose reads represented highly repetitive unmappable sequences which may be a result of the junctions being in close proximity to centromeric regions (see below).

To resolve the genomic architecture at these loci we generated karyotypes from metaphase spreads for representative rearranged clones (refer to [111] Figure 6a and Supplementary Figure 6a), which revealed that in the BFB and chromothripsis (Y8) lineages, the chromosome 12 derivative was likely linked to a copy of chromosome 21 with an intact long arm (q). These observations were confirmed with chromosome painting, demonstrating a derivative chromosome transitioning between 12 and 21 (refer to [111] Figure 6b and Supplementary Figure 6b). Two possible events can explain these findings: the 12-21 fusion could have occurred as an early event during telomere crisis, preceding the divergence of Y8 (chromothriptic) and the BFB lineage; alternatively, independent 21 fusion events stabilized the derivative chromosome 12 following formation

of the distinct junction lineages in Figure 4.11. We consider the first possibility unlikely since the creation of the long-range inversion (j1) and subsequent fold-back junctions in the BFB lineage would require the formation of interstitial 12p breaks on a 12p-21 derivative chromosome. Such breaks are predicted to result in the loss of 21, which would be distal to these junctions on the fusion allele. Furthermore, the acrocentric nature of chromosome 21 would make it more likely to stabilize the overall chromosome architecture, suggesting that an early 12-21 derivative chromosome would be unlikely to engage in the additional SV events observed in the BFB lineage. We therefore consider it likely that each of the BFB cycles and chromothripsis clones were independently resolved through subsequent fusion to 21 (refer to [111] Figure 6c).

Unlike the BFB and chromothripsis clusters, one of the two 12p arm loss lineages (Y11) did not appear to be fused to chromosome 21. In this clone, the derivative chromosome 12 appears to contain a distinct fusion with a longer p-arm (refer to [111]Supplementary Figure 6b). This is consistent with our analysis of the SNV phylogeny, showing Y11 to be mutationally distant from the BFB lineage (Figure 4.12, 4.14). We were unable to further resolve the nature of the stabilization event in this clone. It would be necessary to perform long molecule DNA sequencing across different lineages in order to confirm the distinct nature of the fusion junction in each of the post-crisis clones.

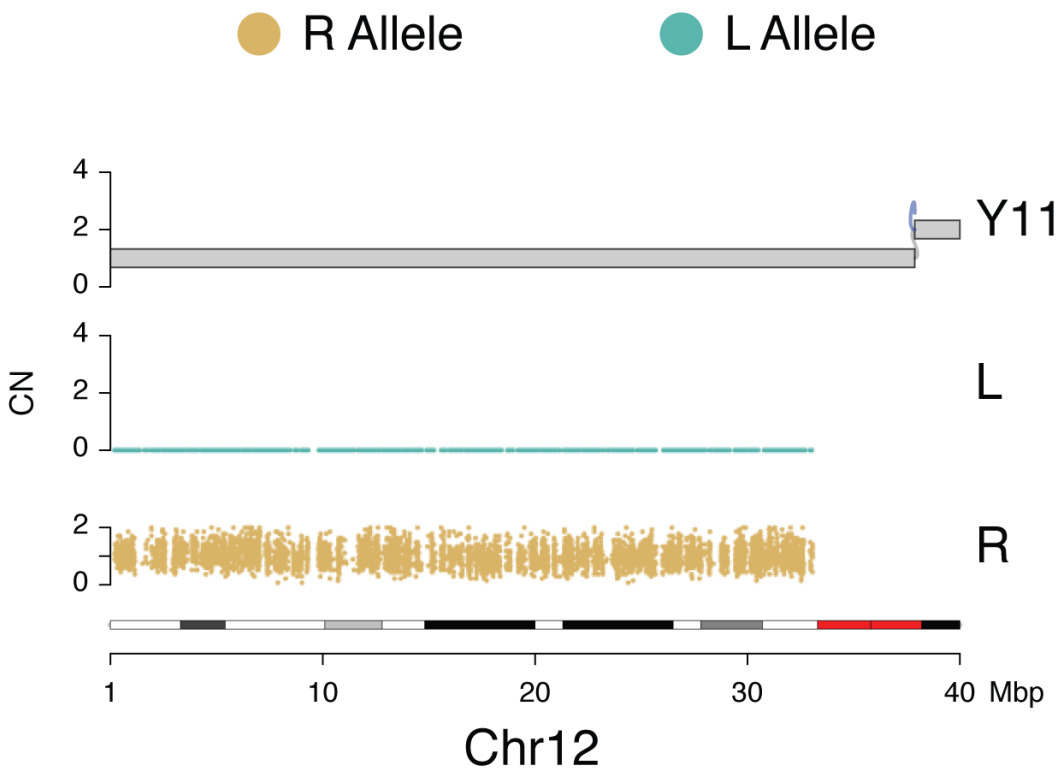

Figure 4.16: **Defining L and R alleles of chromosome 12p based on its loss or retention in arm-loss clone Y11.**

Genomic track plots of parental alleles phased into lost (L) and retained (R) haplotypes on chromosome 12p of clone Y11.

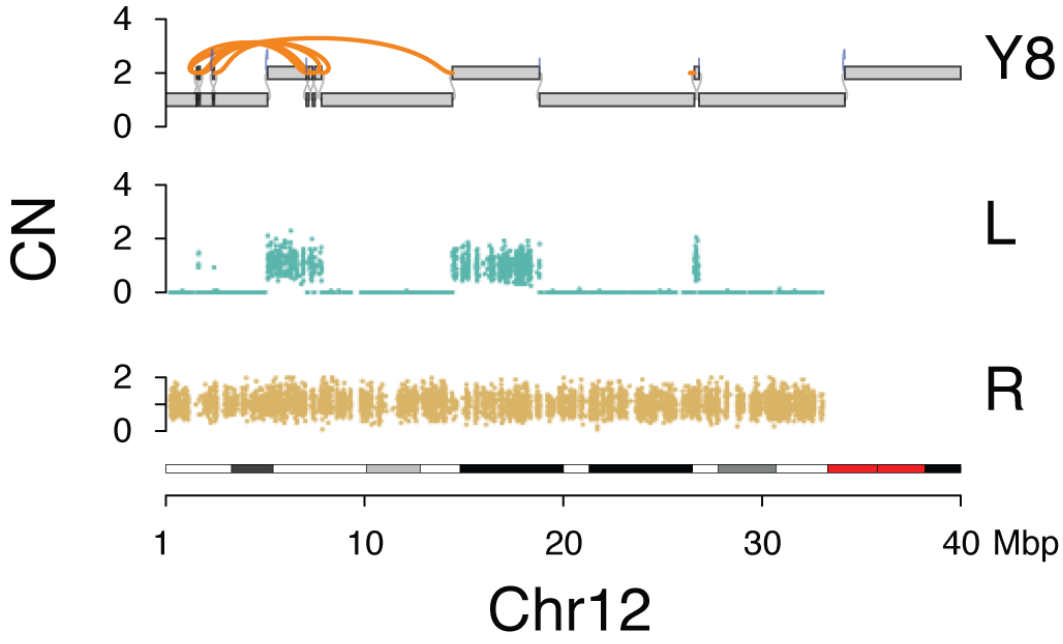

Figure 4.17: **Chromosome 12p haplotype in chromothripsis-like post-crisis clone Y8.**

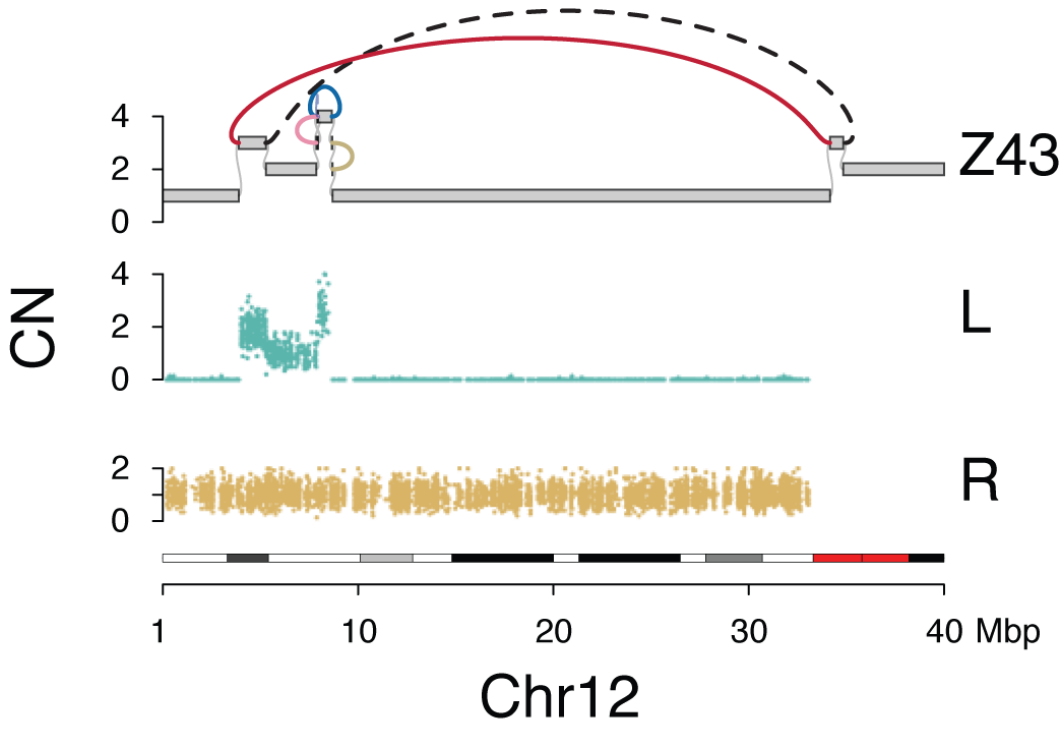

Figure 4.18: **Chromosome 12p haplotype in chromothripsis-like post-crisis clone Z43.**

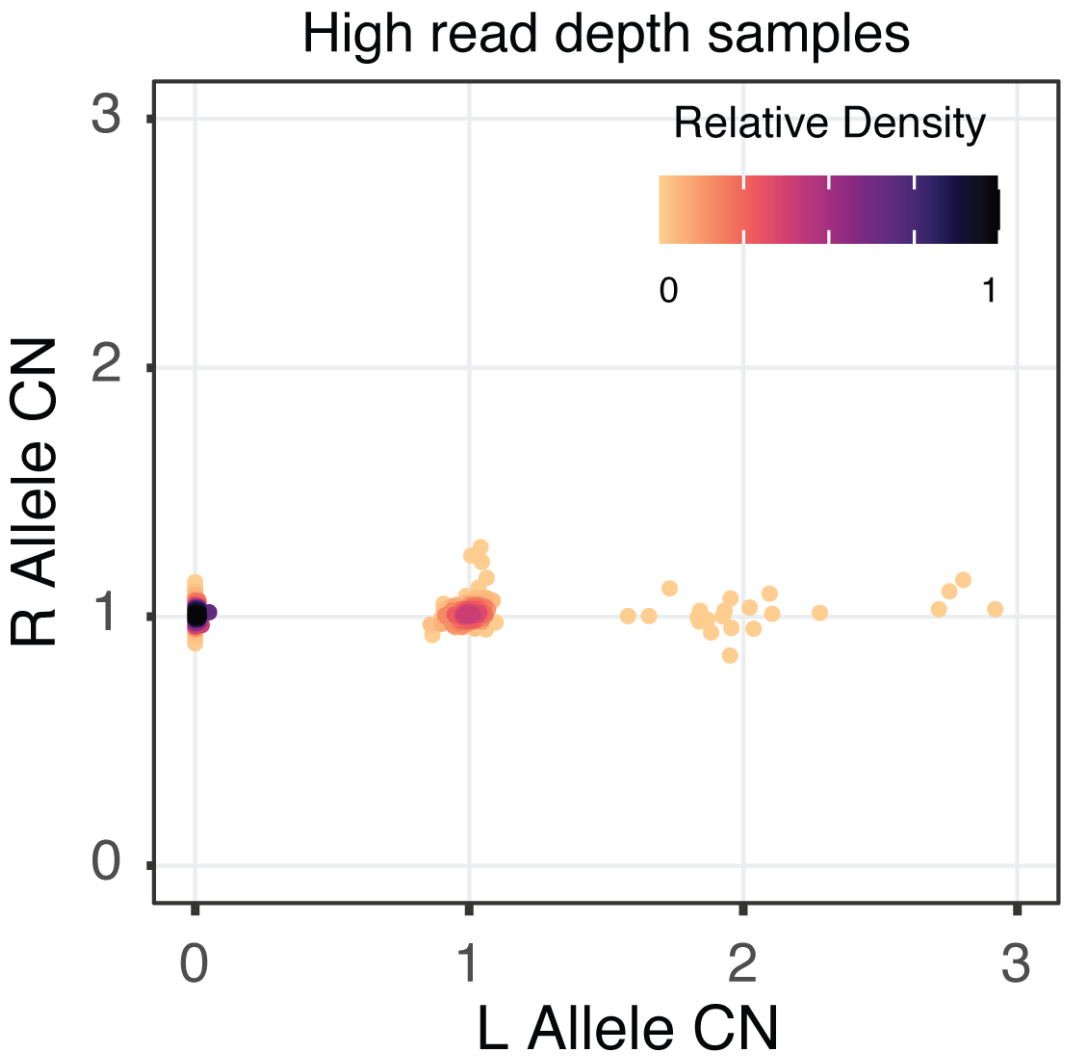

Figure 4.19: **Haplotype CN in high-pass WGS.**

Scatter plot showing purity- and ploidy-transformed L and R haplotype-specific allelic read depth across 12p segments in high-pass WGS-profiled post-crisis clones.

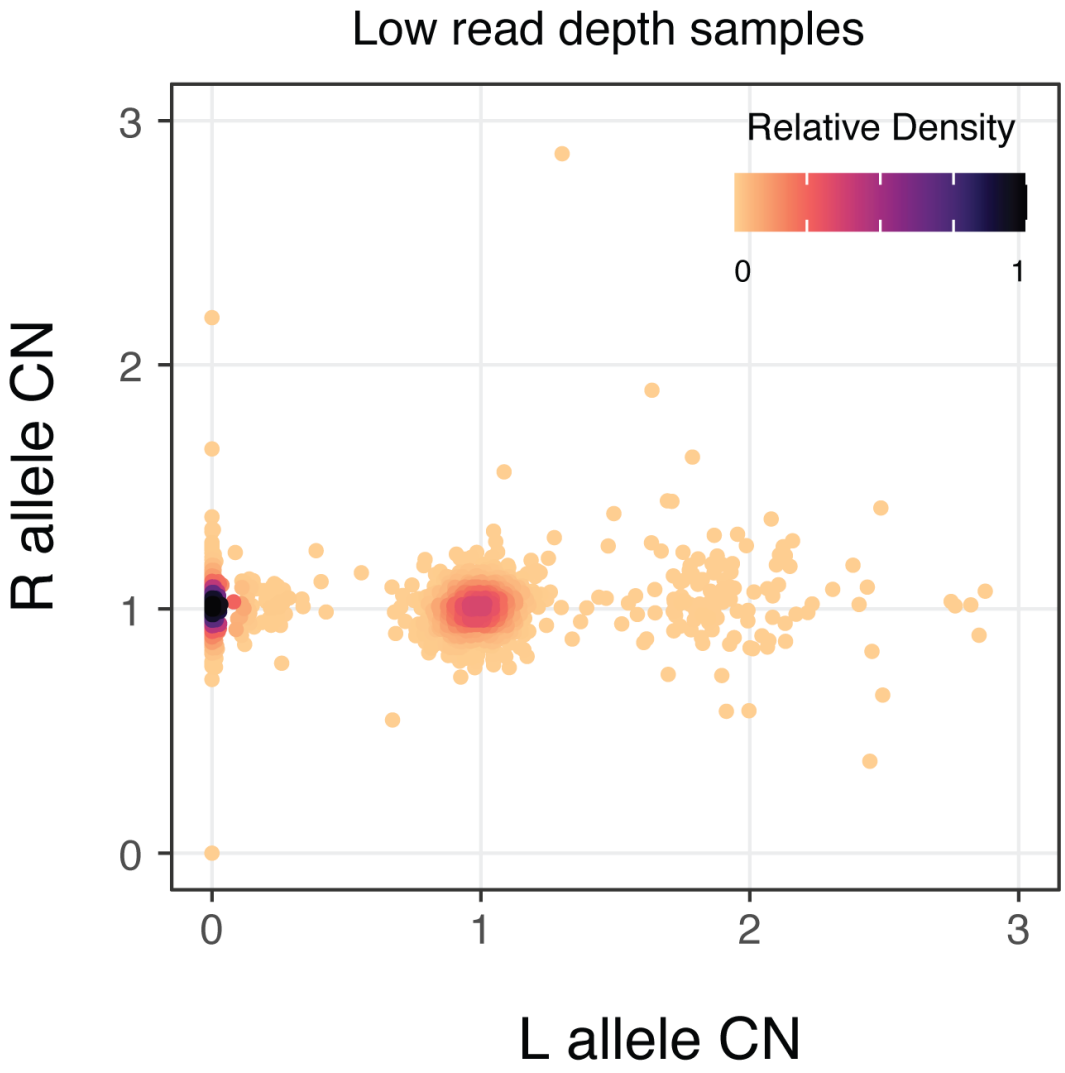

Figure 4.20: **Haplotype CN in low-pass WGS.**

Scatter plot showing purity- and ploidy-transformed L and R haplotype-specific allelic read depth across 12p segments in low-pass WGS-profiled post-crisis clones.

4.10 A short telomere renders 12p vulnerable to telomere attrition

The convergent evolution patterns observed in our system suggests either 12p vulnerability to rearrangements or selection for 12p loss during telomere crisis. We believe strong selection is unlikely, given the existence of day 150 clones with diploid 12p (15.8%, 13/82, with or without 21q gain). The preferential rearrangement of the short arm of chromosome 12 in the post-crisis system could be explained if one of the two 12p telomeres is among the shortest telomeres in the MRC5 parental cells. Attrition of the shortest telomeres is predicted to generate the first telomere fusions and associated rearrangements in the culture.

We first asked whether the same parental allele was targeted across the chromosome 12- associated events in our cohort. Such allele specificity would argue against a selection for loss of 12p sequences since such selection should have occurred without allele preference. We phased heterozygous SNPs on 12p on the basis of whether they belonged to the lost (L) or retained (R) allele on the early 12p arm loss clone, Y11 (Figure 4.16). Analyzing phased SNP patterns across all the high– and low-pass MRC5 clone WGS profiles in our dataset demonstrated that the L allele of 12p was the exclusive target of all chromosome 12 structural variants (Figure 4.19, 4.20). This included the clones from the chromothripsis (Y8, Figure 4.17) and BFB (Z43, Figure 4.18) lineages, which our phylogenetic clustering suggested to be likely independent events on a previously unrearranged chromosome 12 (Figure 4.11). On the basis of these results, we concluded that the short arm of the L allele of 12 was the most vulnerable to rearrangement in the MRC5 parental line.

We next tested whether the preferential 12p events could be due to the presence of a short telomere on one of the 12p alleles. To this end, we combined telomeric FISH with BAC probes specific for chromosome 12 and two other chromosome (6 and 8) that did not show evidence for structural variants in WGS (Figure 4.10). Comparing the ratio of the telomeric signal of the shortest 12p telomeres to the signal of all other telomeres in individual metaphase spreads revealed that one of the 12p telomeres was significantly shorter (data not included, refer to [111] Figure 7e). The shortest telomeres of 6 and 18 (data not included, refer to [111] Supplementary Figure 7b) were also shorter than the median but not to the same extent as 12p. The relative telomere length of the shortest 21p allele showed a heterogeneous distribution that overall was significantly longer than 12p in the parental cells (data not included, refer to [111] Figure 7e). This does not exclude the possibility of 21 becoming critically short at later time points, and indeed the observation of a low percentage of clones in the 5X WGS screening with amplifications of 21q could indicate that this chromosome end did occasionally become deprotected in this population (Figure 4.10). Such deprotection of a chromosome 21 telomere is consistent with chromosome 21 preferentially stabilizing the derivative chromosome 12 (data not included, refer to [111] Figure 6b, Supplementary Figure 6a).

To look for evidence of chromosome 12 being involved in the initial fusion events in this system, we combined a chromosome 12 BAC probe with a centromere probe in MRC5/Rbsh/p21sh/iCRISPRa-TERT cells in crisis (at day 90). Strikingly, we observed a number of instances of chromosome 12 within chromosome fusion events (data not included, refer to [111] Figure 7f). The fraction of chromosome fusions involving chromosome 12 is higher than expected (50% observed versus 4% expected, data not included, refer to [111] Supplementary

Figure 7c). Collectively, these data support the hypothesis that a short telomere on one allele of 12p increased the chance of 12p partaking in a fusion event that preceded subsequent rearrangement lineages.

4.11 Discussion

To precisely define the chromosomal consequences of telomere crisis, we have described the first whole genome profiles of cells emerging from natural telomere crisis, both in the setting of spontaneous and controlled telomerase activation.

Previously published post-crisis cell lines show a wide variety of SV complexity. Some have little to only simple, arm-level CNAs. Some have extensively rearranged chromosomes ending up with complex amplification. However, there is only one instance that fit previously implicated chromothripsis patterns as defined by the gGnome event callers trained from pan-cancer data. Even though several amplicons with fold-back inversions are discovered, they are still relatively primitive compared to the BFB cycles defined in cancer genomes. Hence, canonical cancer complex SVs of chromothripsis and BFB cycles are not a necessary outcome of telomere crisis, even though they are possible. In other words, their presence alone cannot be used as a sufficient indicator that a cancer went through telomere crisis. These cells were not subject to any artificial controls and thus have spent varying time in crisis, potentially with different severity, and natural selection.

Hence we built a new *in vitro* telomere crisis system. By adding an inducible system of *TERT* expression in p16/Rb1 silenced MRC5 cells, we can precisely

time the escape from telomere crisis. Plus, MRC5 cells do not have known defects in major DNA repair pathways, like the majority of cancers at early stages, so that the alterations post-crisis can be directly attributed to the telomere crisis. Indeed, based on our observations of relatively low *TERT* expression, low dicentric chromosome count, we consider our model closer to what commonly happens in cancerous cels. Using this model, by a combination of WGS and cytogenetics, we characterized the start, progression, and termination of the SV evolution during telomere crisis.

In the low-depth WGS screening, we observed the only chromosome arms showing prevalent CNAs are chromosome 12p and 21q. The six clones of 21q amplification is mutually exclusive with the 12p altered groups, in which 21q appears to be diploid. The rest of the clones that are devoid of major CNAs throughout the genome, covering both early and late harvest time, and they represent the clones that escaped telomere crisis readily before any SVs could accumulate. The very existence of unaffected clones even on 150 days in, shows us that the SVs or lack thereof are not the result of positive selection.

To reconstruct the different yet related chromosome 12p structures, we further generalized our JaBbA formulation to penalize unique loose end locations across several genome graphs (see Eq. 4.2), ergo minimizing complexity of a genome graph population. Without this extension, each genome graph may place loose ends based on each coverage profile, resulting in discrepancy between samples for the same true loose end, in turn, misleading the downstream phylogeny inference. This development demonstrated the adaptability of gGnome and bear crucial application in other multiple related samples study

designs, for example, single cell DNA sequencing of heterogeneous cell populations.

The reconstructed graphs showed three main branches of outcomes to chromosome 12p: arm loss with breakpoint hidden within the centromere, chromothripsis, and several stages of BFB cycle. Taking advantage of the complete arm-loss clone Y11, we phased the germline variants on chromosome 12p and found a perfect accordance among all these clones in both low- and high-pass WGS: there is a single chromosome 12 allele that was altered in all these samples. This convinces us the initial trigger, or the first critically short telomere was from chromosome 12p. This is also indirectly shown as one intact chromosome 12 exists in each of these clones by FISH. This point is further verified by telomere binding FISH and found that chromosome 12p does have the shortest telomere when compared to several other chromosomes in MRC5 cell line. This finding is consistent with prior hypothesis that the shortest telomere becomes unstable and initiate the chromosomal alterations during crisis.

To reconstruct a comprehensive evolutionary trajectory, we combined evidence from the exact structural haplotype and genome-wide SNVs. We exhaust all possible haplotypes that can be derived from the graphs and search for a linear combination of them that satisfies each graphs copy numbers. To achieve that through multiple related samples, we have to extend the Eq. 2.5 to the form in Eq. 4.3. Our exhaustive nomination of elementary paths and cycles from a graph, even though guaranteed to be feasible, any cycle is singled out and on face value form itself a circular contig, even for a true tandem duplication. This is biologically improbable so we modified the pool of haplotypes by inserting cycles back to all possible locations among the walks, and fitted copy number

to a more biologically meaning full haplotype repertoire. Thanks to the knowledge of a single chromosome 12 allele being affected, we further simplify the inference by requiring a unrearranged chromosome 12 haplotype to present at copy number 1 in each sample.

While we inferred this tree (Figure 4.11) in a customized deduction, there is great opportunity to build a general-purpose, automated evolutionary trajectory inference method from jointly inferred haplotypes. Interestingly, the two arm-loss configurations are actually distant on the SNV phylogenetic tree, indicating convergent evolution. Y15 is closer to the BFB clones than to the other arm-loss Y11 clone, which was further supported by a small simple deletion event on the q-arm of chromosome 12. Assuming a relatively constant rate of mutation, the diverging process of the BFB clones happened in a shorter time frame.

Since all of these data are sequenced from stable, supposedly isogenic cell populations, after telomere crisis, there should not be ongoing genome instability, and it requires each of them to be stabilized in some way. In other words, the damaged chromosome 12p must have regained functioning telomere. Judging from the karyotyping, most of these clones lack a copy of chromosome 21, even though in WGS chr21 seems intact in copy number. Supported by bi-color FISH of both chromosomes 12 and 21, we found the acrocentric chromosomes have been fused to the damaged chromosomes 12. Contrary to our first hypothesis that this fusion happened in ancestral state and early on, the topology of junctions gained through the process requires the breakage between chromosome 12p and 21q if it was the case. We deem this breakage and re-fusion highly improbable, so we have to land in the conclusion that these fusions happened independently, implying striking convergent mechanism to stabilize a structurally unstable chromosome.

In conclusion, our data reveal that telomere crisis can instigate a wide spectrum of Structural variants in the viable descendants of this genomic trauma.

First, our results indicate that cells can emerge from telomere crisis with minimally altered genomes. Second, BFB cycles and chromothripsis are not a universal hallmark of post-crisis cell lines. Third, our results indicate that natural telomere crisis can manifest as a focal and diverse cascade of SV events converging on a single chromosome arm. Since no single class of Structural variant appears to be a hallmark of past telomere crisis, other genomic insignia will have to be identified in order to determine whether a given cancer has experienced telomere dysfunction in its proliferation history.

CHAPTER 5

WHOLE-GENOME CHARACTERIZATION OF LUNG ADENOCARCINOMAS LACKING ALTERATIONS IN THE RTK/RAS/RAF PATHWAY *

In this final chapter, I describe the collaborative effort within the Cancer Genome Atlas (TCGA) consortium to characterize the whole genomes of lung adenocarcinoma lacking canonical pathogenic alterations in RTK/RAS/RAF pathway as determined by whole exome sequencing (WES) and transcriptome sequencing (RNA-seq).

Individual contributions: The Cancer Genome Atlas consortium conceptualized, initiated the project, and collected data. Jian Carrot-Zhang, Nicolas Robine, and Marcin Imielinski, led the parts of the study shown in here. Matthew Meyerson, Ramaswamy Govindan, and Marcin Imielinski oversaw the project and the final manuscript. I executed the parts of the whole genome analysis presented here and prepared the results within the manuscript [152].

5.1 Introduction

As one of the most common and deadliest cancer [153], lung adenocarcinoma has been known to be driven by genetic alterations (e.g. mutations, fusions, amplifications, deletions) in receptor tyrosine kinases (RTK) and its downstream RAS/RAF/MEK cascade proteins, that eventually activate MAPK signaling pathway [154]. Previously, a series of TCGA studies estimated 70-80%

*Carrot-Zhang, J. et al. Whole-genome characterization of lung adenocarcinomas lacking alterations in the RTK/RAS/RAF pathway. Cell Rep. 34, 108784 (2021)

of LUAD cases harbor drivers in the RTK/RAS/RAF pathway using WES and transcriptome sequencing (RNA-seq) [155, 156, 120]. Based on these knowledge, many small molecule inhibitors have been developed to target various proteins along this pathway including *EGFR*, *BRAF*, *ALK*, *RET*, *ROS1*, *KRAS*, and provide unprecedented clinical benefit for LUAD patients [157]. Thus, it is crucial to design therapeutic regime based on accurate information about if and what RTK/RAS/RAF pathway alterations (RPA) are present and likely functional within a tumor.

Apparently, the remaining 20-30% of cases, which we refer to as RTK/RAS/RAF pathway alteration-negative or RPA(-) LUADs, pose a major clinical challenge in precision thoracic oncology [155]. One immediate question is whether RPA(-) LUAD driven by some distinct RTK/RAS/RAF-independent pathways, or whether the profiling technology (WES and RNA-seq) missed relevant events due to technical factors like uneven coverage or low sample purities.

Whole genome sequencing offers a chance to gain a better picture of the variants within these tumors. We anticipate the addition of WGS data to at least 1) either verify the RPA(-) status or recover previously missed known RPAs, 2) nominate potential SV and non-coding drivers hard to capture with WES, 3) determine the extent and type of genomic instability to compare to RPA(+) LUADs. Our genome graph paradigm offers a unique advantage to not only enhance the accuracy of identifying candidate alterations, but also reveal their underlying mutational mechanism. To that end, we performed WGS on LUAD samples from The Cancer Genome Atlas (TCGA) cohort that had appeared

in previous analyses to lack an RTK/RAS/RAF pathway-activating alteration [155] and comprehensively characterized their variant landscape in Chapter 5.

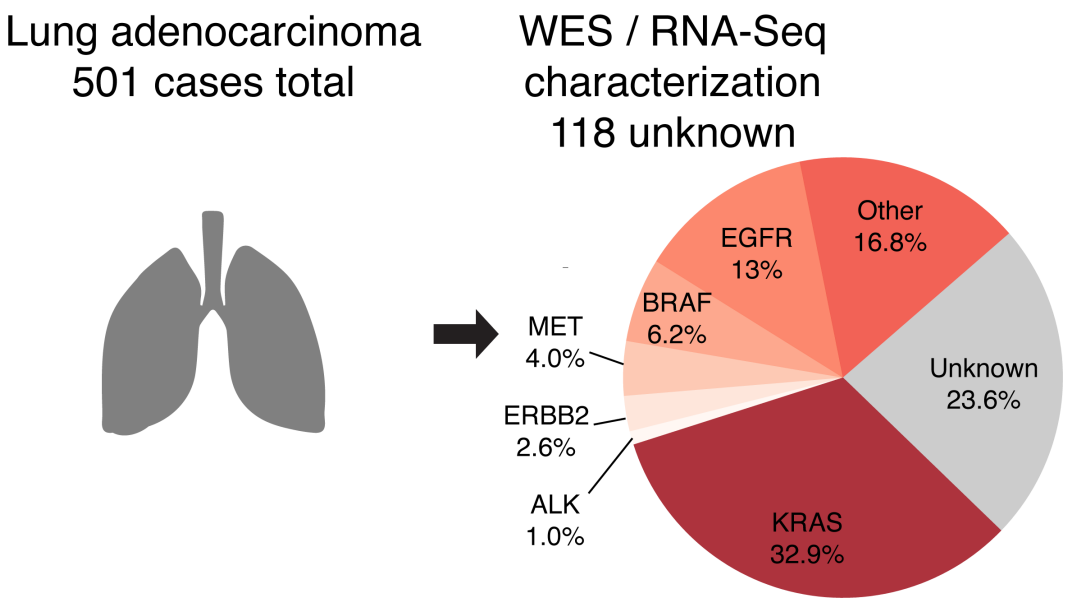

Figure 5.1: **Driver alterations composition in 501 lung adenocarcinoma of TCGA.**

About 23% of TCGA LUAD cases have not been associated with an known alteration in the RTK/RAS/RAF pathway.

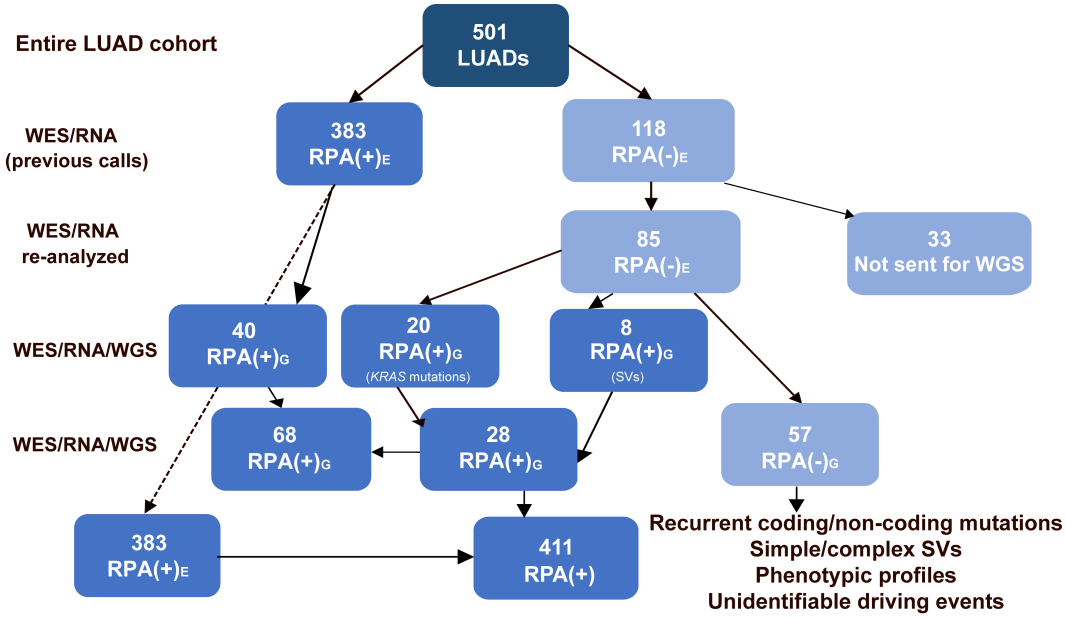

Figure 5.2: **Classification of RPA(+) and RPA(-) LUADs.**

From the entire LUAD cohort, 383 RPA(+)$_E$ and 118 RPA(-)$_E$ samples are identified based on WES and RNA-seq data, and 85/118 RPA(-)$_E$ samples are sent for WGS. Our WGS analyses re-classify 28/85 RPA(-)$_E$ cases into RPA(+)$_G$. Among the 383 RPA(+)$_E$ cases, 40 cases have WGS data from a previously published TCGA study (Imielinski et al. 2017). Together, we use 57 RPA(-)$_G$ cases, 68 RPA(+)$_G$ cases and a total of 411 RPA(+) cases for downstream analyses.

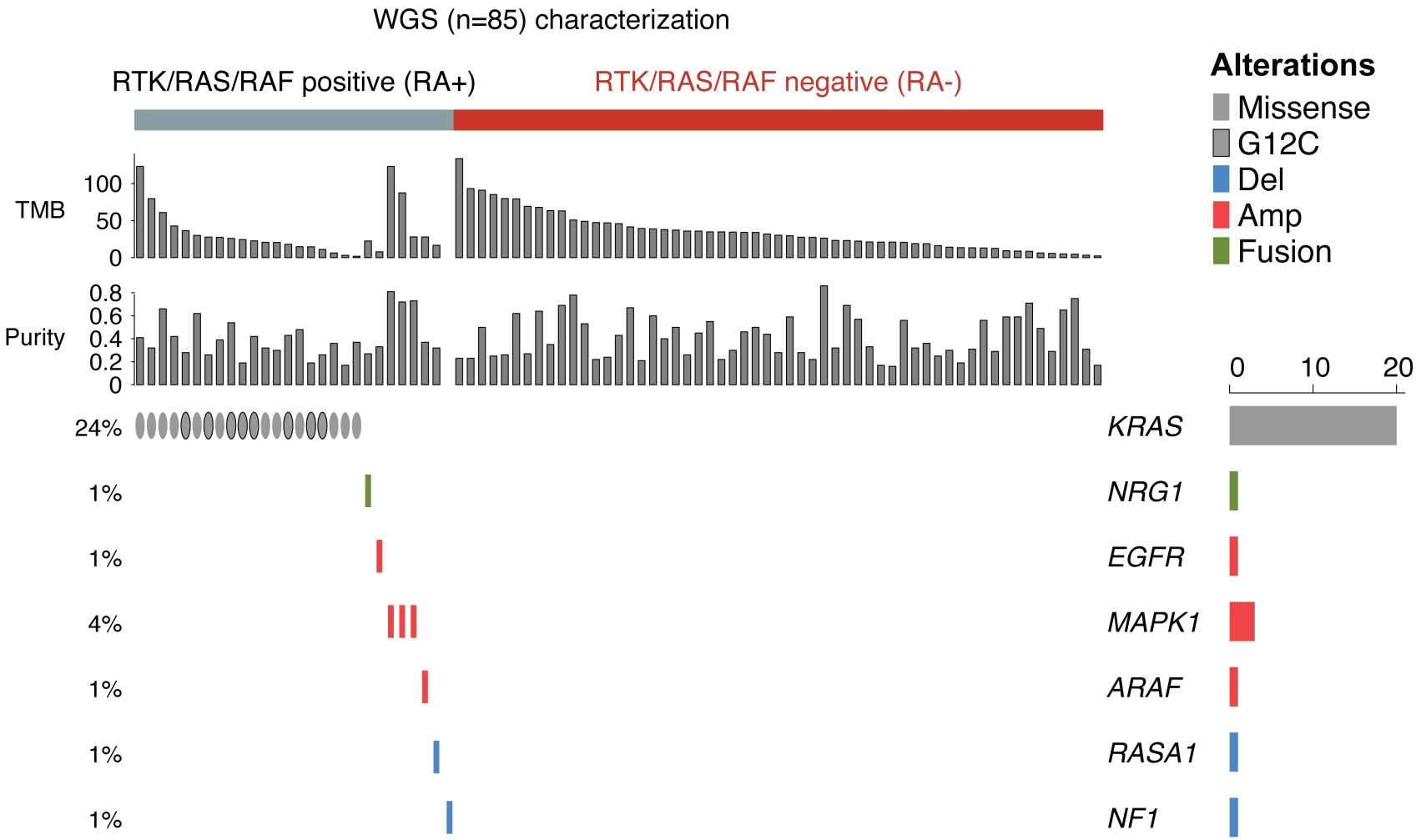

Figure 5.3: **RPA identified in WGS of the 85 RPA(-)$_E$ samples.**

5.2 Identification of RPA(-) LUADs

A previously published TCGA study of lung adenocarcinoma [155] identified 383 of 501 (76%) of LUAD cases as RTK/RAS/RAF-alteration-positive, or RPA(+), using WES and RNA-seq (data not included, refer to [152] Table S1). Samples with an activating mutations in *KRAS*, *EGFR*, *BRAF*, *ERBB2*, *MET*, *RIT1*, *NRAS*, *RAF1*, *HRAS*, *ARAF*, *MAP2K1* or *SOS1*; loss of function mutations in *NF1* or *RASA1*; fusions in *ALK*, *ROS1*, *RET*, *MET* or *NTRK2*; and amplification of *EGFR*, *ERBB2*, *KRAS*, *MET*, *FGFR1* or *MAPK1* were classified as RPA(+) by WES and RNA-seq, which we henceforth designate as RPA(+)$_{\mathrm{E}}$. Among the remaining 118 LUADs, we performed WGS for 85 tumor-normal pairs (Figure 5.1, 5.3, 5.2) at average 73.5-fold (± 7.9 s.d.) and 37.0-fold (± 5.8 s.d.) coverage for tumor and matched normal samples, respectively, followed by somatic single-nucleotide variant (SNV), indel, copy number, and structural variant (SV) analysis (Appendix 6.5).

Our multi-step analysis schema is shown in Figure 5.2. In the first step, we re-analyzed the 85 RPA(-)$_{\mathrm{E}}$ cases to determine if there was truly no evidence of coding mutations in the RTK/RAS/RAF pathway. Surprisingly, we found that 20 of the 85 cases harbored *KRAS* hotspot mutations, with 8 samples showing the recently targetable p.G12C mutation [158, 159] (Figure 5.3). Re-examination of the WES data for those samples confirmed the mutation calls for 16 of the 20 samples, but the read support was insufficient to enable high-confidence variant calling without invoking the WGS information (refer to Figure 1C of [152]). The poor WES coverage for those *KRAS* mutations was likely due to low capture efficiency [160] for the first coding exon of *KRAS*, which contains codons 12 and 13.

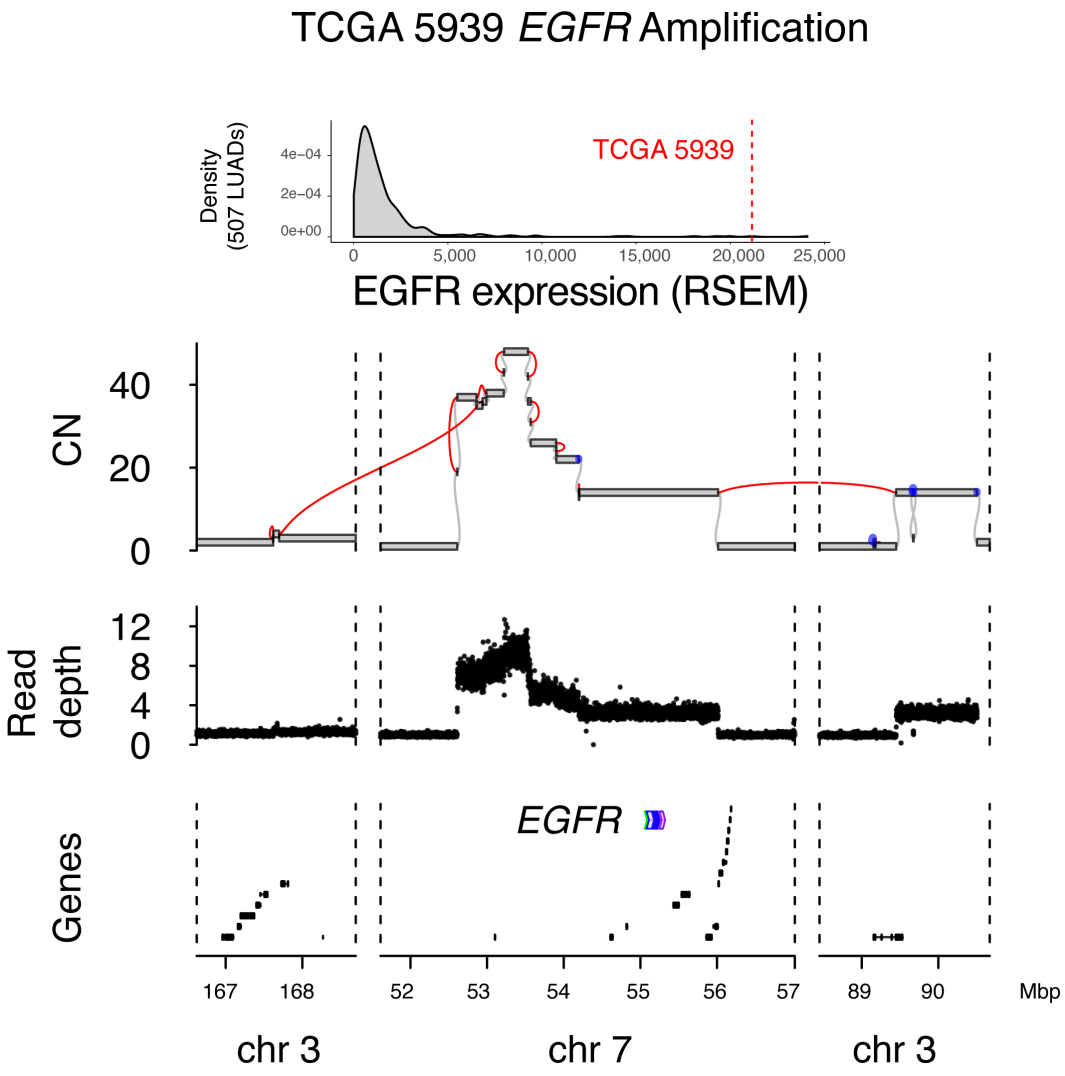

Figure 5.4: **An example of *EGFR* amplification and over-expression by BFB.**

Top panel, *EGFR* RSEM among RNA-seq of LUAD samples. Bottom panel, purity-adjusted copy number and SV junctions (red lines) support a BFB cycles event underlying the amplification. Lower panels indicate WGS read depth and gene location in the region. CN, copy number.

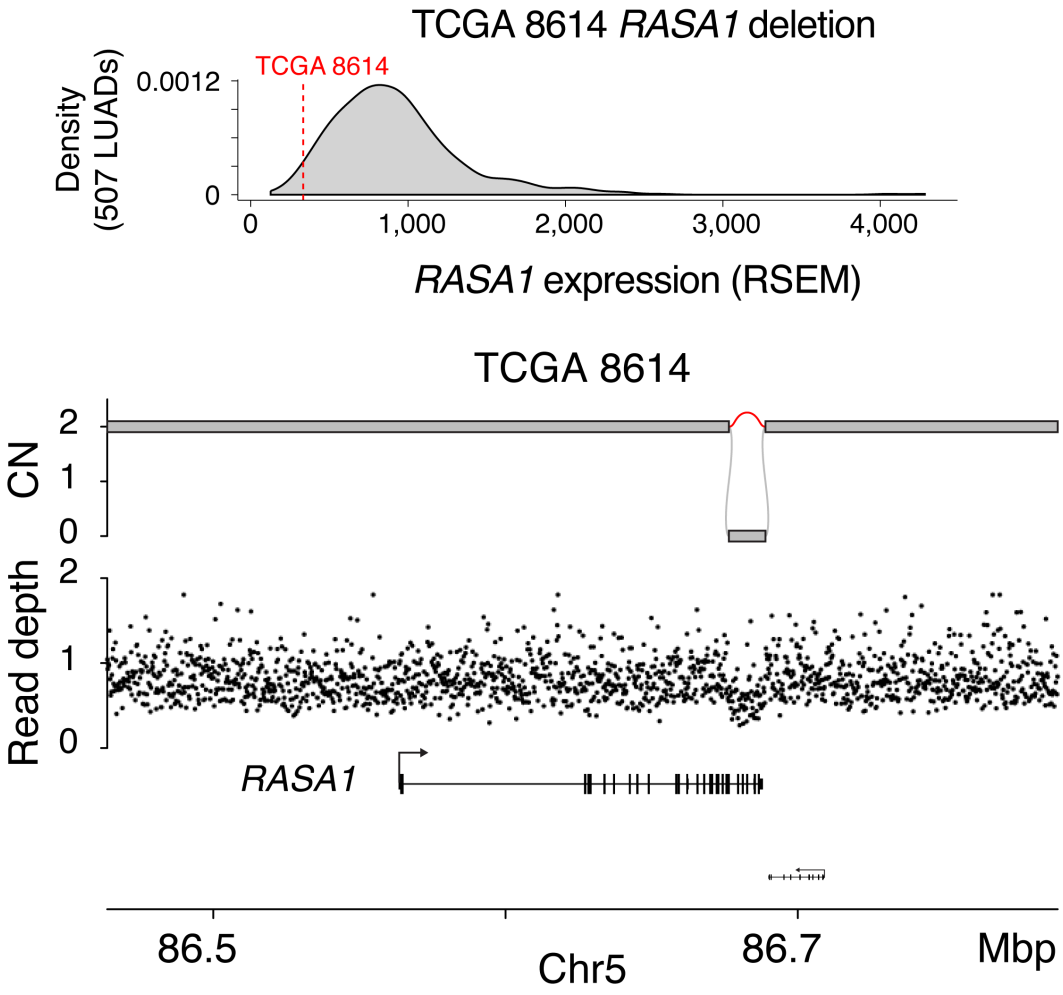

Figure 5.5: **An example of *RASA1* deletion and under-expression.**

Top panel, *RASA1* RSEM among RNA-seq of LUAD samples. Bottom panel, deletion spanning from exon 21 to the end of the gene (12 kbp).

Among the 85 samples, we also identified eight cases with somatic SV and copy number alterations (SCNAs) in known RTK/RAS/RAF pathway members. Among those alterations, we found complex SVs driving high-level amplification and over-expression of oncogenes including *EGFR* (n=1) and *MAPK1* (n=3) (Figure 5.3). Further classification of complex SVs showed the *EGFR* amplification to be driven by a breakage-fusion-bridge cycle (BFB cycles) (Figure 5.4). We also found deletions coupled with loss of heterozygosity (LOH) resulting in decreased expression of *RASA1* (n=1) or *NF1* (n=1) (Figure 5.5, also refer to [152] Table S2), both of which negatively regulate RTK/RAS/RAF signaling. The focal deletion (379bp length) in *NF1* affected only a single exon that is not well-represented in WES or RNA-seq, highlighting the advantage of WGS for identification of focal SV events (refer to [152] Table S2). We also identified one case with *ARAF* amplification and one case with *NRG1* fusion, which are alterations previously shown to activate RTK/RAS/RAF signaling in LUADs [161, 162]. Interestingly, we found amplification and over-expression of *SOS1* in one case (TCGA-62-8399, Figure 5.3). Although *SOS1* mutations have been shown to activate RTK/RAS/RAF signaling [163], the role played in RTK/RAS/RAF activation by the amplification described here is unclear.

Overall, we identified 28 (33%) additional RPA(+)$_G$ cases among the 85 RPA(-)$_E$ that had undergone WGS (Figure 5.2). We labeled the remaining 57 cases as RPA(-)$_G$ LUADs because they lacked an RTK/RAS/RAF pathway alteration identified by WGS (as well as WES and RNA-seq). The remainder of the results reported here focus on that subset.

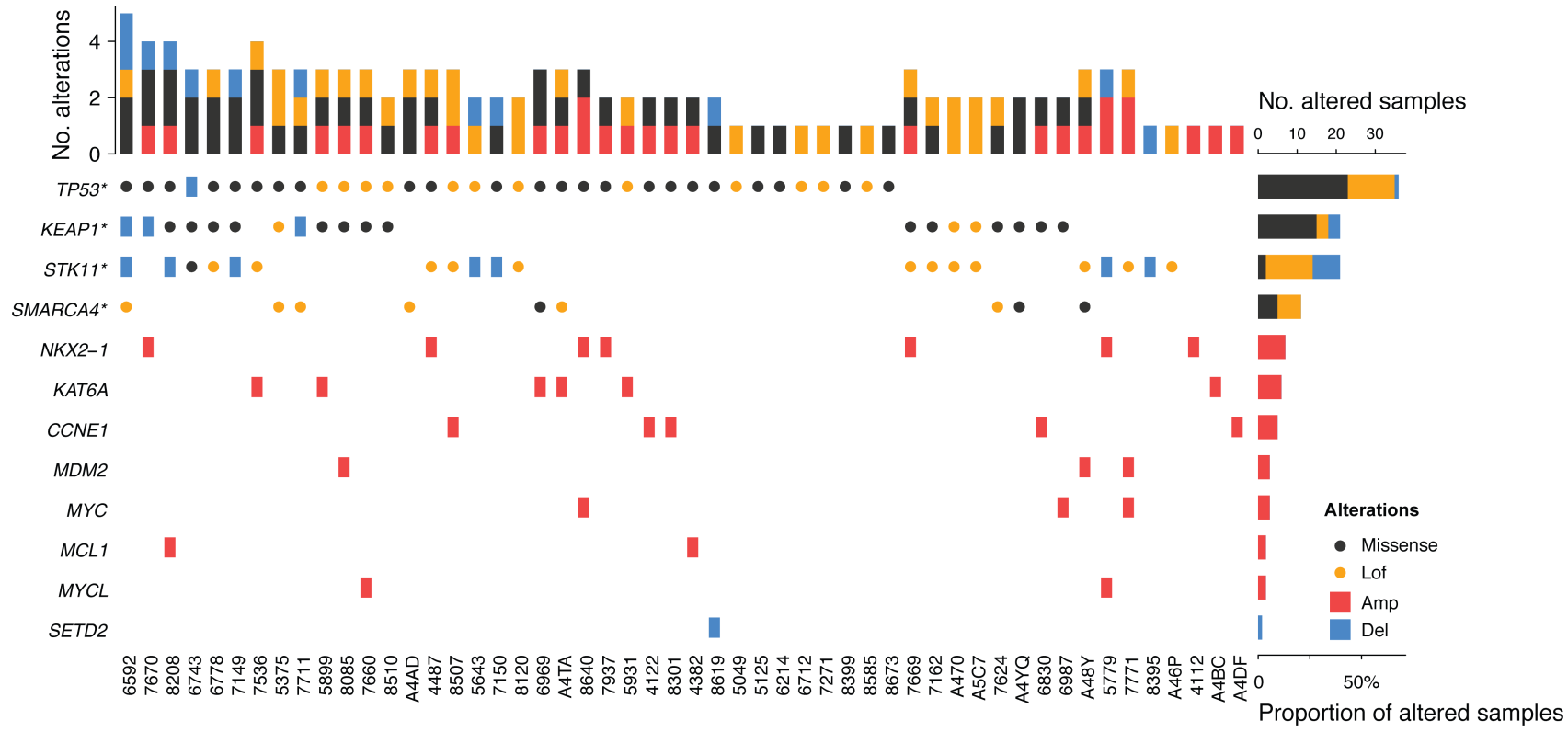

Figure 5.6: **Alterations outside the conventional RTK/RAS/RAF pathway in RPA(-) samples.**

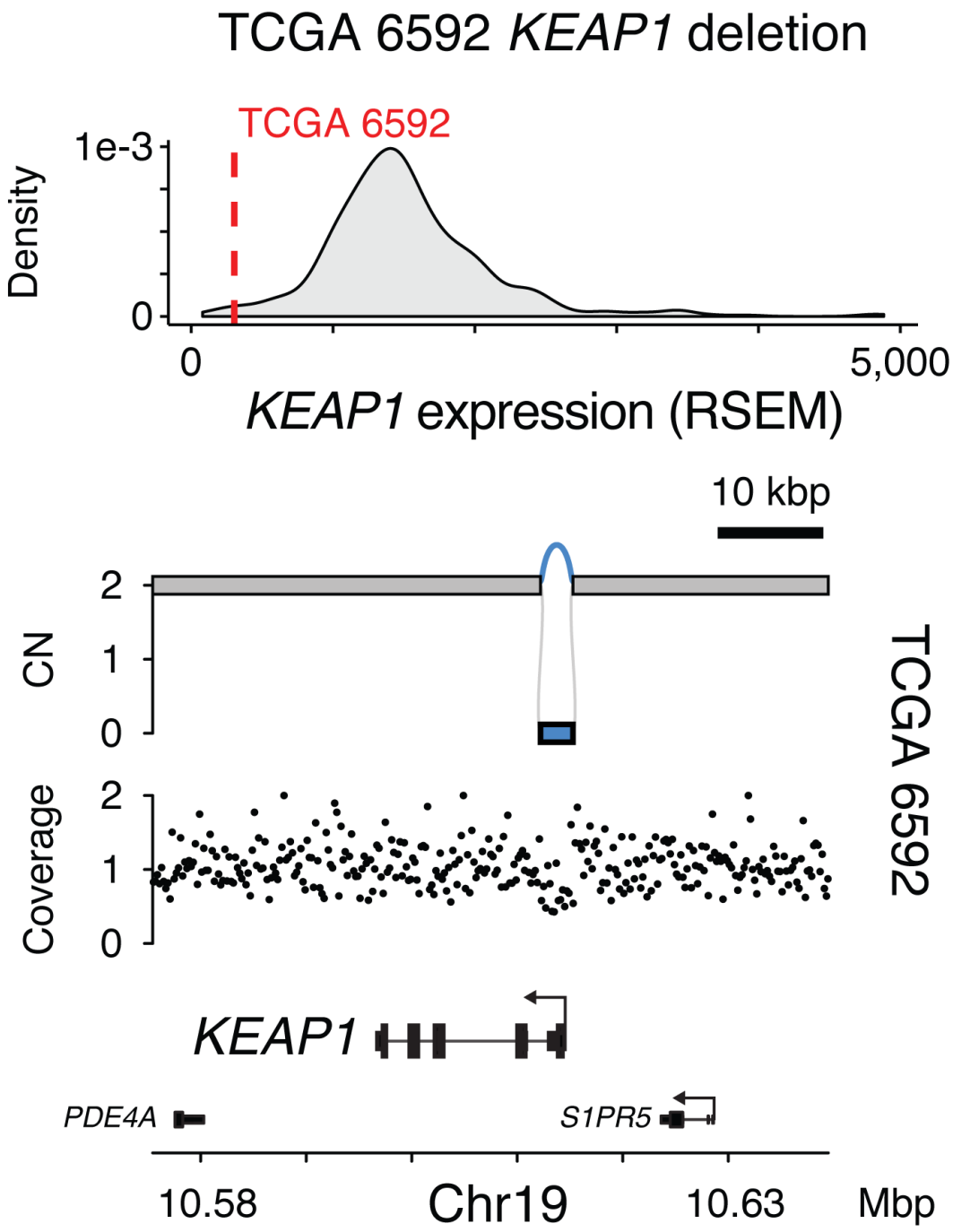

Figure 5.7: **An example of *KEAP1* deletion and under-expression.**

Top panel, *KEAP1* RSEM among RNA-seq of LUAD samples. Bottom panel, 3kbp deletion covering the transcription start site.

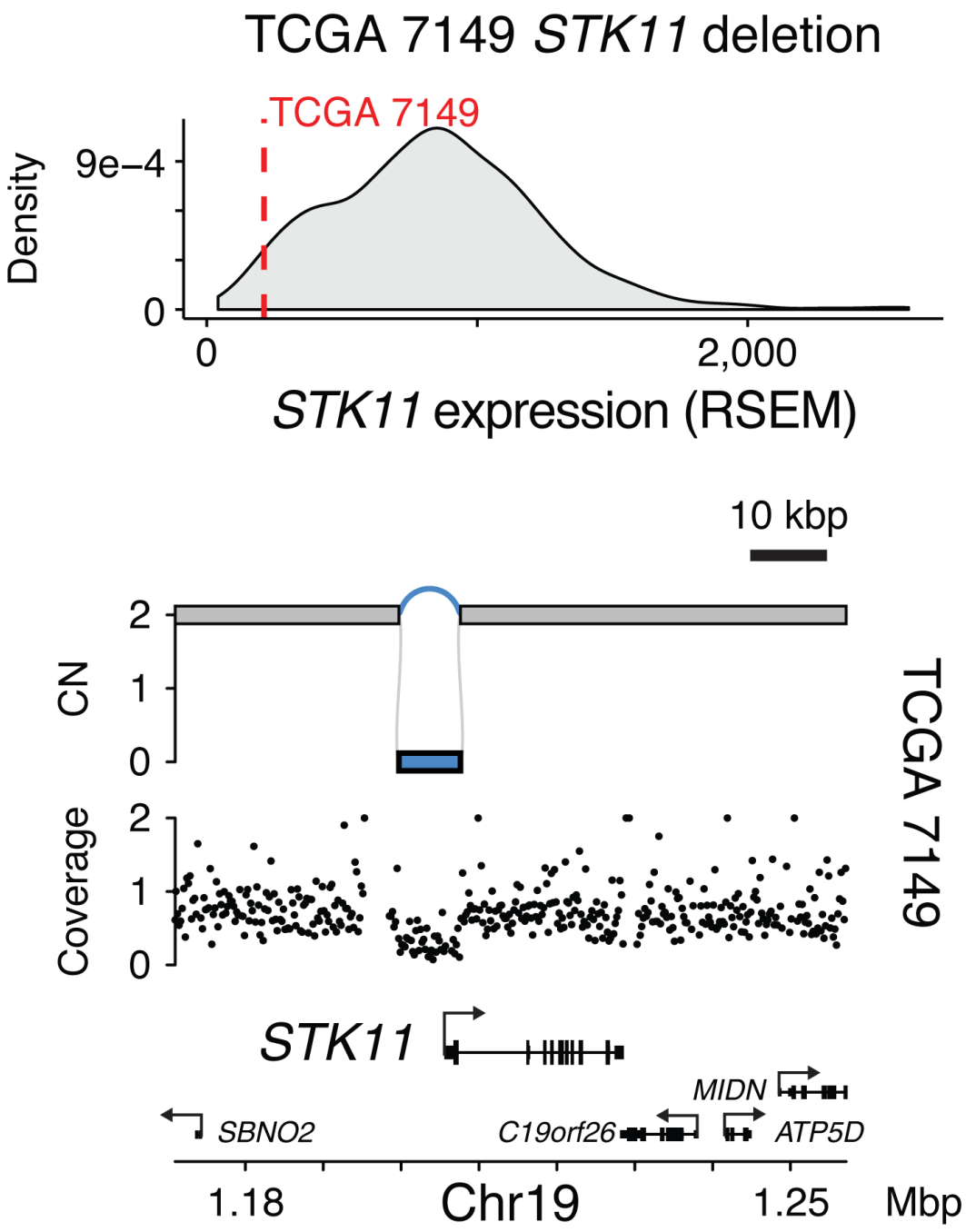

Figure 5.8: **An example of *STK11* deletion and under-expression.**

Top panel, *STK11* RSEM among RNA-seq of LUAD samples. Bottom panel, 8kbp deletion covering the transcription start site.

5.3 Recurrent coding alterations in RPA(-)$_G$ LUADs

Having identified the 57 RPA(-)$_G$ LUAD cases, we sought to define their protein-coding driver alteration landscape. Among the coding sequences of 14,987 known protein-coding genes with a median expression RSEM ≥ 1 in all of the TCGA LUAD tumors, we analyzed protein-altering indels and SNVs across 57 RPA(-)$_G$ LUADs to identify recurrently mutated genes. The algorithm used employs a gamma-Poisson regression background model [164] correcting for known covariates of LUAD mutation density (e.g. chromatin state, replication timing, GC content, see Appendix 6.5). We found four tumor suppressor genes, *TP53*, *STK11*, *KEAP1* and *SMARCA4* significantly mutated in the RPA(-)$_G$ samples (FDR<0.1, Figure 5.6). GISTIC analysis [165] of WGS-derived single copy number alterations (SCNAs) identified significantly deleted regions harboring *SETD2* and significantly amplified regions harboring *NKX2-1*, *KAT6A*, *CCNE1*, *MDM2*, *MYC*, *MCL1*, and *MYCL* (FDR<0.1, Figure 5.6). Of note, we did not identify novel genes that were significantly mutated or amplified/deleted in the RTK/RAS/RAF pathway; in other words, we were unable to identify new putative driver alterations in an obvious candidate member gene of the RTK/RAS/RAF pathway, possibly owing to the limited sample size.

Our SV analysis (see Appendix 6.5) identified simple, focal deletions (<100kbp) targeting *STK11* (Figure 5.8) and *KEAP1* (Figure 5.7), coupled with LOH, leading to the decreased expression of these genes (refer to [152] Table S2). In two cases (TCGA-86-7711 and TCGA-50-6592), the focal deletions targeted the promoter/transcription start site of *KEAP1* without altering the coding regions but resulted in reduced expression of *KEAP1* (Figure 5.7, also refer to [152] Table S2). Alterations in *STK11* in KRAS-driven LUADs have been shown

to be associated with immune exclusion and a poor response to immunotherapies [166]. We found that loss-of-function events in *STK11* to be anti-correlated with the computationally estimated fraction of leukocytes in the tumor (refer to [152] Figure S2D).

As opposed to 411 RPA(+) cases from the full TCGA LUAD cohort (including the 28 cases rescued from the RPA(-)$_E$ category by our WGS analysis), we found a significant enrichment of *TP53* mutations ($P = 5.5 \times 10^{-4}$, OR=2.97, Fisher's exact test), *KEAP1* mutations ($P = 4.3 \times 10^{-4}$, OR=3.06) and *SMARCA4* ($P = 3.4 \times 10^{-4}$, OR=4.16) in the RPA(-)$_G$ cases (Figure 5.9) for genes listed in Figure 5.6. When expanding the analysis to 239 COSMIC cancer genes, we found additional enrichment of mutations in *NRG1* (Figure 5.10, $P = 1.2 \times 10^{-5}$, OR=12.0), *ESR1* ($P = 4.5 \times 10^{-4}$, OR=11.4), *BLM* ($P = 1.1 \times 10^{-3}$, OR=12.4) and *FOXO3* ($P = 2.3 \times 10^{-3}$, OR=9.29, Figure 5.9,). The RPA(-)$_G$ samples also showed significantly higher tumor mutation burden (TMB) in a linear regression model controlled for tumor purity (Figure 5.11). We found that recent smokers (last smoking year less than 15) were enriched in the RPA(-)$_G$ group compared to the RPA(+) group, although mutagen activity associated with tobacco smoking activity (COSMIC SNV signature 4) was not different between the RPA(-)$_G$ group and RPA(+)$_G$ group, controlling for tumor purity (Figure 5.12). We did not find additional differences between RPA(+)$_G$ and RPA(-)$_G$ LUADs in their molecular/clinical features (leukocyte fraction, genome doubling, degree of aneuploidy, age of diagnosis and genetic ancestry [152].

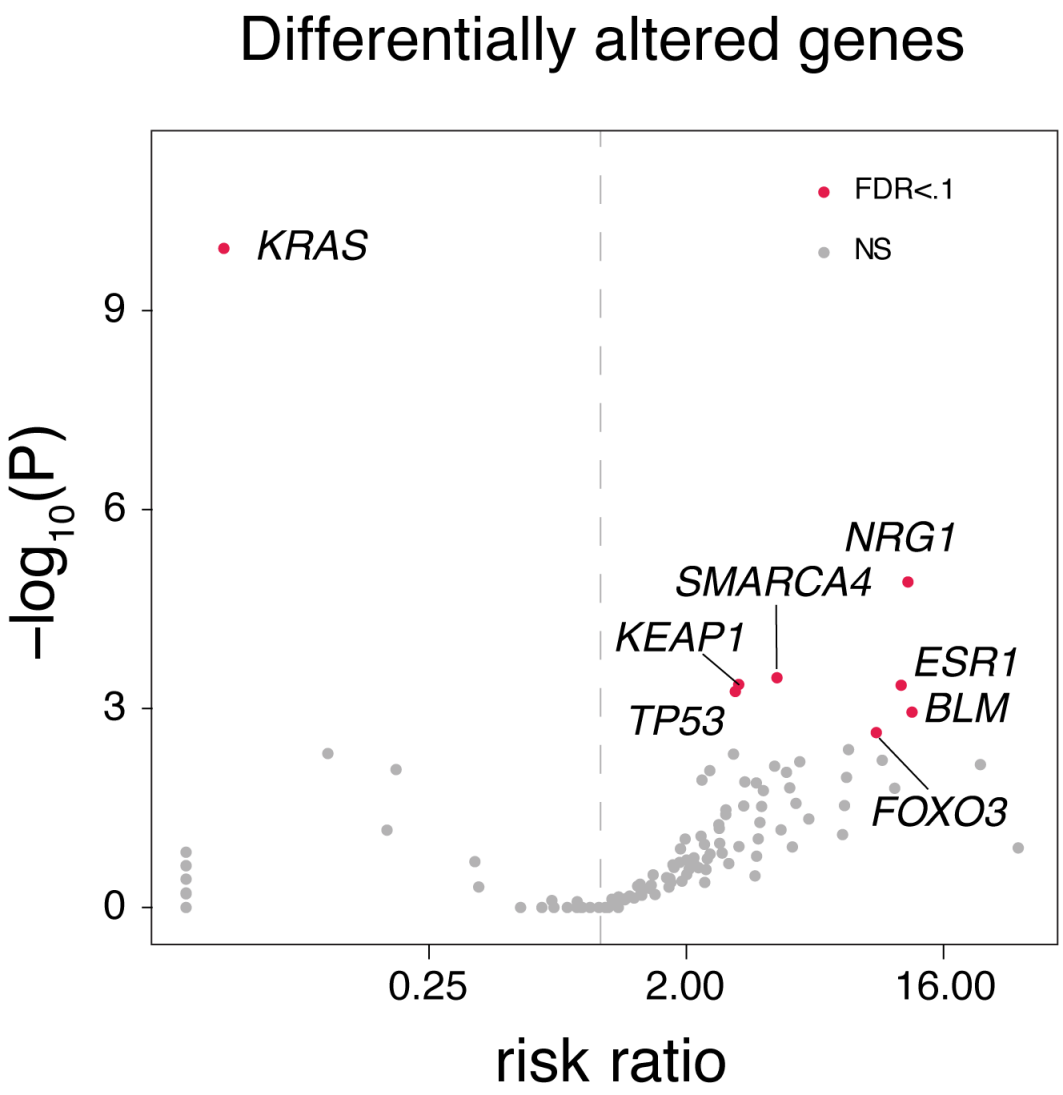

Figure 5.9: **Differentially altered driver genes in RPA(-) versus RPA(+) samples.**

P values are obtained from Fisher's exact test. NS: not significant.

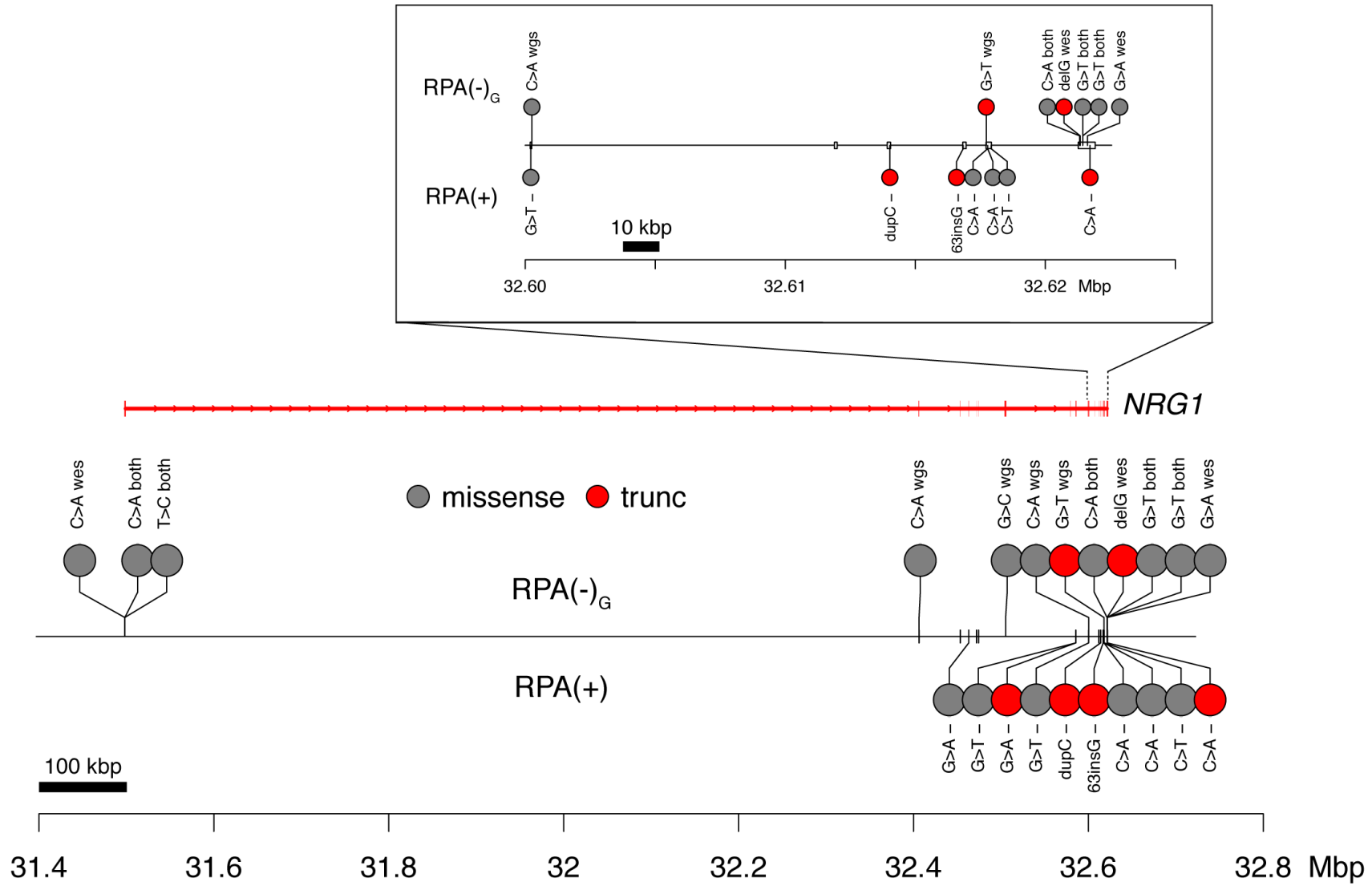

Figure 5.10: ***NRG1* mutations in RPA(-) versus RPA(+).**

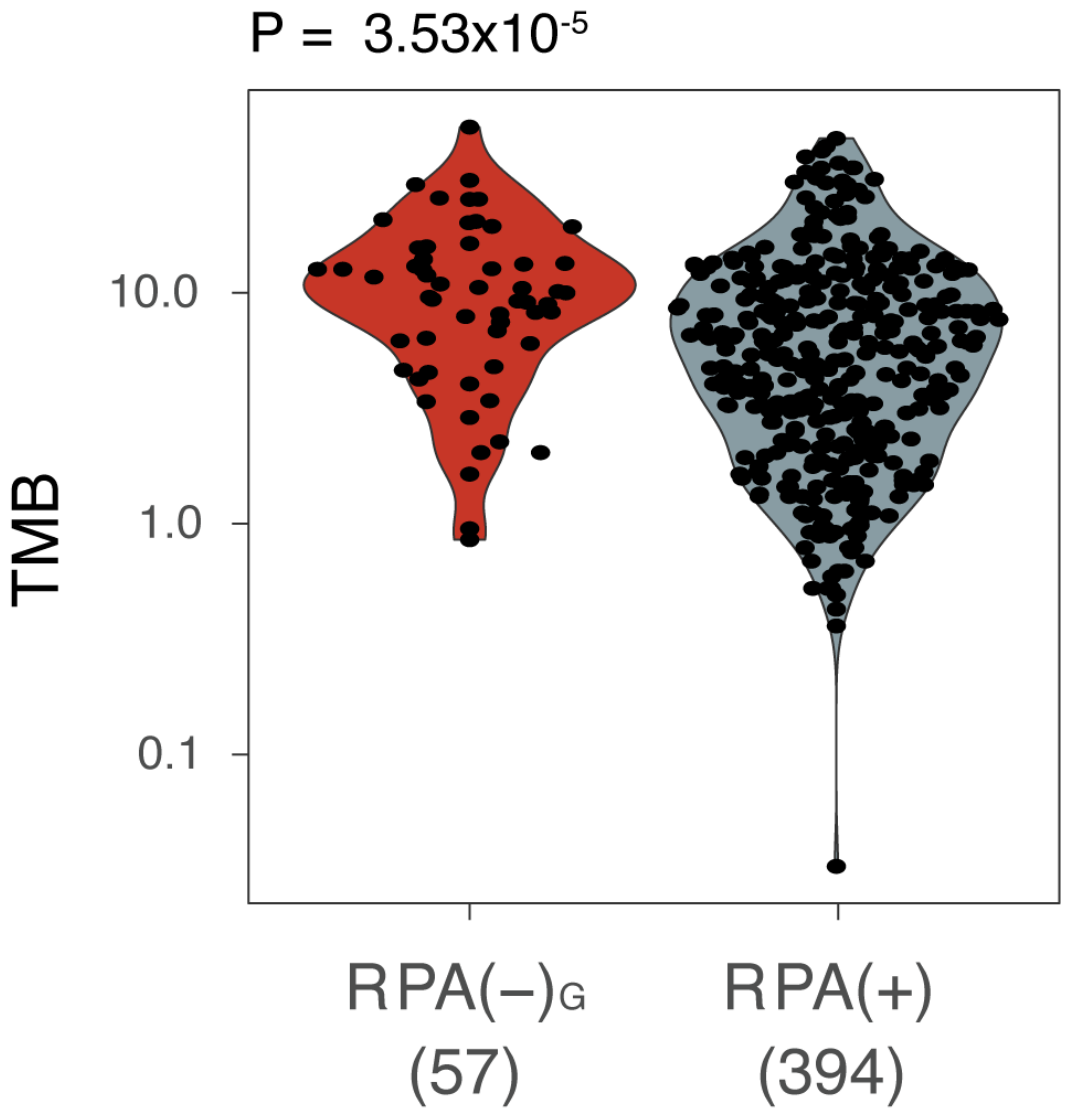

Figure 5.11: **Higher TMB in RPA(-) than RPA(+) samples.**

P value is calculated from a Wilcoxon rank test. The violin plot reflects kernel density estimations.

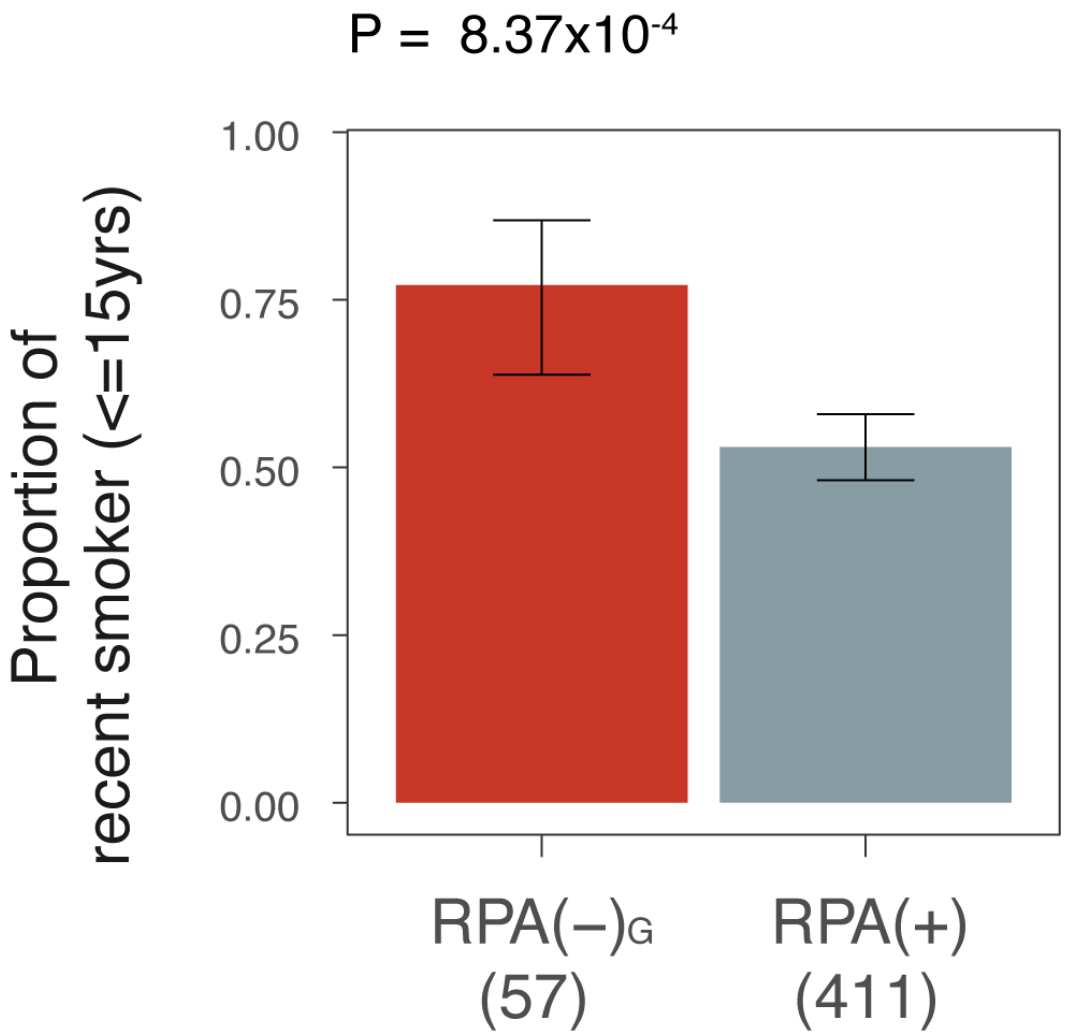

Figure 5.12: **RPA(-) enriched in patients who smoked within 15 years from diagnosis.**

P value is obtained from a two-proportions z-test. Error bars represent the 95% confidence intervals from the two-proportions z-test.

5.4 Recurrent non-coding alterations in RPA(-)$_G$ LUADs

We next asked whether the RPA(-)$_G$ cases harbored recurrent novel SNVs or indels outside of the coding genome. We focused the search on regions nominated by a recent TCGA ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) study that identified regions of open chromatin, and required that the region be identified in at least 2 of the 44 LUAD samples subject to ATAC-seq [167]. Open chromatin regions are associated with active promoters, enhancers, and transcription factor binding sites, and thus may be targets of positive somatic selection [168, 169]. When we analyzed 60,572 total mutations in 139,841 LUAD-specific open chromatin regions (2.7% of the genome) using gamma-Poisson regression with known covariates of LUAD passenger mutation density (Appendix 6.5), we found three loci that were nominally enriched (FDR<0.25) in SNVs or indels, located near *ILF2*, *CUL2*, and *TSN* (Figure 5.13). Applying the intuition that expressed and dosage sensitive genes might be targets of non-coding alteration, we examined genes that were consistently expressed across TCGA LUAD (median RSEM>10) and recurrently amplified in RPA(-)$_G$ samples (see Appendix 6.5, [124]). This analysis yielded *ILF2* as the sole significant peak (FDR<0.1, Figure 5.14).

The promoter region of *ILF2* was mutated in six out of 57 cases (P=$2.7\times10^{(}-6$), coefficient=2, Gamma Poisson Regression; Figure 5.15). We used the FunSeq2 method to annotate the sequence motifs bound by transcription factors [168]. All six mutations lay in the "sensitive" and "ultrasensitive" (i.e., highly conserved) regions of the genome [107]. One mutation (chr1:153643633, G->A) was predicted to disrupt a *HOXB6* motif and another mutation (chr1:153643690, G->T) was predicted to disrupt an *NR3C1* motif. We did not observe any muta-

tional signature enriched in the six mutations. Moreover, RPA(-)$_G$ cases harboring *ILF2* promoter mutations showed increased expression of *ILF2*, compared to ILF2-wildtype cases (Figure 5.16), or cases harboring other mutations within the +/-10kbp window of the promoter region (Figure 5.17, 5.18), after controlling for the local copy number of *ILF2* and tumor purity. On the other hand, non-coding mutations near *CUL2* did not affect *CUL2* expression, and intronic mutations in *TSN* showed a non-significant trend towards an increase in *TSN* expression (P=0.066). *ILF2* is located in chr1q21.3, which is frequently amplified in LUADs. In myeloma, increases in *ILF2* expression through amplification have been shown to promote tolerance of genomic instability and drive resistance to DNA damaging therapies, through dysregulation of RNA splicing and DNA damage response pathways [170]. Consistent with that role of *ILF2*, we found *ILF2* expression to be associated with increased SV burden (P=0.01, coefficient=0.9, negative binomial regression).

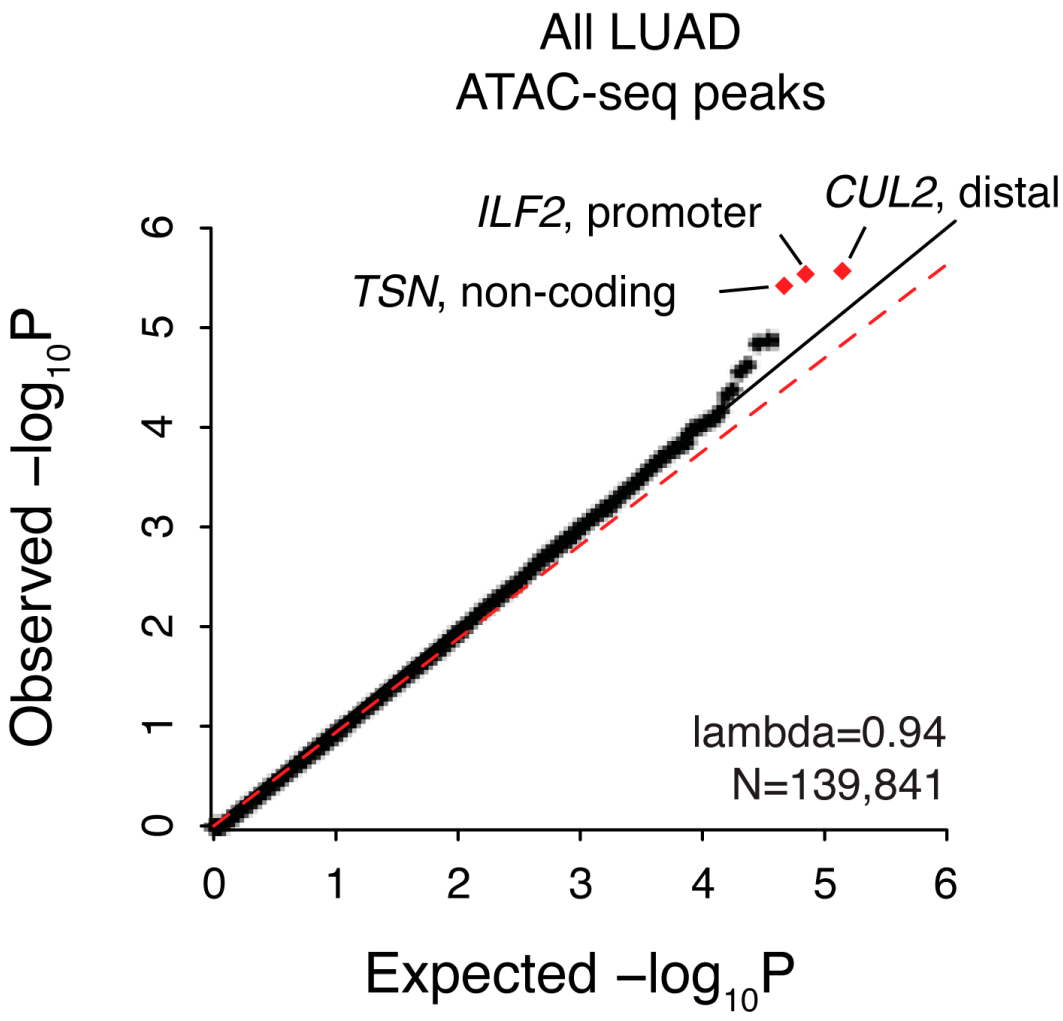

Figure 5.13: **Recurrent non-coding mutations within LUAD-specific ATAC-seq peaks in RPA(-)**

Three genes with non-coding mutations nominated through recurrence analysis across LUAD-related ATAC peaks. Red dots indicate loci with FDR < 0.25.

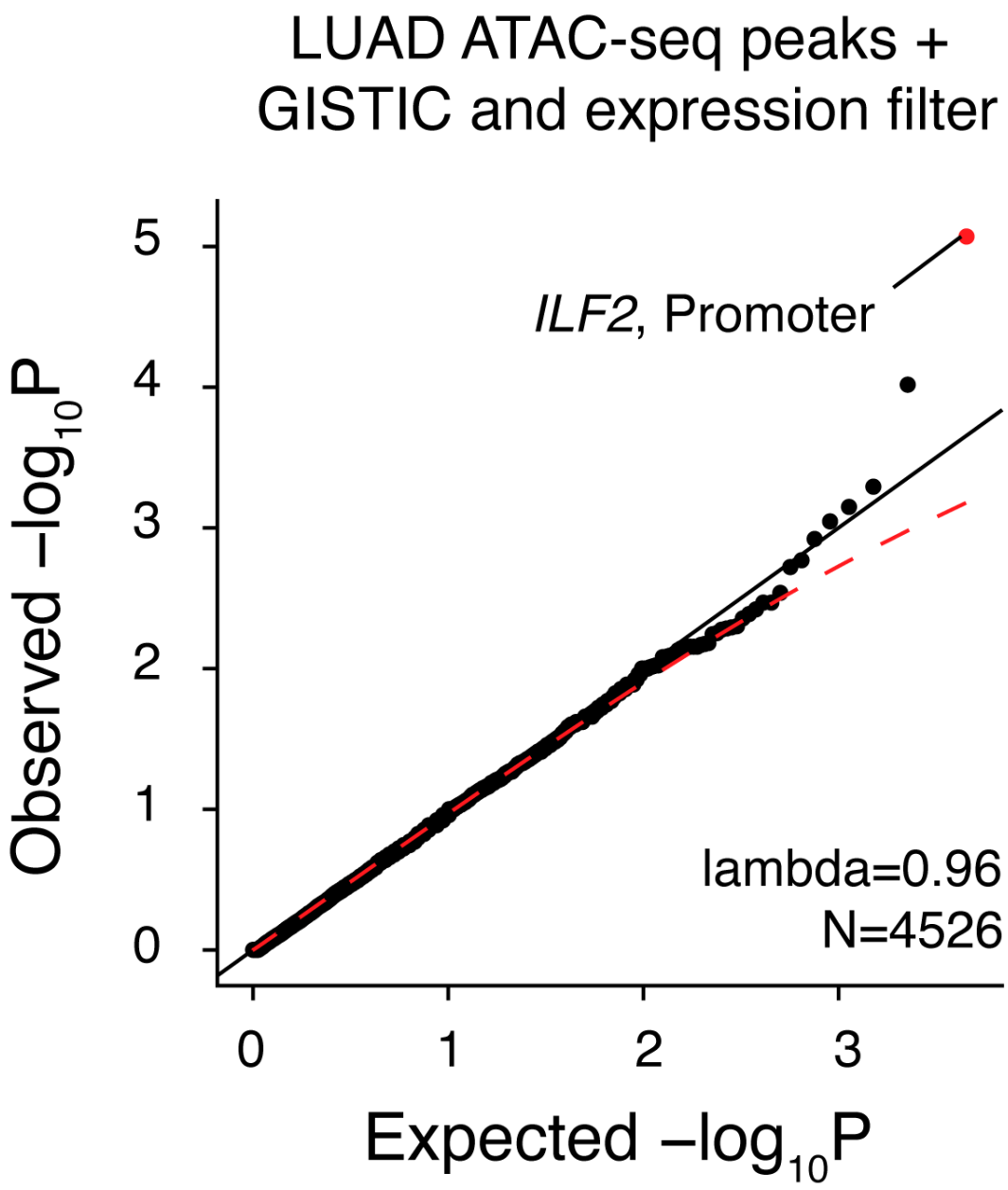

Figure 5.14: **Recurrent non-coding mutations within the intersect of LUAD-specific ATAC-seq peaks and LUAD-specific recurrent SCNA peaks in RPA(-)**

Same as Figure 5.13, but restricted to ATAC-seq peaks in genes with RSEM $\geq$ 10 across TCGA LUAD and recurrently amplified in RPA(-)$_G$ samples. Red dots indicate FDR < 0.1.

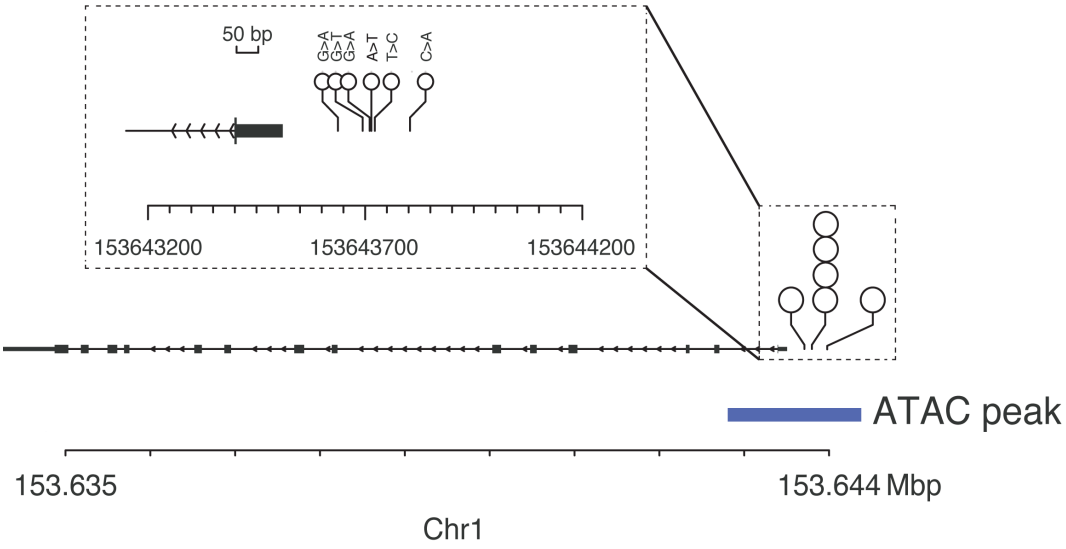

Figure 5.15: **Recurrent promoter mutations near ILF2 gene**

Among 57 RPA(-)$_G$ samples, 6 SNVs are observed in the promoter region of *ILF2*; all are located within ATAC-seq peaks.

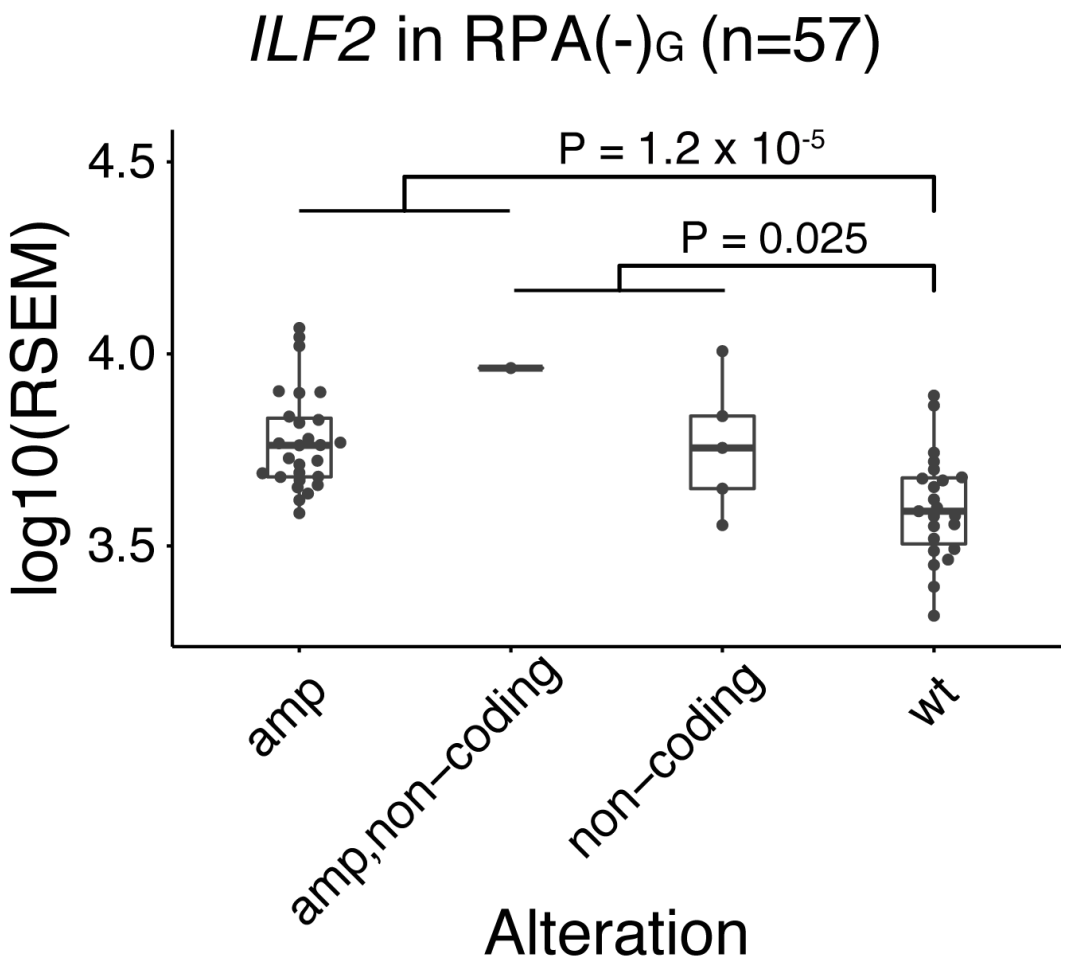

Figure 5.16: **Promoter mutations and amplifications both increase *ILF2* expression compared to wildtype.**

P values are calculated from linear regression analysis correlating expression, adjusting for the local copy number of *ILF2* and purity. Boxplot shows median, interquartile range, and 1.5 times the interquartile range.

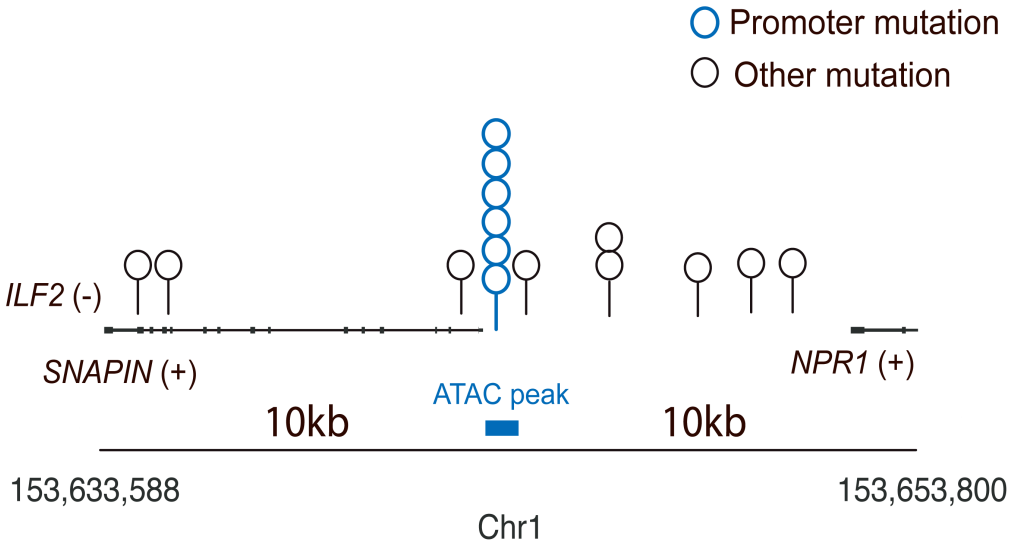

Figure 5.17: **Mutations around 10kb window from the *ILF2* promoter peak.**

Nine other SNVs among 57 RPA(-)$_G$ LUAD samples are observed in the 10kbp+/- window outside of the promoter region of *ILF2*. Promoter mutations are indicated in blue circles. Other mutations are indicated in black circles.

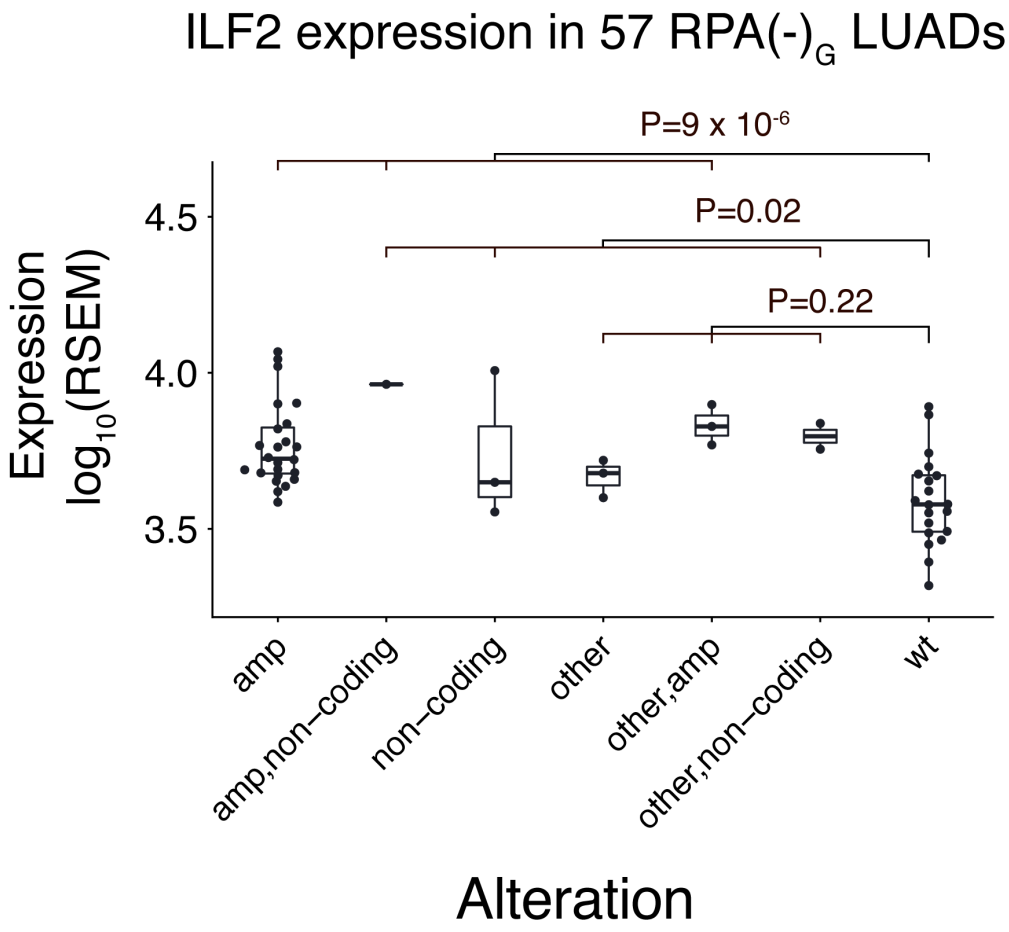

Figure 5.18: **Expression of *ILF2* with respect to alterations.**

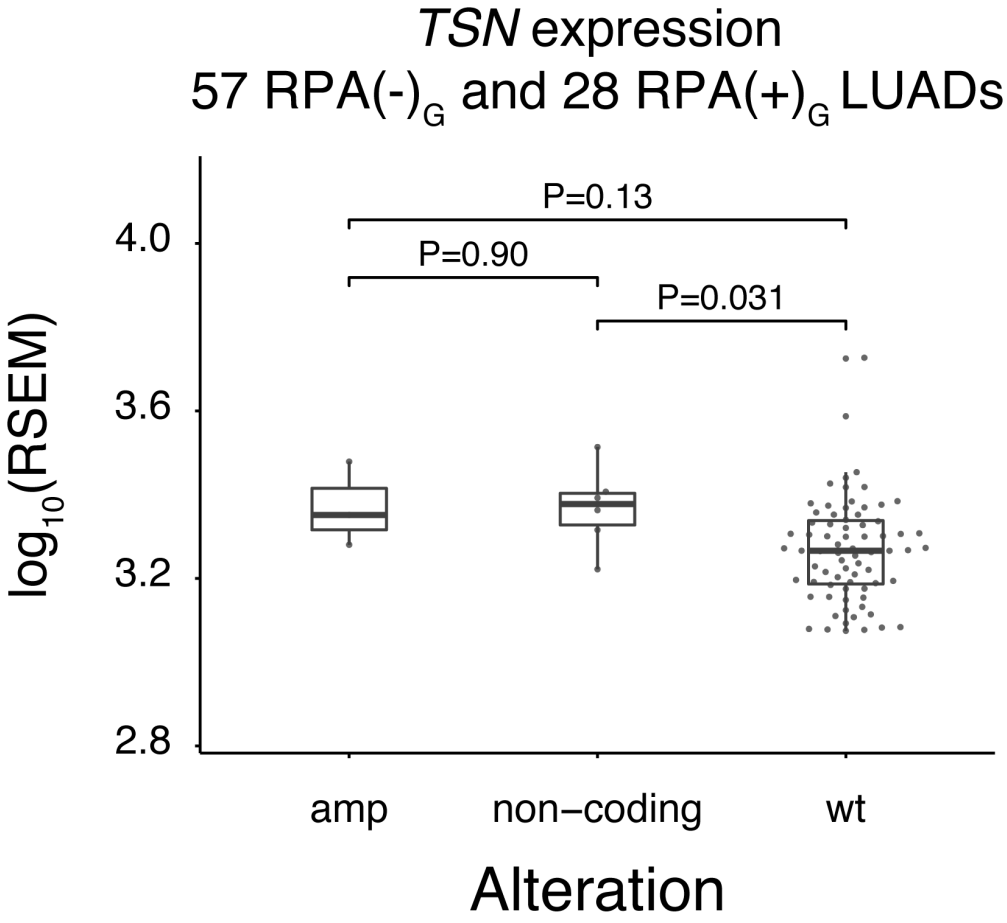

Figure 5.19: **Expression of *TSN* with respect to alterations.**

5.5 Complex SV patterns in RPA(-)$_G$ LUADs

By integrating read depth changes with rearrangement breakpoint locations to generate junction-balanced genome graphs (JBGG, see Subsection 2.1.5), we analyzed genome graphs [83] of the 57 RPA(-)$_G$ cases to identify complex patterns of Structural variant. Analysis of subgraphs (https://github.com/mskilab/gGnome) identified multiple instances of complex amplicons (18 double minute, 5 BFB cycles, 5 tyfonas, 52 pyrgo), as well as simple duplications (mean=16.3 per sample, Figure 5.20). Tyfonas are recently identified complex amplicon comprising hundreds of high junction copy number (JCN) and fold-back inversion junctions [83] that are enriched in cancers such as dedifferentiated liposarcomas and acral melanomas. As an example, we show an amplification of *NKX2-1* driven by tyfonas (Figure 5.22). Pyrgo, which comprise "towers" of low copy duplication junctions [83], were also found to drive the amplification of LUAD loci including *NKX2-1* (Figure 5.23).

Genes located inside double minute, BFB cycles, and tyfonas events were markedly enriched in expression outlier genes ($P < 1 \times 10^{-16}$, Mann-Whitney U test) relative to genes involved in pyrgo and simple duplication events (Figure 5.26), suggesting that the former events were retained in the cancer cell due to the growth-promoting effects of altered gene expression. Although none of these complex SV types were correlated with *TP53* mutations, which are thought to generate genomic instability, there was a significantly higher incidence of simple deletions observed in the *TP53*-mutant cases ($P = 1 \times 10^{-4}$, Mann-Whitney U test, Figure 5.21). That association held true when including the 68 RPA(+)$_G$ samples and controlling for purity and RPA status ($P = 2 \times 10^{-4}$, coefficient=12, linear regression).

Double minutes were the most common complex SV type seen in the RPA(-)$_G$ cases (12 of 57 samples, Figure 5.20). Like extrachromosomal circular DNA segments, double minutes do not segregate symmetrically; thus, their dosage per cell is exquisitely responsive to selection pressure [45, 38]. As a result, at least one of the genes in any given double minute likely contributes to tumor development. To leverage that intuition, we focused on a relatively small double minute identified in case TCGA-55-5899 (Figure 5.24). We found that this double minute fused and amplified multiple focal regions on chromosome 13 spanning 1.0 Mbp, and resulted in the high-level gain (>10 copies) of three intact genes (*UBL3*, *SOX21*, and *LIG4*). Two of these, *UBL3* and *LIG4*, were over-expressed in TCGA-55-5899 relative to the full LUAD cohort with RNA-seq data (Figure 5.25). Because we did not observe any genes to be amplified and over-expressed in more than one RPA(-)$_G$ case (refer to [152] Table S3), larger numbers of cases would have to be analyzed to gain an understanding of the possible role genes amplified by double minutes play in driving RPA(-) LUADs.

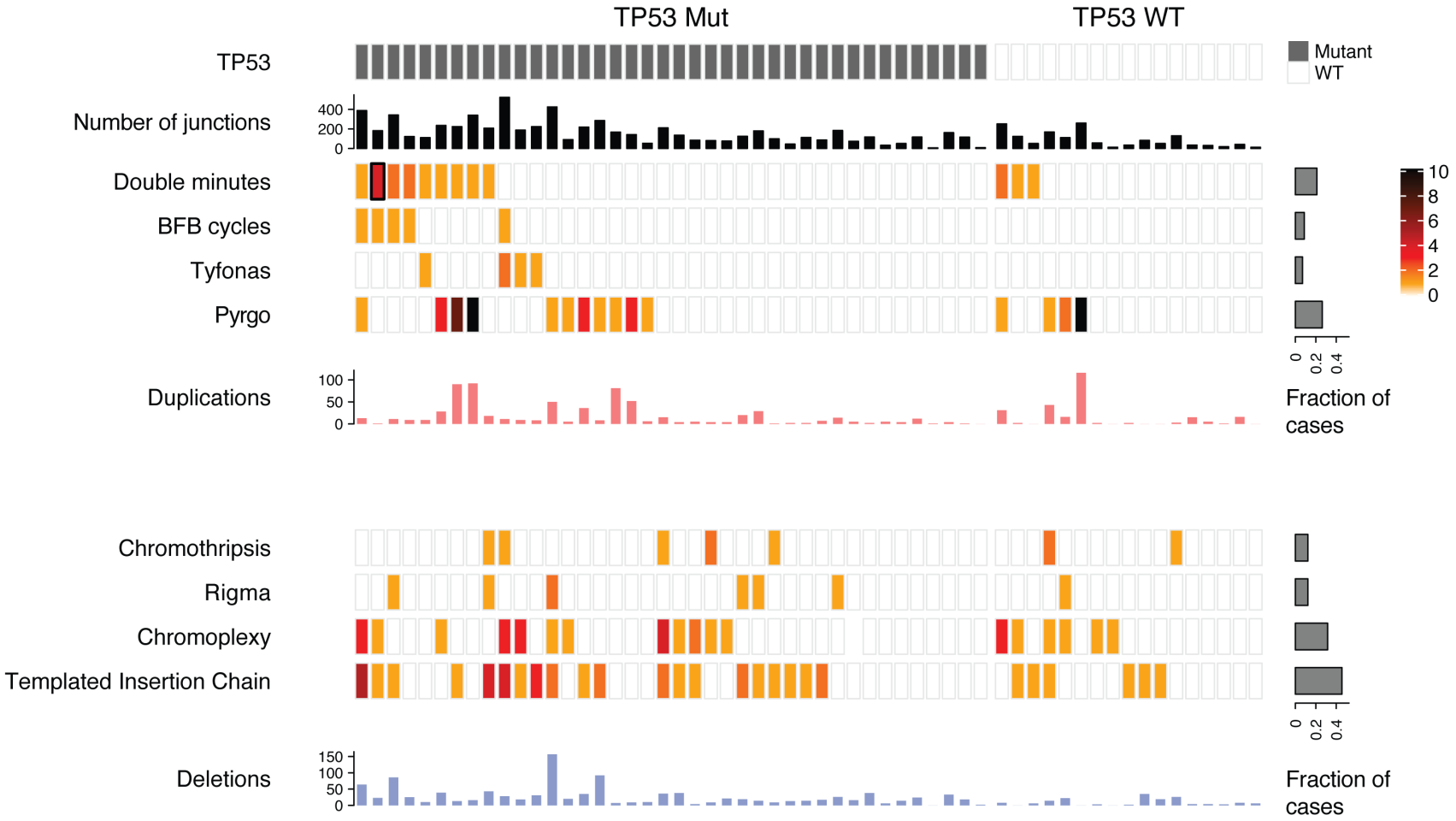

Figure 5.20: **Burden of simple and complex SV events in RPA(-) samples.**

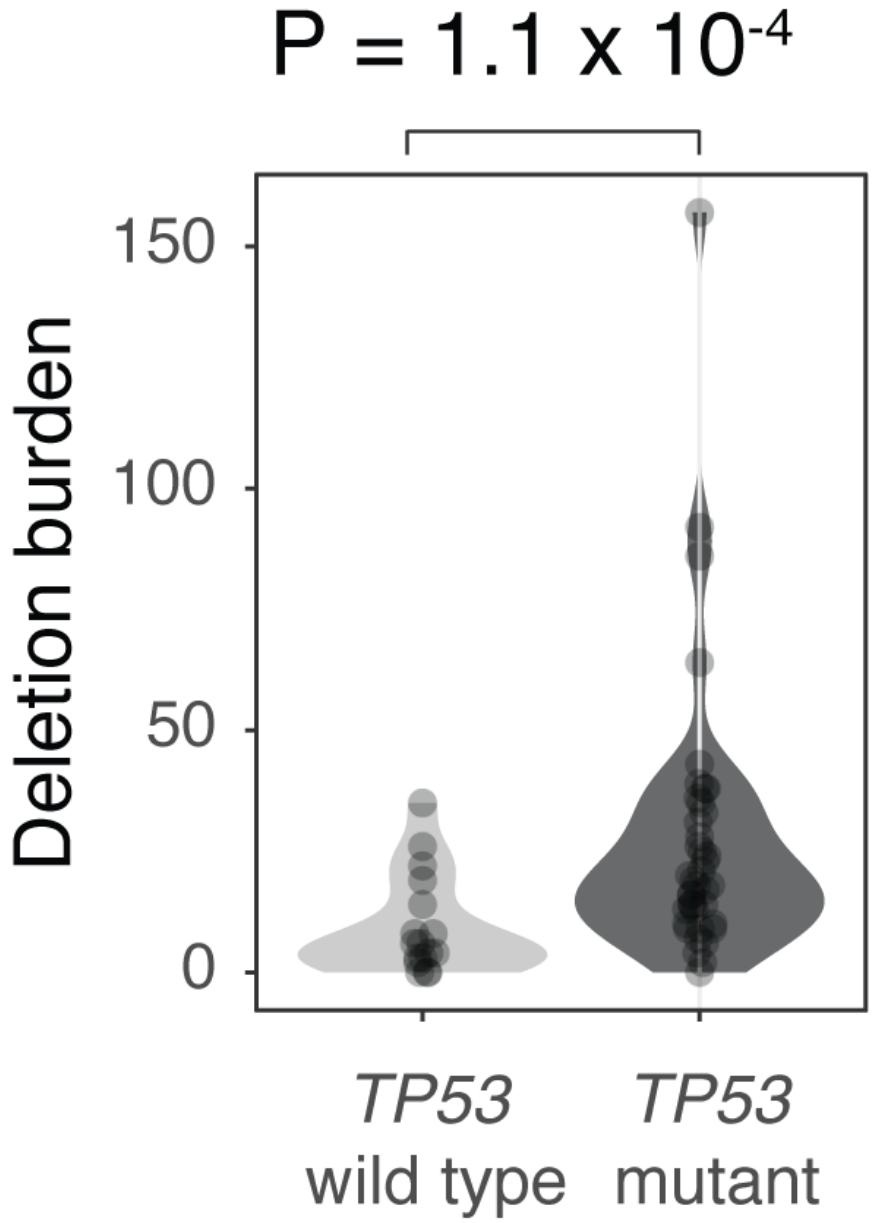

Figure 5.21: **Excess deletion burden associated with *TP53* loss of function alterations.**

P value is obtained from Mann-Whitney U test. Violin plots reflect kernel density estimations.

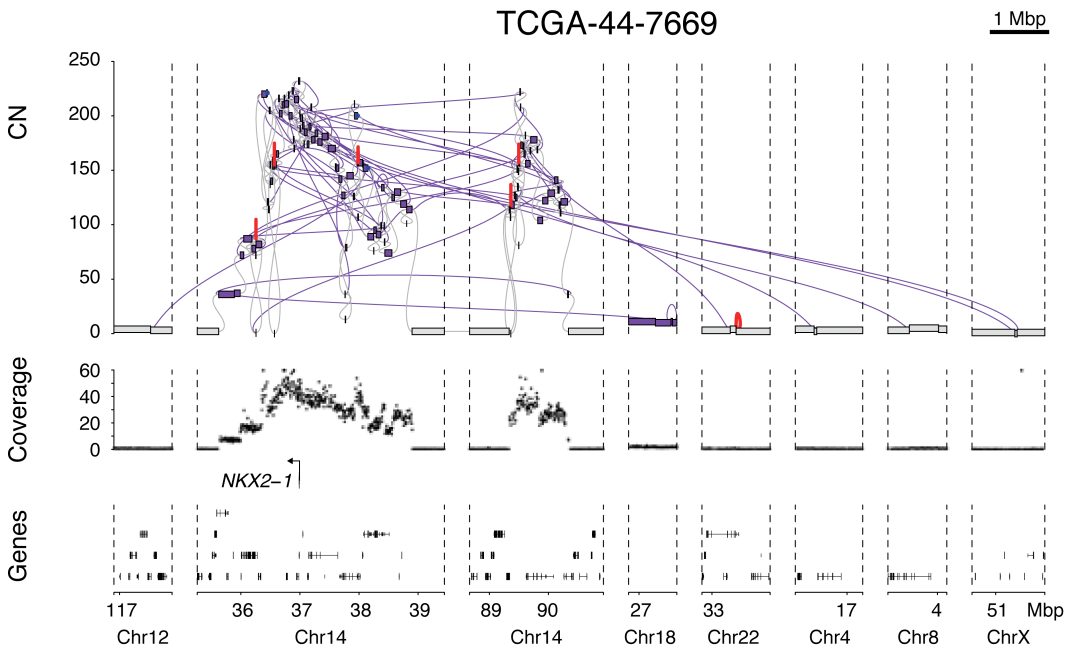

Figure 5.22: ***NKX2-1* amplification by tyfonas.**

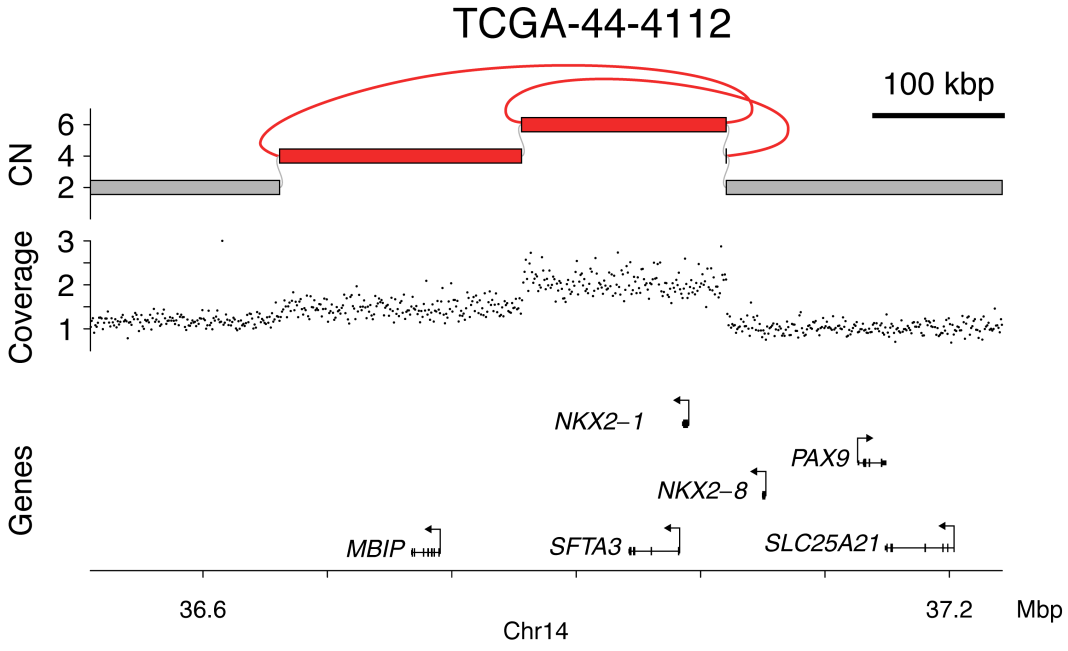

Figure 5.23: ***NKX2-1* amplification by pyrgo.**

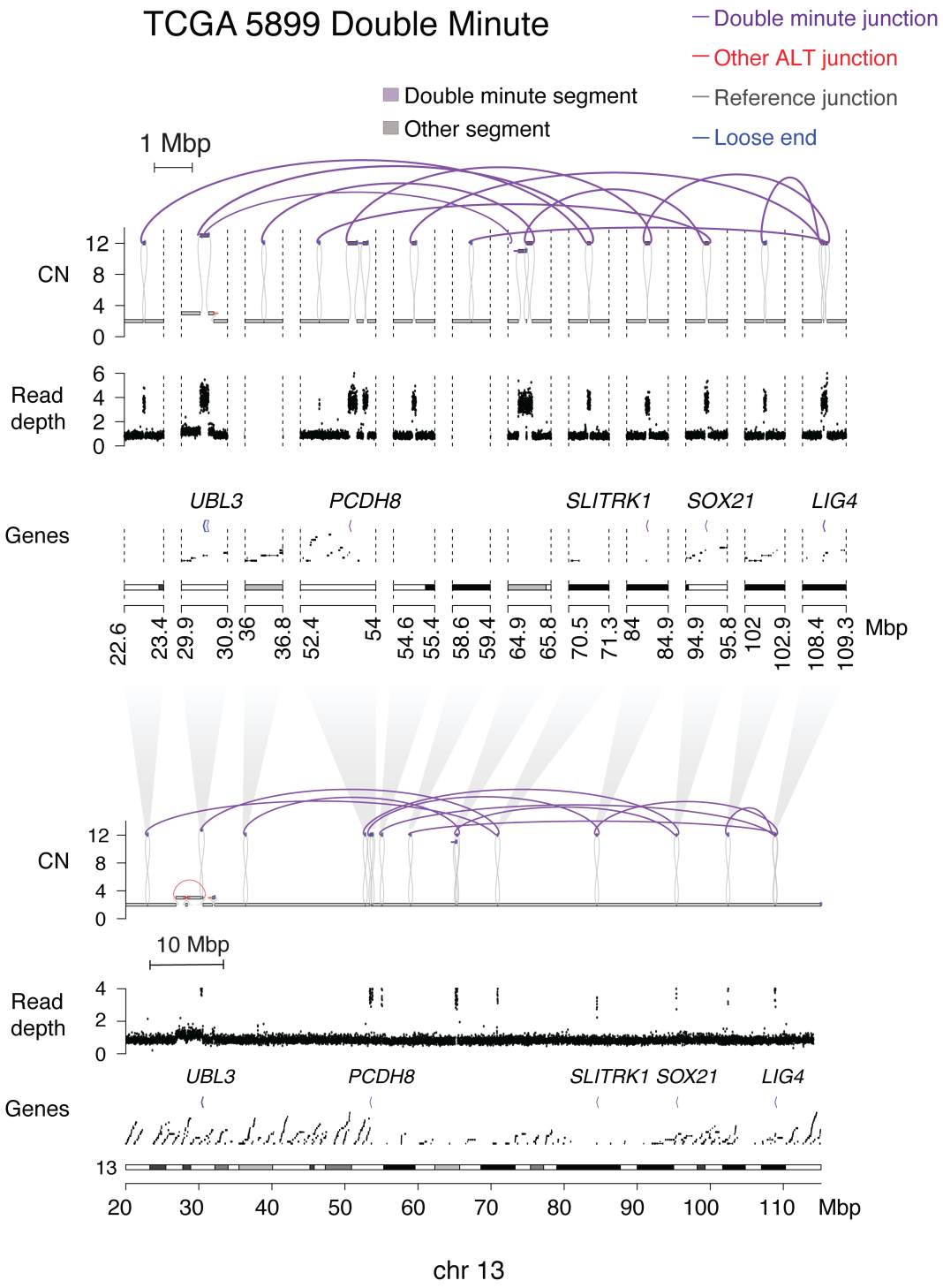

Figure 5.24: **An example of double minute (DM) amplifying multiple distal loci.**

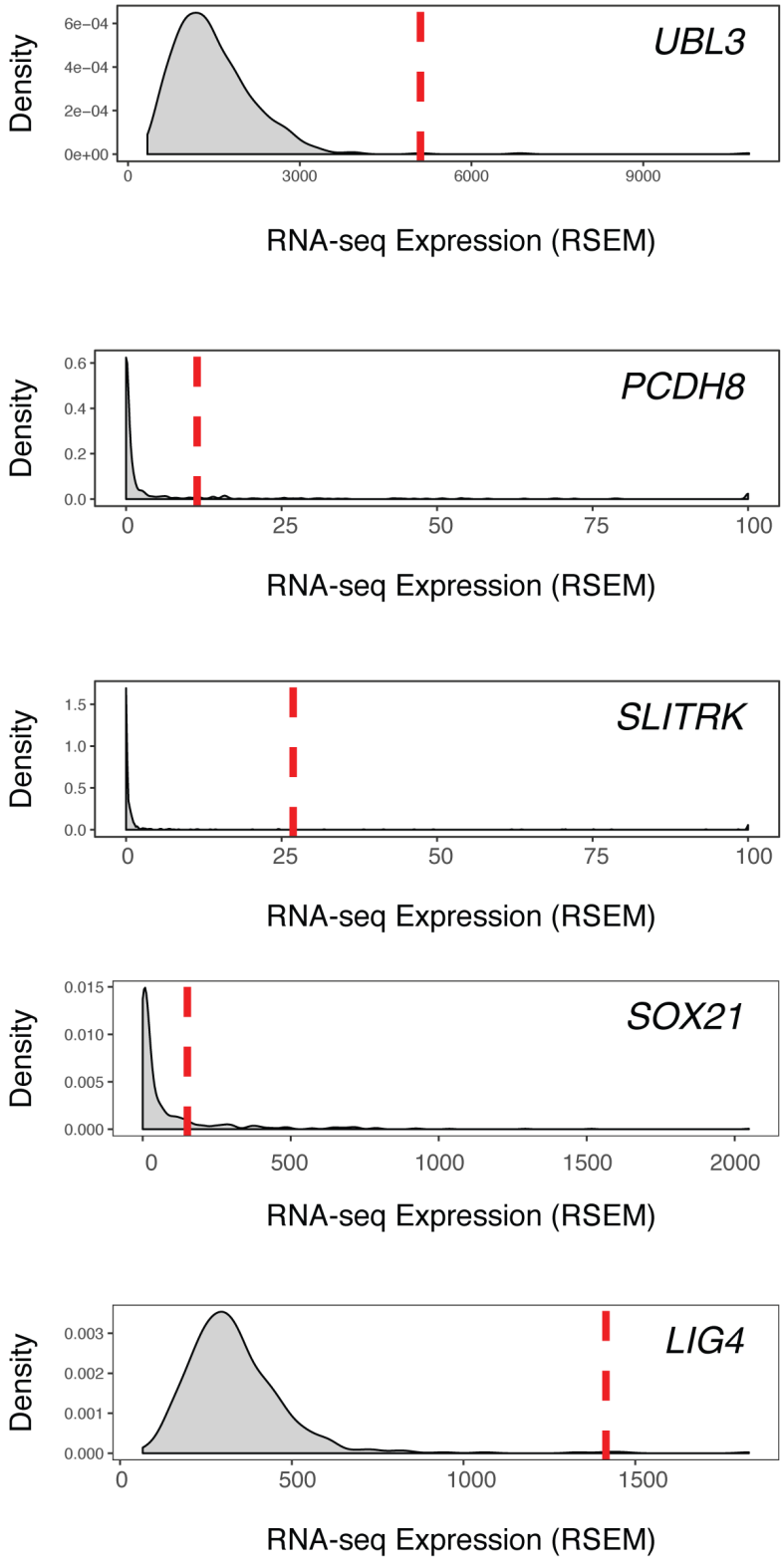

Figure 5.25: **Over-expression of genes amplified in the DM.**

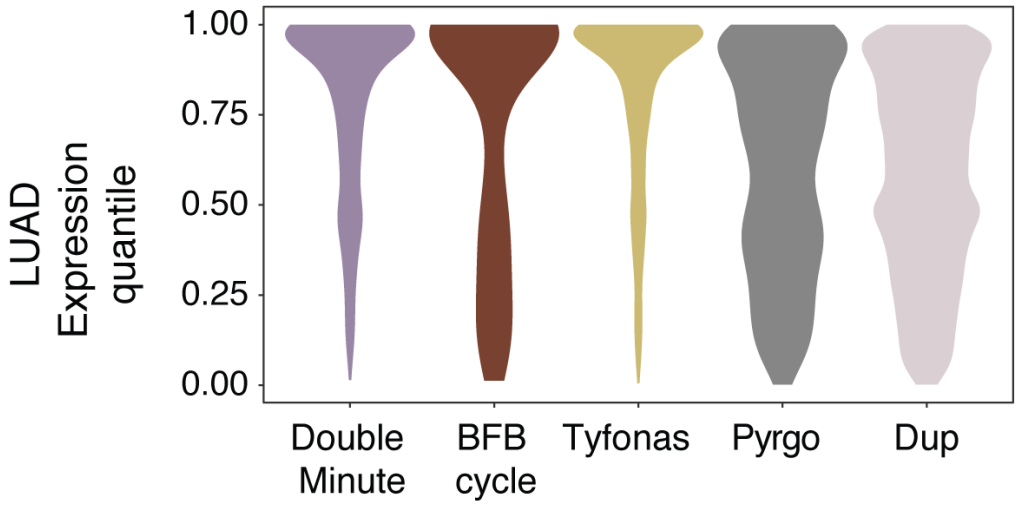

Figure 5.26: **Over-expression of genes amplified by any of the three complex amplification events.**

5.6 Discussion

Although large-scale genomic studies have shown LUADs to be molecularly heterogeneous, the majority of LUAD cases share the common feature of RTK/RAS/RAF signaling activation through a genetic driver [154, 171]. A key question, however, is whether RPA(-) LUADs – which, by definition, lack RTK/RAS/RAF drivers – represent a biologically distinct entity. Our results suggest that they are heterogeneous but that they do share common biological features, including a high frequency of *TP53* mutations and high mutation burden. Those features tend to distinguish RPA(-) LUADs from their RPA(+) counterparts.

A key confounding factor when we try to define RPA(-) LUADs as a distinct entity is the technical limitation of reliably detecting RTK/RAS/RAF pathway genomic lesions in impure tumor samples. Strikingly, 28 of 85 cases in the present study that were previously found negative for any RTK/RAS/RAF lesion by WES and RNA-seq pipelines were subsequently shown by our WGS analysis to harbor a somatic RTK/RAS/RAF driver. The high prevalence of overlooked *KRAS* mutations is explained in part by low tumor purity and/or decreased probe affinity for the GC-rich exons in *KRAS* during WES library preparation [160].

The recent discovery of small molecules with activity against *KRAS* p.G12C LUAD [158, 159] highlights the importance of precise identification of mutations at that locus. However, many of the challenges could potentially be overcome by higher read depth (e.g. >40X minimal on-target coverage), which is now routinely achieved by some clinical-grade target capture assays [172, 173]. Nev-

ertheless, the high rate of missed RTK/RAS/RAF lesions in our WGS cohort tells a cautionary tale: false-negative calls should always be considered even when working with high-quality datasets.

Eight of the 28 rescued RPA(+)$_G$ cases in our WGS cohort harbored cryptic SV lesions (protein-coding fusions, SCNAs), which represent alternative mutational mechanisms for (in)activating RTK/RAS/RAF genes (e.g. *EGFR* BFB cycles, focal deletions of *NF1* and *RASA1*). WGS is naturally adapted to detect such complex or subtle structural alterations [83]. We leveraged that capability to identify a spectrum of SV patterns among the 57 RPA(-)$_G$ cases. Notably, 9/85 (11%) samples in the cohort harbored focal deletions in *STK11* that were undetected by WES. Alterations in genes such as *STK11* and *KEAP1* have come into focus as possible prognostic and/or predictive biomarkers in patients with lung cancer [174, 166], inclusion of full genomic capture probe sets for those genes may become necessary in the near future to identify samples with alterations accurately. Double minutes, the most prevalent SV type among complex SVs in the RPA(-)$_G$ cases have recently been implicated in genomic plasticity, oncogene selection, and chromatin evolution [45]. Further studies analyzing larger cohorts will be valuable for dissecting the role they may play in driving RPA(-) LUAD biology in greater detail.

If the 57 RPA(-)$_G$ cases identified in this study indeed represent a distinct biological entity of RTK/RAS/RAF-independent LUADs, then what pathway drives them to proliferate? Do RPA(-) LUADs show distinct genetic or therapeutic vulnerabilities? Although our analyses have nominated candidate drivers in LUAD (e.g. *ILF2*), it is unclear how tumors harboring an alteration in such genes would phenocopy the proliferative effect of RTK/RAS/RAF alteration.

Perhaps the frequent SV-driven loss of tumor suppressors (e.g. *STK11*) or amplification of genes operating downstream of RTK/RAS/RAF signaling (e.g. *MYC*), which we observed in our RPA(-)$_G$ LUADs, can cooperate to fill the missing role [175]. It is also possible that a small subset of RPA(-) tumors are still RTK/RAS/RAF-driven but activate the pathway through epigenetic dysregulation, or genetic alterations in genes that are less frequently altered in LUAD, and therefore not detected by the statistical methods employed in this study. Alternatively, the biology of RPA(-) LUADs may resemble that of other cancer types in which RTK/RAS/RAF alterations are rarely seen and are marked by early *TP53* loss and high TMB; those tumor subtypes may take a different evolutionary path that is less dependent on the sustained proliferative signaling that RTK/RAS/RAF activation provides [176, 177, 178]. Such a path may accumulate key genetic alterations in a different order than RPA(+) tumors [117], begin in an alternate cell of origin, or undergo lineage switching in the course of evolution.

Taken together, our findings suggest that RPA(-) LUAD is likely to represent a heterogeneous entity and that WGS of much larger cohorts of RPA(-) and RPA(+) LUADs would be necessary to fully address the nature of their underlying biology.

CHAPTER 6

CONCLUDING REMARKS

Merely over a decade young, genomic sequencing has already revolutionized our approach to study cancer and will surely continue to grow, fueled by the ever decreasing sequencing cost and the pressing need to optimize clinical benefits [5]. The study of somatic variant patterns has delivered ample successes in elucidating genome maintenance mechanisms and translating to clinical utility. The extraordinary complexity of SVs in cancer genomes pose the next major analytical challenge. In this thesis, I have shown the definition, implementation, and inference of genome graphs for analyzing complex SVs. I have used it to categorize complex amplicon, stratify pan-cancer patients, and reconstruct exact SV evolution in cells surviving telomere crisis. I have also demonstrated its utility in whole genome characterization of cancers without known driver alterations. As discussed in each chapter, all these parts have room for improvement, but the current results still suffice to show the potential of such a general, flexible, and powerful analysis paradigm. Looking beyond these results, in this concluding chapter, I draw up 5 future areas in which this genome graph paradigm will make great impact: 1) systematic discovery of SV signatures, 2) resolving complex haplotypes by integrating emerging long-molecule profiling technologies, 3) dissecting the intra-tumoral heterogeneity of SVs, 4) reconstructing the evolutionary trajectory of SV mutagenesis, and 5) eventually translating to applications in clinical sequencing.

6.1 Systematic discovery of SV classes and signatures

Mutation signatures have been instrumental in today's cancer genomics research as they provide an interpretable summary of the footprints of all mutagenesis pathways active in the life history of the tumor sample. As we have discussed in the introduction (1.1), the extraction of mutation signatures can be seen as finding latent combinations of distinct variant classes from an observed count matrix.

While there are exciting new algorithms that can better extract signatures given such a count matrix, especially the ones that balance the abundant SNVs and rarer SVs [66] and ones that integrate genomic covariates [16], the embedding of SVs into a system classification is still the utmost limiting step, due to their intrinsic complexity. Such a system would cover the complete spectrum of SVs and have clear distinction between variant classes. To achieve this, we must ask the fundamental question: are there more distinct classes of SVs? It spawns at least two major inquiries: 1) developing methods to nominate SV classes from existing WGS data, and 2) profile new samples that would most likely reveal new patterns.

Thus far there have been two systematic methods to codify complex SV patterns. The *top-down*, like I have demonstrated in Chapter 3 and [83], takes whole genome graphs, breaks down to clustered subgraphs, extracts various relevant summary statistics from candidate subgraphs, and let the groups of similar recurrent patterns emerge unsupervisedly. On the other hand, the *bottom-up* approach like described in [41] enumerates all possible configurations of junctions and copy number resulting from a given number of junctions, then search

against the observations in pan-cancer whole genomes to quantify their abundances. They both have advantages and limitations of their own.

The top-down approach is hollistic and allows similar yet not exactly isomorphic subgraph candidates to group together, but relies on careful choice of summary statistics based on human knowledge. The bottom-up approach is precise and exhaustive, but the number of potential motifs grows super-exponentially with the number of junctions, which is why [41] only builds up to 5 junction patterns. However, more complex patterns are prevalent and not all can be broken down to the current repertoire. Plus, as we have explored in [44], there are blindspots of short-read sequencing where a junction breakend cannot be accurately mapped, and these loose ends require the method to consider incomplete matching. More importantly, as we have demonstrated throughout this thesis, there are many complex SVs that result from a mixture of different processes. This not only means different processes affecting the same locus and interfere each other's classification, but also naturally compound processes that are multi-step. For example, in our proposed model of tyfonas, there are initial phases of chromothripsis-like shattering, intermediate BFB-like extrachromosomal amplification, as well as final stabilization by fusing new chromosomal ends. These nuances should not be collapsed under one term if we want to precisely reflect the distinct processes underlying the formation of the final outcome. There are also patterns during initiation stages of a process like the SV evolution immediately following telomere crisis 4, where a subgraph clearly resembles a chromothripsis or BFB yet because it is not as extensively rearranged as commonly seen in tumor genomes, they do not pass the current threshold to be formally classified. Hence, the next method would ideally be able to be sen-

sitive to the hallmarks of distinct processes, flexible to the scale of the events, and tolerable to the incompleteness and errors of real data.

In essence, from the perspective of rearranged genomes as a graphs, the task of identifying distinct yet recurrent SV patterns is equivalent to network motif mining [179]. It was first introduced in [**?**], as "patterns of inter-connections occuring in complex networks at numbers that are significantly higher than those in randomized networks", and motifs of up to three vertices were exhaustively searched in cellulr signaling and ecological food webs generating new insights into the structures of such complex systems. However, a general algorithm for exhaustive and exact motif mining is inhibitively expensive [180], as there are at least two computationally intensive steps in any such algorithms: subgraph searching and motif matching. Given a target motif, locating a potential occurence of it in the input graph gives a search space that grows exponentially with the size of the input. Subsequently, each candidate subgraph needs to be matched to the target motif, which is a known NP-hard problem [181]. Moreover, because of the aforementioned inevitable incompleteness of real graph data, an approximate matching method is necessary, where structurally resembling yet non-isomorphic candidates can be assigned to the same class.

These challenges have been tackled with methods allowing strictly defined fuzzy rules, such as varying node and edge labelings [182] or fixed number of edge loss [183]. Promisingly, with the recent rapid advancement in graph representation learning, which learns lower dimension embeddings of graphs, [179] has formalized the motif mining problem from the intensive search and match into more efficient *approximate motif mining* problem. They proposed and implemented *motiFiesta* algorithm, which iteratively collapses connected nodes

so that intermediate nodes represent subgraphs in the original. The selection of the next nodes (subgraphs) to collapse is based on structural similarities and concentration. It was shown to recognize similar and recurrent motifs (e.g. star, clique, barbell) efficiently on large datasets. Such an algorithm sheds light on applying graph representation learning techniques to genome graphs as well.

In terms of new samples, it may seem granted the number of WGS-profiled tumors keep growing in the near future. The completion of PCAWG marked an important milestone of whole genome profiling of primary cancers [2], that for the first time, all variants including SVs from various cancer types are harmonized and available for a systematic analysis. More recently, there has been a study of substitution mutation signatures in a cohort of more than 12000 prospectively collected new pan-cancer WGS samples [184], the largest to date. Yet, there is plenty to consider on what type of samples to sequence next, with what type of technology (discussed in the next sections), to optimize the gain of insights and clinical benefit.

First of all, metastatic cancers are the leading cause of deaths in patients, and can rise out of therapies and surgeries. The Hartwig Medical Foundation profiled the whole genomes of 2520 solid metastatic tumors and did not find striking gene that is specifically associated with metastasis not primary cancers. find much higher whole genome doubling rate than reference primary tumors plus wide-spread CNAs [185]. Moreover, previous WGS studies and experimental systems have pointed to complex amplifications as a prevalent mechanism to sustain mitotic growth or gain therapeutic resistance [110, 186, 187, 29]. For instance, in one study [187], more than half of metastatic castration-resistant prostate carcinomas harbor high CN amplification of androgen receptor gene

(*AR*). In a experimental system, Shoshani et al. [29] found HeLa cells respond to increasingly aggressive methotrexate treatment by shattering and forming extrachromosomal DNA at the previously mildly amplified homogeneous staining regions covering the resistence gene *DHFR*. The aggresive phenotype and the strong selection effect from prior treatments makes metastatic tumors a promising source to learn more about the mutagenesis pathways of SVs.

Secondly, pediatric tumors have shown very distinct driver landscape from adult tumors [188] and numerous reports have pointed to complex SV events producing putative drivers like *MYCN* amplification and *PRDM6* enhancer hijacking [189]. One in six pediatric neuralblastoma carries the prototypical *MYCN* amplification, and in [37] the authors resolved the structure of the highly rearranged, ecDNA amplicon. Combining epigenetic profiling, they also pinpointed the cis-regulatory elements co-amplified with the *MYCN* gene either from its own vincinity or hijacking from distal locus, revealing novel principles of the formation of these driver amplifications. Generally speaking, the young patients have not acquired as many gradually accrued mutations as adult patients, leaving catastrophic rearrangement events an attractive candidate as the driver and iliciting the hypothesis that they might have resulted from very different mutagenesis pathways.

Last but not least, beyond tumor tissues from biopsy, germline, other pathological tissues, and even healthy tissue are also informative to the study of mutational processes of SVs. In the germline genome of autism spectrum disorder and other developmental abnormalities various *de novo* complex SV patterns have been categorized including chromothripsis [190]. Meanwhile, somatic mosiacism in various normal tissues have been characterized and surprinsingly

cancer driver alterations are common even in healthy tissues [191]. In human fetal brains, simple and complex SVs are detectable and can be timed as early as 14 weeks after conception [192]. Such studies are still very challenging owning to the low variant allele frequencies in normal tissue, low sequencing depth, and the relatively high error rate of single cell DNA sequencing. Hence most pioneering studies are focused on SNVs and INDELs, yet detecting traces of junctions and CNAs is apparently the next step.

6.2 Resolving complex haplotypes with long-range sequencing and mapping

While viewing SV event classes as graph motifs on genome graphs has great potential, it is not to be ignored that it is only a compression of actual DNA sequence which are represented as the walks along genome graphs. The same graph can be deconvolved into combinations of haplotypes in a large number of ways, and sometimes this phasing are crucial to differentiate the underlying mutational processes. For example, in homologous recombination deficient (HRD) tumors, with the combination of short-read and linked-read WGS, it is shown that distinct quasi-reciprocal SVs can take either cis or trans phases, and are associated with BRCA1 or BRCA2 deficiency despite looking identical on un-phased genome graphs. Thankfully, technology advancements in long-range sequencing or mapping are a natural complement to overcome the limited power of short-read sequencing in phasing the complex SVs.

Recently, long-molecule profiling has become the frontier in characterizing SVs, for obvious superiority over short-read data which often fall short of

uniquely mapping breakends in repetitive regions [44]. With longer reads, most of the community expect better precision and even better recall in identifying the junctions in a sample [193]. While this is true, we argue that the main strength of long-molecule profiling is to phase and resolve the order of multiple junctions that constitute the derivative sequence by a complex SV event [44, 83, 194, 102]. The recent achievement of a complete human reference genome sequence by assembling multi-platform long-molecules [195] has shown the possibility of obtaining the exact sequences in highly repetitive regions like centromeres. Furthermore, by combining long-read sequencing with single-cell template strand sequencing (Strand-seq), Ebert et al. have reported a haplotype-resolved human germline structural variants and small variants in 35 individuals [196]. In the meantime, new algorithms for assembling haplotype-resolved genomes like Hifiasm [197] and for subsequently compiling them into pangenome graphs like Minigraph [74], show that we are approaching a new era to directly assemble the derivative sequences altered by SVs.

Nonetheless, the current long-read technologies are still progressing towards affordability, longer-ranges, and higher throughput. The cost of any long-molecule profiling is still not on par with the already widely used short-read sequencing, making it harder to scale up and achieve the higher number of samples that SV analyses require. In terms of the length of the profiled molecule, both mainstream long-read sequencing technologies, by Pacific Biosciences and Oxford Nanopore, are still in the median range of tens of kilobases, so that fully resolving long-range complex structures drive the cost even higher. Optical mapping such as Bionano can achieve about 200 kilobases, more suitable for resolving the complex events, yet it is not sequencing and cannot provide base-level resolution. Due to lower number of reads profiled, the coverage data

from long-molecule technologies is not as robust for copy number analysis as short-read counterparts. There have been creative sample preparation and algorithmic enhancements [198, 199] for copy number inference on long-read sequencing platforms, but they have to be adapted to sequence shorter fragments, practically abandoning the particular strength of long-reads. In sum, at least in the near term, using long-read to resolve the structure of complex SVs still requires integration with short-read WGS.

To that end, our genome graph paradigm has built a native way to incorporate the unphased yet more accurate JBGGs from short-read sequencing, with the long-molecule data to reconstruct optimal haplotypes, exemplified by Figure 3B and 6E in [83]. AmpliconReconstructor attempts at the same task but only for amplified regions [194]. RCK is another tools that aims to achieve this [102]. The comparison of performance is left for future work, but our tools are immediately applicable to such benchmarks.

6.3 Uncover SV heterogeneity with multi-sample or single cell WGS

Genetic heterogeneity is a prerequisite for evolution and subclonal SVs have been extensively documented within various tumors [110, 106]. For instance, Watkins et al. [110] characterized intra-tumor copy number heterogeneity from multi-sample, longitudinal biopsies of 394 tumors across 22 tumor types, indicating the continuous selection of CNAs throughout evolution.

Currently, the estimation of cancer cell fraction (CCF) of SVs are individual

junction-focused. One of the latest methods, SVclone [200], uses variant supporting read count over total read count at breakends as the main data source to estimate junction allele frequency, analogous to the methods for SNVs. It makes effort to rescue reads lost due to the split mapping at the breakpoint, as well as allows for correction of copy number if external CN input is given, and achieved satisfiable results. However, like most early SV analytical methods, it ignores the intrisic coupling of junction and copy number, and the connections between junctions. Within JaBbA framework, I can see future improvements incorporating the CCF of each clone as a parameter for junctions, and allow the total CN or allelic CN summed across clones to be real valued (between integers) while restricting the current integer and linear constraints within each clone. The clonal structure can be informed by SNV allele frequency clustering, which provides much more power. Such a model if implemented would output optimal major and sub- clonal genome graphs from single or multiple bulk WGS of tumor samples, greatly empowering the dissection of SV intra-tumoral heterogeneity.

With respect to sampling strategies, multiple spatio-temporal sampling study design is gaining popularity as it taps into the change of SV patterns in related lineages. For example, in [106] we sampled Barrett's esophagus tissues from a prospective patient cohort at two or three time points during the course of disease, from upper and lower esophagus separately. We found abundant rigma events in both the patients that eventually developed esophageal carcinoma and those who did not, at both time points, indicating that rigma is a ubiquitously active mutagenesis process in esophagus regardless of cancer outcome [106]. Yet, the presence of BFB cycles is significantly associated with cancer outcome,

implying that the activation of a mutational process that generates BFB cycles triggers the malignant transformation. In another study,

Promising even greater resolution, single-cell WGS is providing a lens into the intra-tumoral heterogeneity of SVs [201, 113, 114]. Zaccaria and Raphael proposed an innovative method CHISEL to infer allelic-specific copy numbers by aggregating weak signals of SNPs . In general, the analysis of single cell profiling will assign each cell a cluster based on the lower-resolution CN segmentation, and then aggregate the reads in each cluster to form *pseudo-bulk* that represent a homogeneous clone. Our approach applied to the isogenic clones in the post-crisis MRC5 cells are immediately compatible with these data. We expect the combination of the two to yield unprecedented clarity to the dynamic of SV evolution.

6.4 From patterns to mechanisms and back: the trajectory of SV evolution

The other side of the problem to SV pattern identification in real genomes is the fundamental theory of the genesis of SVs. Such theories first drew interets of studies with the goal to infer ancestral genome organization from multi-species synteny data [94]. Greenman et al. set out on a series of papers to rigorously define the features of the outcomes given some known operations on the genomic sequence, for example BFB cycles [202]. These are fundamental conceptual constructions that will be extremely useful in generalizing to real patterns which are often way more complex. Our genome graph paradigm provides a genericb framework to simulate *de novo* sequences of accumulating SVs. We have used its

results in our benchmarking of JaBbA Section 2.2.3. Admittedly, there are many patterns in the result that look biologically improbable, e.g. large portions of loss among many chromosomes, but it proves the feasibility and value doing simulations by incrementally adding more biologically meaningful processes and constraints.

6.5 SVs as biomarkers in clinical sequencing

WGS is becoming recognized as a powerful alternative to the more popular and thus far more cost-effective capture-based platforms [203]. Not only it uncovers all classes of known drivers at a pristine accuracy, the variant patterns derived from vastly more abundant passenger variants can indicate the vulnerability of the tumor. They can serve as a new kind of biomarkers for DNA repair defects or immunogenecity.

Our genome graph paradigm not only provides a sound theoretical foundation for complex SV analysis, but also contribute to the accessibility by inventing easy-to-use, interpretable interfaces for the community. Future plans (already underway) includes a tumor board (a data dashboard regarding the genomic characters of a tumor sample) that compiles putative CNA drivers (amplification of oncogenes or deletion of tumor suppressor genes), potential enhancer hijacking events, complete dictionary of SV events along with SNV/INDEL mutation signatures, and clinically relevant derivative metrics like HRD score (such as HRDetect). If transcriptomic data is available, the associated expression level of any interesting gene will also be demonstrated against a background of

public pan-cancer database to argue for functional impact of observed genomic variants.

To help clinical sequencing reach even wider patient populations, WGS of plasma cell-free DNA (cfDNA) are drawing more attention. A number of recent studies point to its use in minimal residual disease (MRD) monitoring [204, 205], profiling of metastatic diseases [206], or even early detection of tumor cells in healthy population [207, 208]. In all of these applications, it is almost impossible to identify the exact breakends of a somatic junction as the read depth for tumor derived DNA fragments are too shallow. However, a considerably detailed copy number profile are repeatedly shown feasible to inferred. A possible way to transfer the learnings in tumor tissue WGS is to use the genome graphs and SV patterns identified as training labels, and ask a model to predict the presence of a particular pattern from downsampled, lossy data of cfDNA. If realized, this imputation of SV pattern based on sparse CN will add crucial information to the tumor genome profiling from merely a tube of peripheral blood drawn from the patient.

All in all, I strongly believe that this thesis proved the genome graph paradigm a rigorous, pragmatic framework for the analysis of SVs in cancer genomes. I wish it will soon be applied by the community to answer various questions about SV mutagenesis and cancer etiology, and hopefully make long-lasting impacts on the treatment of cancer patients.

APPENDIX A: CHAPTER 2 SUPPLEMENTARY MATERIALS AND METHODS

Genome simulations

Simulated sequencing of rearranged tumor samples for benchmarking were first constructed as haplotype-specific genomic sequences in `.fasta` format, then read sampling, alignment, and finally mixed with normal reads to a series of purity levels. In the first step we generated a set of *de novo*, forward simulated, rearranged cancer genomic sequences from an initial set of input junctions (`SimBLE`, `https://github.com/mskilab/sim.ble`). SimBLE iterates through simulated cell cycles to gradually incorporate the input junctions into the derived genome from previous steps until exhausting the input junction set, while keeping track of the actual rearranged haplotypes. As a result, it generates a coherent sequence of the rearranged genome guided by the haplotypes encoded in the reference genomic ranges. We then simulated sequencing reads from this FASTA file with ART read simulator [209] to an average depth of 40X and aligned them to the reference genome hg19 to obtain the simulated BAMs. Trivially, the reference genome itself is also subjected to the same *in silico* sequencing to provide as normal controls. We did 40 distinct simulations with different random subsets of somatic junctions (from 5 to 333 total) identified in the HCC1143 breast cancer cell line.

In addition to these *de novo* simulated BAMs, we also obtained WGS for HCC1954 breast cancer cell line and HCC1954BL the corresponding normal fibroblast cell line. To simulate stromal admixture, we combined tumor and normal simulated (or HCC1954) BAM files, then downsampled, and mixed tumor

and normal reads across ten tumor DNA proportions from 0.1 to 1.0. We created four technical replicates for each of these ten purity levels, which yielded 40 pairs of tumor and normal BAM files for the 40 distinct simulated genomes and HCC1954, respectively.

APPENDIX B: CHAPTER 3 SUPPLEMENTARY MATERIALS AND METHODS

Structural variant event classification

To identify simple and complex structural variant events in the genome graph output of JaBbA, we implemented a set of classifiers. These are implemented in the `events` function in the `gGnome` package (freely available from: `https://github.com/mskilab/gGnome`), which calls the classifier functions below. Procedures followed to discover each event type are described below.

Rigma and simple deletions

Rigma candidates were nominated as clusters of at least 2 overlapping DEL-like junctions with JCN less than ploidy and sizes between 10 kbp and 10 Mbp. `fishHook` was used as a Poisson model on a per sample basis to statistically nominate regions of the genome enriched with DEL-like junctions using sliding windows of 1 Mbp (500 kbp stride) while correcting for the occurrence of other junction types. From the model, regions with a significant enrichment of $FDR < 0.5$ were chosen. Any significant regions that were adjacent were collapsed into contiguous regions. Candidate deletion clusters that fall within these statistically enriched collapsed regions were then marked as rigma events. The remaining isolated low-JCN deletions were called simple deletions. This caller is implemented as the `del` function of `gGnome`.

Pyrgo and simple duplications

The procedure to identify pyrgo and simple duplications was analogous to the rigma identification above, but with the use of DUP-like junctions instead of DEL-like junctions. The same filters were applied as above. For candidate duplication clusters that pass these filters, fishHook was again used to statistically nominate genomic regions enriched for duplications while correcting for the occurrence of non-duplication junctions. Candidate duplication clusters that fall within the statistically enriched windows were then marked as a pyrgo. Collapsing was performed as for rigma. Remaining isolated low-JCN duplications were called simple duplications. This caller is implemented as the `dup` function of `gGnome`.

Double minute, BFB cycles, and Tyfonas

To nominate amplicons (amplified clusters), weakly connected components were identified among vertices with copy number above 2×ploidy in JaBbA graphs. The resulting 12,588 subgraphs were further subsetted to only those whose maximum JCN > 7, leaving 1,703 such high level amplicons with at least one high copy junction. Four curated features were then used to annotate each of these amplicons, 1) maximum segmental/interval copy number 2) the sum of all fold back inversion JCN divided by maximal interval copy number, 3) the ratio between maximum JCN and maximum interval copy number, and 4) the number of junctions with elevated JCN (thresholded on JCN > 3). To separate these amplicons, hierarchical clustering with Euclidean distance and complete linkage was performed on the basis of these features.

To assess the most optimal number of amplicon clusters, k clusters from 2 to 15 were tested via a bootstrapping procedure. Briefly, for each k, the amplicons were clustered across 100 bootstraps in which 75% of the data was sampled and assessed using function "clusterboot", from the `clusterboot` R package. For each bootstrap and setting of k, k clusters were computed and a Jaccard similarity index was computed between every observed cluster and its most similar cluster in the sample. The stability of each observed cluster associated with a given k was computed as the average Jaccard similarity index across all bootstraps for that observed cluster. The cluster stability of the solution with parameter k was then computed as the average stability of all observed clusters associated with k.

Using the elbow method, $k = 4$ was identified as the parameter setting with the maximal cluster stability. Furthermore, within each original cluster for a given k, a mean Jaccard score was computed to assess the level of evidence for a cluster. Clusters with mean score > 0.75 were considered stable, meaning there was sufficient evidence in the data to support the cluster as a true pattern, while those with mean score < 0.6 were likely artefactual.

Clustering was performed by building an unsupervised classifier that combines hierarchical clustering with a decision tree. Specifically, a decision tree was trained using recursive partitioning (function `rpart` in the `eponymous` R package) from the hierarchical clustering result based on the above features. The decision tree arrived at a fold back JCN ≥ 0.5 to distinguish tyfonas/BFB cycles amplicons from the double minute/Other amplicon categories. Out of the amplicons with fold-back JCN ≥ 0.5, amplicons with total number of junctions with high JCN (# ≥ 26) were called tyfonas and the rest were called BFB cycles. Within

amplicons containing fold-back JCN < 0.5, those amplicons with ≥ 31 high copy junctions distinguished otherwise unspecified amplicons ("Other") from regular double minutes, respectively. This algorithm was then used to nominate the tyfonas, double minutes, BFB cycles, and the unspecified amplicons ("Other") among high level amplicons. These decision tree-based calls were then used for the subsequent analyses on these amplification events (see below). Although 4 was the optimal number of clusters, only three of the clusters (tyfonas-, double minute- and BFB cycles-labeled amplicon clusters) showed a mean Jaccard score of > 0.75, with a fourth cluster scoring 0.67. Due to the inconclusiveness of this fourth cluster within the amplicon data, this pattern was excluded from further analyses. This caller is implemented as the `amp` function of `gGnome`.

Chromothripsis

Subgraphs that best fit the features of chromothripsis patterning described in [210] were identified. Candidate clusters (i.e. weakly connected components) of genomic intervals overlapping the footprints of at least three ALT junctions were nominated. These were further filtered to those clusters whose segmental copy numbers occupied a narrow range of states. Specifically, each interval within every footprint of a candidate cluster was subject to three criteria: 1) occupancy of at most 3 different copy states (for a diploid genome, adjusted up proportionally to the ploidy of the genome), 2) composition of at least 8 segments, and 3) having an interval copy number at the width-weighted 99^{th} percentile that did not exceed 4 (for a diploid genome, or adjusted to the ploidy of the genome). Clusters that survived this three-step filter were required to contain junctions with a near-uniform mixture of basic configurations (χ^2 test, p-

value > 0.001). Out of the final remaining clusters, several further criteria were required to yield the final chromothripsis calls as follows: each cluster must have 1) at least 7 internal junctions, 2) no fewer than two sub-100kbp footprints and no more than four ≥ 100 kbp segments within the cluster, and 3) on average at least 3 junctions with inter-breakend spans that overlap one another. This caller is implemented as the `chromothripsis` function of `gGnome`.

Chromoplexy

Chromoplexy as first described in [35] can be identified by a series of reciprocal (or nearly-reciprocal, with small deletion bridges between adjacent breakends) long-range junctions (span more than 10 Mbp on the reference). Accordingly, chromoplexy was identified from a pool of low-JCN (≤ 3) edge clusters in which junction breakends are no further than 10 kbp away from the next junction breakend. Edge clusters that contained at least 3 long-range junctions, and whose footprints occupied at least 3 discontiguous genomic territories separated by >10 Mbp on the reference were called as chomoplexy events. This caller is implemented as the `chromoplexy` function of `gGnome`.

Templated insertion chains (TICs)

We nominated templated insertion chain (TIC) events that result from short insertions involving more than 1 junction, usually with a gain of copy of all loci within the event, or linking disparate loci through shorter segmental "hops" within the genome [41]. To capture this pattern, the following procedure was used: Breakends identified across the whole genome were ordered along ref-

erence coordinates, and pairs of breakends that fall within an interval of ≤ 500 kbp and ≥ 50 kbp of each other were kept as candidates. These candidate breakend pairs comprise a "+"breakend followed by a "−" breakend in the direction of increasing reference coordinates, with each breakend originating from a different junction. A new graph was constructed with vertices comprising each junction and edges comprising adjacencies linked by candidate breakend pairs. All walks through the graph that traverse through at least 2 vertices/junctions were obtained. ALT junctions within each walk traversed on this graph were then labeled as a unique TIC event. This caller is implemented as the `tic` function of `gGnome`.

Other simple events

Inversions, inverted duplications, and translocations were also called as part of the compendium of SV event types in this study. Inversions and inverted duplications were both defined as pairs of overlapping, oppositely oriented, INV-like junctions of the same JCN as well as equal left and right vertex CN, with no third junction or a loose end interfering. We defined translocations as single or reciprocal pairs of TRA-like junctions connecting two different reference chromosomes that are not connected by any other junctions. This caller is implemented as the `simple` function of `gGnome`.

Cancer genome analysis

Harmonized pipelines for junction detection, high-density read depth estimation, purity / ploidy inference, somatic SNV / indel calling, and loss of function

mutation calling were applied to the full dataset of 2,813 WGS cases as described below. Several standard analyses for mutational interpretation, signature inference, and APOBEC analyses were performed.

Junction detection

High- and low-confidence somatic junction calls for tumor / normal pairs were obtained using `SvABA` [46], based on the optimal settings for input into JaBbA (see above). For samples (e.g. CCLE cell lines) that lacked a paired normal sample, we used HCC1143BL WGS as the normal / constitutional reference sample.

To further eliminate constitutional junctions, we used `SvABA` to obtain normal / constitutional junction calls for 1,017 TCGA tumor/normal pairs and construct a panel of normals (PON). All somatic junction candidates within 1 kbp of a PON junction (determined by overlap at both breakends) were flagged as constitutional and removed. The remaining `SvABA` calls were treated as "high-confidence" calls if they passed `SvABA` internal filters. The remaining calls were treated as "low-confidence" (see "JaBbA model fitting" section above).

Read depth

Coverage profiles were derived using fragCounter, which counts the midpoints of proper pairs in 200 bp bins tiling hg19 using `samtools` (v1.9) and corrects the binned read counts by LOESS (function `loess` within R base package) (see "read depth preprocessing" section for mathematical details). We then segmented tumor / normal ratios using Circular Binary Segmentation [100] (segment function in DNAcopy R package). For samples lacking a matched normal

(i.e. CCLE cell lines), a composite of the 1,017 TCGA normal coverage profiles was used, comprising the average of the 200 bp bins across all autosomal chromosome coordinates.

Quality control

From the 2,813 WGS sequenced samples, we excluded 34 samples for which JaBbA optimization did not converge and one sample which exceeded RAM requirements during SV classification. Among 69 graphs which did not converge on first pass (< 16G RAM, < 24 hours compute time), 35 successfully completed with increased memory and run time (< 500G RAM, > 24 hours of compute time). One of these 35 failed SV classification due to the number of incorporated junctions exceeding RAM requirements. Upon inspection, graphs that failed to converge at the JaBbA were associated with noisy read depth data and many segments (> 130,000 internal vertices).

Purity and ploidy estimation

For all cases with both a tumor and normal WGS profile, read counts supporting germline heterozygous SNP alleles were obtained by intersecting SNV sites present in both tumor and normal with SNPs from HapMap 3.3. Purity and ploidy estimates were obtained for all samples through `Sequenza` [211], `TITAN` [212], `Ppurple` (`https://github.com/mskilab/Ppurple`), or a custom least squares grid search `ppgrid` (available through JaBbA). Consensus purity and ploidy was determined by curation from the panel of the three calls across all tumor normal pairs. `ppgrid`, which does not require germline SNP

sites, was used for CCLE cell lines to obtain purity and ploidy estimates since germline SNP sites were unavailable.

Somatic SNVs and indels

To obtain somatic SNV/indel calls `Strelka2` [213] was run under paired (i.e. tumor / normal) mode with default parameters using hg19-based references. Somatic SNVs and indels were obtained only for those cases where tumor and normal BAMs were available (2481 out of 2813 cases). In addition to the recommended filters, a universal mask (https://github.com/lh3/sgdp-fermi/releases/download/v1/sgdp-263-hs37d5.tgz) was used to remove common artifacts in low-mappability regions, described in [214]. After these initial filters, only sites determined to pass `Strelka2`'s quality filter (i.e. sites where the "FILTER" field was marked as "PASS") were considered, yielding a high quality set of somatic SNV calls. Variant annotations were obtained for SNV/indel using SnpEff with the GRCh37.75 database.

Germline SNVs and indels

For constitutional SNV/indel calls `Strelka2` [213] was run as above, except in normal-only mode. Germline SNVs and indels were obtained only for non-cell line samples within the study. The universal mask was also used for germline SNV/indel calls. Additional filters restricted the germline variants used to those which met the following criteria: i) sites that do not overlap with common variants i.e. those variants that matched coordinates and ALT alleles with sites from the normal ExAC population

that have a minor allele frequency of >1% (`ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/subsets/`), ii) high quality sites as determined by `Strelka2`'s quality thresholds (i.e. sites in which the "FILTER" field was marked as "PASS"), and iii) sites that overlapped and matched ALT alleles with known pathogenic variants from ClinVar annotations (`ftp.ncbi.nlm.nih.gov:/pub/clinvar/vcf_GRCh37`). Variant annotations were obtained for this final, high quality set of germline SNVs/indels using SnpEff as above. For constitutional heterozygous SNP calls (see "Postprocessing" section above), we used HapMap 3.3 defined on hg19 (https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html). We assessed tumor and normal SNP counts using Rsamtools (R / Bioconductor) and nominated constitutional heterozygote sites as any site with > 0.2 and < 0.8 ALT variant allele fraction in the normal sample.

Mutation interpretation

Alteration status of genes was obtained across the cohort. Homozygous deletions, heterozygous deletions, and amplifications were detected from absolute copy number calls available from the JaBbA graphs. Filtered SNV/indel annotations falling within protein coding regions, were considered only if they constituted somatic missense or truncating events, germline pathogenic variants, or germline truncating variants. To obtain loss-of-heterozygosity calls allele-specific copy number was called in cases with both tumor and normal samples. With genome-wide germline heterozygosity coverage for these samples, allele specific copy number was measured using JaBbA using the `jabba.alleles()` function within the `JaBbA R` package. For CCLE, our calls only included the publicly available mutational drivers.

Driver amplification analysis

For the scope of amplification drivers, we selected Cancer Gene Census genes that intersect with the pan-cancer GISTIC amplification peaks in [124]. Any overlap between the genomic footprint of a double minute, BFB cycles, or tyfonas event with an amplified driver gene (thresholded as having a segmental CN > twice ploidy) was counted to calculate the frequency of each event upon one of these genes.

Protein-coding gene fusions across complex SV types

We obtained expressed protein-coding gene fusion calls from a published TCGA RNA-seq analysis [215]. Coordinate overlaps between expressed (> 10 supporting RNA-seq reads) protein-coding gene fusion calls and SV event footprints from JaBbA graphs were tallied across 870 intersecting tumors. The number and genomic density of unique, expressed fusion products attributed to each event type were plotted in Figure 3.2. Enrichment between tyfonas and other SV event categories (double minutes, BFB cycles, chromothripsis, other) was assessed using Wilcoxon rank-sum test.

APPENDIX C: CHAPTER 4 SUPPLEMENTARY MATERIALS AND METHODS

WGS basic data processing

Reads were aligned to GRCh37/hg19 using the Burroughs-Wheeler aligner (bwa mem v0.7.8, http://bio-bwa.sourceforge.net/). Best practices for post-alignment data processing were followed through use of Picard (https://broadinstitute.github.io/picard/) tools to mark duplicates, the GATK (v.2.7.4) (https://software.broadinstitute.org/gatk/) IndelRealigner module, and GATK base quality recalibration.

Variant rearrangement junctions were identified using SvABA [46] and GRIDSS [47] with standard settings. For MRC5 samples, the somatic variant setting of each tool was used, with the ancestral MRC5 line as the matched normal. SvABA was applied using a panel of normals (PON) that was constructed by running SvABA to obtain constitutional junction calls for 1,017 TCGA tumor/normal pairs (TCGA dbGaP: phs000178.v11.p8). For GRIDSS, a PON was obtained from the Hartwig Medical Foundation (https://nextcloud.hartwigmedicalfoundation.nl). 1 kbp binned GC and mappability corrected read depth was computed using fragCounter (https://github.com/mskilab/fragcounter). Systematic read depth bias was subsequently removed using dryclean (https://github.com/mskilab/dryclean).

Low-pass WGS clustering

Genome-wide binned read depth was aggregated across 118 low pass WGS clones across 10 kbp bins by taking the median of 1 kbp binned normalized read depth from dryclean (see above). To minimize read depth noise in unmappable regions, recurrent (>10% of the cohort) low-quality coverage regions (defined in [83]) are combined with regions bearing consistently high variance in our high-pass sequencing dataset (standard deviation >0.3 for bin value over the mean in 100 kbp windows). Hierarchical clustering was then applied on the genome-wide Euclidean distance of bins, with "method = ward.D2" option. Six clusters were identified following dendrogram inspection.

Junction balance analysis

Preliminary junction balanced genome graphs were generated for MRC5 and SV40T cell lines from binned read depth and junction calls (see above) using JaBbA (https://github.com/mskilab/JaBbA)[83]. Briefly, 1 kbp binned read depth output from dryclean was collapsed to 5 kbp and JaBbA was run with slack penalty 500. gGnome (https://github.com/mskilab/gGnome) was used to identify complex structural variant patterns. Genome graphs and corresponding genomic data (e.g. binned coverage, allelic bin counts) were visualized using gTrack (https://github.com/mskilab/gTrack).

Loose end classification

Each loose end in each MRC5 genome graph was analyzed to identify a clone specific (i.e. absent in the ancestral MRC5 line) origin for the mates of high mapping quality (MAPQ=60) reads mapping to the location and strand of the loose end. These mates were assessed for neo-telomeric sequences by counting instances of 11 permutations of a 12-bp telomere repeat motif (TTAGGGTTAGGG) (using the R / Bioconductor Biostrings package) in the mates. The mates were also assembled into contigs using fermi [70] aligned using bwa mem [137] via the RSeqLib R package [150] to hg38 (which contains a more highly resolved centromere build) to characterize novel repeat (e.g. centromere) fusions (GATK Human reference genome, hg38, data bundle including Homo_sapiens_assembly38.fasta, gs://gcp-public-data–broad-reference). The loose end loci were also assessed through overlap with the hg19 repeatMasker database (human_g1k_v37_decoy.repeatmasker) for the presence of reference annotated repeats that might explain the absence of a mappable junction explaining the copy number change.

SNV phylogeny

To compute an SNV phylogeny across MRC clones, we first identified SNV that were acquired in MRC5 clones relative to the ancestral MRC5 line using Strelka2 [213] (https://github.com/Illumina/strelka) under paired (i.e. tumor / normal) mode with the clone as the "tumor" and the MRC ancestral line as the "normal" sample and default parameters and GATK hg19 resource bundle (Genome Analysis Toolkit GATK Resource Bundle for hg19; gs://gatk-legacy-

bundles). Acquired SNVs were first filtered according to the Strelka2 PASS filter as well as additional filters (MQ = 60, SomaticEVS>12, total ALT count > 4) yielding 27,220 total unique variants across the 13 MRC5 clones. Reference and variant allelic read counts were assessed at each SNV site (via the R / Bioconductor Rsamtools package, version 3.6.1, http://www.r-project.org/) across all 13 clones. We then further required a >0.5 posterior probability of a variant being present in a sample, by assuming Binomial likelihood of variant read count and using the aggregated allele frequency in all samples as the prior, resulting in the final 14,970 unique mutations. The binary matrix of clones by SNV loci. was then used to derive a neighbor-joining phylogenetic tree using the R / Bioconductor package ape. Following tree construction, we associated each SNV with its most likely phylogenetic tree branch by comparing the binary incidence vector associated with each SNV with the binary incidence vector associated with each tree branch, and finding the closest branch using Jaccard distance, only linking SNV to branches when the SNV was within <0.1 Jaccard distance of the closest branch, thus producing the groupings of SNVs in Figure 4.12.

Parental SNP allelic phasing and imbalance

Germline heterozygous sites in the parental MRC5 line were identified by computing allelic counts at HapMap sites (GATK human reference genome, hg19 data bundle, hapmap_3.3.b37.vcf) and identifying loci with variant allele fraction >0.3 and <0.7. Y11, a clone with loss of a single allele at 12p, was chosen to phase parental SNPs on 12p. At each locus, the allele (reference or alternate) with a 0 read count was assigned to the "L" (lost) haplotype and the other allele was assigned to the "R" (retained) haplotype. (All heterozygous SNP loci in the

region contained exactly one allele with a 0 read count). L and R allelic counts were then computed at these sites across all 13 high pass WGS and 131 low pass WGS samples. These counts were divided by the genome wide mean of heterozygous SNP allele counts (in these 100% pure and nearly diploid samples) to derive the absolute allelic copy number [53].

SNV clustering

Inter-SNV distances were computed for all pairs of reference adjacent acquired SNVs associated with each MRC5 clone and visualized as rainfall plots. Runs of two or more SNVs with inter SNV distances < 2 kbp were nominated as clusters. Two distinct SNV clusters were identified on chromosome 12p across the 13 clones.

APPENDIX D: CHAPTER 5 SUPPLEMENTARY MATERIALS AND METHODS

Sample selection and whole-genome sequencing

Eighty-five samples were selected for WGS, among 118 previously whole-exome sequenced TCGA LUAD samples that were negative for 1) activating mutations in *KRAS*, *EGFR*, *BRAF*, *ERBB2*, *MET*, *RIT1*, *NRAS*, *RAF1*, *HRAS*, *ARAF*, *MAP2K1* and *SOS1*; 2) loss-of-function mutations in *NF1* and *RASA1*; 3) fusions in *ALK*, *ROS1*, *RET*, *MET* and *NTRK2*; and 4) amplification in *EGFR*, *ERBB2*, *KRAS*, *MET*, *FGFR1* and *MAPK1*. The same criteria were applied to re-identify RPA(-) samples in the WGS analysis, except that over-expression (defined as a z-score greater than 1.96 for gene expression among the full TCGA LUAD samples) was additionally required to qualify an amplification as an oncogenic driving event. DNA was sequenced using the Illumina HiSeq platform. Paired-end sequencing reads were aligned to hg19 using BWA (v0.6.2) [216] aln and processed through NovoSort (v1.03.01) to mark PCR duplicates (http://www.novocraft.com/products/novosort/), then through GATK (v3.4) [217] for indel realignment (jointly for the normal and tumor samples). Base quality scores were recalibrated with GATK.

Identification of somatic mutations and SCNAs

Somatic SNVs were called by MuTect (v1.1.7) [218], Strelka (v1.0.14) [219] and LoFreq (v2.1.3a) [220]. Somatic indels were called by Strelka, Pindel (v0.2.5) [221] and Scalpel (v0.5.3) [222]. Variants called by only one of the three callers

were filtered out. Final VCFs were formatted to pass the EBI validator (v 0.4.3) (https://vcftools.github.io/perl_module.html). Variants were annotated for their effect (non-synonymous coding, nonsense, etc.) using snpEff [223] based on human genome annotations from ENSEMBL. We further annotated the variants using snpEff, snpSift [223] and GATK VariantAnnotator module with information from COSMIC [224], 1000 Genomes Project [225] , ExAC [226], CIViC [227], and UniProt [228]. Non-coding mutations were further annotated by Funseq2 [168]. Mutational signatures were analyzed using SignatureAnalyzer [229].

Data on genetic ancestry, genome double, aneuploidy, leukocyte fraction and other clinical features were downloaded from the Genomic Data Commons (https://portal.gdc.cancer.gov/). GISTIC 2.0 [165] was used to identify significant SCNAs using copy number segments generated by Titan [51]. High-level amplification was defined by log2-transformed copy number ratios > 1. To calculate the allelic fraction of *KRAS* mutations in WGS and WES, we applied a custom script counting reads supporting the altered alleles and the reference alleles, respectively [230]. Reads with based quality and mapping quality lower than 30 were removed.

Identification of Structural variants and genome graph reconstructions

Somatic aberrant junctions (i.e. pairs of strand-specific disconnected loci that form neo-adjacencies in the cancer genome) were identified by SvABA v1.1.3 [97], using the default setting for tumor-normal pairs. We then used JaBbA,

a junction balance analysis [83] to reconstruct a genome graph for each sample through the application of a maximum likelihood model to high-resolution binned (200bp) normalized read depth and unfiltered SvABA junctions (after exclusion of small (<1kbp) deletion-like junctions). The read depth input to JaBbA was calculated as the ratio between tumor and normal sample's WGS read counts in all 200bp genomic bins, corrected for guanine/cytosine content and 100mer mappability. Subsequently, we used Circular Binary Segmentation [231] with αparameter at 1×10^{-5} to derive a primary segmentation for each sample. The segmentation was later combined with SvABA-identified aberrant junctions to build the genome graph. The affine mapping between read depth signal and integer copy number is dictated by the hyperparameters ploidy and purity. We used the published purity and ploidy values from the GDC (https://portal.gdc.cancer.gov/). Any sample missing from that resource was supplemented by Sequenza (Favero et al., 2015), based on the allelic read counts at germline heterozygous sites using Samtools [216]. A total of 13 types of simple and complex SV events were annotated and visualized using gGnome (https://github.com/mskilab/gGnome.git), with the same default parameters as described in Hadi et al. 2020. SV burden per sample was defined by the number of junctions of simple SVs in that sample. Genome graphs drawn in copy number over genomic coordinates plots are made with gTrack (https://www.github.com/mskilab/gTrack) and gGnome (https://www.github.com/mskilab/gGnome).

Oncoprints, mutation barplots, and expression quantiles

Genomic alterations affecting genes in the cohort were plotted with ComplexHeatmap [232] with the aforementioned definitions for SNVs and CNAs. Tumor mutation burden was calculated by dividing the total number of SNVs in the eligible mutation calling region proposed in [137] by the total width of these regions (2429.397 Mbp). Expression quantiles and density plots are made with the gene's RSEM (RNA-seq by Expectation Maximization) values of the full set of 507 LUAD RNA-seq in TCGA. Mutation barplots ("loliplot") are made with trackViewer package (http://bioconductor.org/packages/release/bioc/html/trackViewer.html).

Differential alteration frequencies between RPA(-)$_G$ and RPA(+)$_G$

The frequency of genomic alterations in various genes were compared between RPA(-)$_G$ and RPA(+)$_G$ using Fisher's exact tests with false discovery rate (FDR) threshold below 0.1. To maximize statistical power, we only considered the variant types that can be detected both through WES and WGS, and compared the frequency of alteration in the RPA(-)$_G$ group (N=57) to the rest of all the TCGA LUAD samples with WES-based variant calls from the PanCanAtlas (N=411, http://api.gdc.cancer.gov/data/1c8cfe5f-e52d-41ba-94da-f15ea1337efc). Same method was used for clinical or molecular features, including smoking history, age of diagnosis, leukocyte fraction, genome doubling, degree of aneuploidy, genetic ancestry, primary disease stage (data available through GDC) using Fisher's exact test or Wilcoxon's Rank test. TMB comparison was performed based on WES-defined TMB values using linear regression controlling for tumor purity.

Recurrence analysis of genomic alterations

A gamma-Poisson regression framework (fishHook) that takes into account different confounders that affect the mutation count in a population was used [164]. fishHook allows for defining the genomic region of interest, called the hypotheses set, to be used for recurrence analyses. For coding mutation analysis, models were fitted with gene bodies as the hypotheses set. Non-synonymous, missense and truncating mutations were tested separately. We also corrected for multiple hypotheses on 47 genes identified by prior genomic studies on LUAD [155, 156, 120]. For non-coding mutation analysis, lung-specific ATAC-

seq peaks (https://gdc.cancer.gov/about-data/publications/ATACseq-AWG) defining open chromatin regions [167] were used as the hypotheses set. Lift over was used to map hg38 coordinates to hg19. ATAC peaks occurring at least 2 of 44 LUAD samples were used, and peaks within 100bp distance were merged. The model was fitted with the following covariates of the neutral mutation density: 1) Fraction of heterochromatic regions in each query interval. Heterochromatin annotation obtained from chromHMM on A549 cell line from Epigenomics Roadmap [233]; 2) Gene expression data for LUAD for A549 cell line; 3) GC content in reference genome; 4) Replication timing from normal human epidermal keratinocytes; 5) DNA accessibility annotation from DNase-seq for A549 cell line. Besides genome-wide hypotheses, we also tested within the subset of ATAC peaks that overlaps recurrently amplified regions [155] and with putative target gene median RSEM>10 among LUAD samples.

Associations between non-coding mutation and target gene expression are evaluated through fitting an ordinary linear model to log RSEM values with the non-coding mutation presence and amplification status of *ILF2* gene.

Quantification and Statistical Analysis

All statistical tests are carried out using R (v3.6.1) and Bioconductor (v3.10) and listed within the figure legends and Results. Fisher's exact tests are executed with "fisher.test" function, Wilcoxon's Rank test with "wilcox.test", ordinary linear models with "lm".

BIBLIOGRAPHY

[1] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, March 2011.

[2] ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82–93, February 2020.

[3] Theodor Boveri. Concerning the origin of malignant tumours by theodor boveri. translated and annotated by henry harris. *J. Cell Sci.*, 121 Suppl 1:1–84, January 2008.

[4] Matthew Meyerson, Stacey Gabriel, and Gad Getz. Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.*, 11(10):685–696, October 2010.

[5] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–724, April 2009.

[6] Ke Shi, Michael A Carpenter, Surajit Banerjee, Nadine M Shaban, Kayo Kurahashi, Daniel J Salamango, Jennifer L McCann, Gabriel J Starrett, Justin V Duffy, Özlem Demir, Rommie E Amaro, Daniel A Harki, Reuben S Harris, and Hideki Aihara. Structural basis for targeted DNA cytosine deamination and mutagenesis by APOBEC3A and APOBEC3B. *Nat. Struct. Mol. Biol.*, 24(2):131–139, February 2017.

[7] Ludmil B. Alexandrov, Serena Nik-Zainal, David C. Wedge, Samuel A. J. R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Børresen-Dale, Sandrine Boyault, Birgit Burkhardt, Adam P. Butler, Carlos Caldas, Helen R. Davies, Christine Desmedt, Roland Eils, Jórunn Erla Eyfjörd, John A. Foekens, Mel Greaves, Fumie Hosoda, Barbara Hutter, Tomislav Ilicic, Sandrine Imbeaud, Marcin Imielinski, Natalie Jäger, David T. W. Jones, David Jones, Stian Knappskog, Marcel Kool, Sunil R. Lakhani, Carlos López-Otín, Sancha Martin, Nikhil C. Munshi, Hiromi Nakamura, Paul A. Northcott, Marina Pajic, Elli Papaemmanuil, Angelo Paradiso, John V. Pearson, Xose S. Puente, Keiran Raine, Manasa Ramakrishna, Andrea L. Richardson, Julia Richter, Philip Rosenstiel, Matthias Schlesner, Ton N. Schumacher, Paul N. Span, Jon W. Teague, Yasushi Totoki, Andrew N. J. Tutt, Rafael Valdés-Mas, Marit M. van Buuren, Laura van't Veer, Anne Vincent-Salomon, Nicola Waddell, Lucy R. Yates, Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium,

ICGC PedBrain, Jessica Zucman-Rossi, P. Andrew Futreal, Ultan McDermott, Peter Lichter, Matthew Meyerson, Sean M. Grimmond, Reiner Siebert, Elías Campo, Tatsuhiro Shibata, Stefan M. Pfister, Peter J. Campbell, and Michael R. Stratton. Signatures of mutational processes in human cancer. *Nature.*, 500:415, 2013.

[8] Rémi Buisson, Adam Langenbucher, Danae Bowen, Eugene E Kwan, Cyril H Benes, Lee Zou, and Michael S Lawrence. Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science*, 364(6447), June 2019.

[9] Michael S Lawrence, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Chip Stewart, Craig H Mermel, Steven A Roberts, Adam Kiezun, Peter S Hammerman, Aaron McKenna, Yotam Drier, Lihua Zou, Alex H Ramos, Trevor J Pugh, Nicolas Stransky, Elena Helman, Jaegil Kim, Carrie Sougnez, Lauren Ambrogio, Elizabeth Nickerson, Erica Shefler, Maria L Cortés, Daniel Auclair, Gordon Saksena, Douglas Voet, Michael Noble, Daniel DiCara, Pei Lin, Lee Lichtenstein, David I Heiman, Timothy Fennell, Marcin Imielinski, Bryan Hernandez, Eran Hodis, Sylvan Baca, Austin M Dulak, Jens Lohr, Dan-Avi Landau, Catherine J Wu, Jorge Melendez-Zajgla, Alfredo Hidalgo-Miranda, Amnon Koren, Steven A McCarroll, Jaume Mora, Ryan S Lee, Brian Crompton, Robert Onofrio, Melissa Parkin, Wendy Winckler, Kristin Ardlie, Stacey B Gabriel, Charles W M Roberts, Jaclyn A Biegel, Kimberly Stegmaier, Adam J Bass, Levi A Garraway, Matthew Meyerson, Todd R Golub, Dmitry A Gordenin, Shamil Sunyaev, Eric S Lander, and Gad Getz. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, July 2013.

[10] Kadir C Akdemir, Victoria T Le, Justin M Kim, Sarah Killcoyne, Devin A King, Ya-Ping Lin, Yanyan Tian, Akira Inoue, Samirkumar B Amin, Frederick S Robinson, Manjunath Nimmakayalu, Rafael E Herrera, Erica J Lynn, Kin Chan, Sahil Seth, Leszek J Klimczak, Moritz Gerstung, Dmitry A Gordenin, John O'Brien, Lei Li, Yonathan Lissanu Deribe, Roel G Verhaak, Peter J Campbell, Rebecca Fitzgerald, Ashby J Morrison, Jesse R Dixon, and P Andrew Futreal. Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. *Nat. Genet.*, October 2020.

[11] Pablo E García-Nieto, Erin K Schwartz, Devin A King, Jonas Paulsen, Philippe Collas, Rafael E Herrera, and Ashby J Morrison. Carcinogen susceptibility is regulated by genome architecture and predicts cancer mutagenesis. *EMBO J.*, 36(19):2829–2843, October 2017.

[12] Marcin Imielinski, Guangwu Guo, and Matthew Meyerson. Insertions and Deletions Target Lineage-Defining Genes in Human Cancers. *Cell*, 168:460–472, 2017.

[13] Paz Polak, Rosa Karlić, Amnon Koren, Robert Thurman, Richard Sandstrom, Michael S. Lawrence, Alex Reynolds, Eric Rynes, Kristian Vlahoviček, John A. Stamatoyannopoulos, and Shamil R. Sunyaev. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature*, 518:360, 2015.

[14] Xueqing Zou, Gene Ching Chiek Koh, Arjun Scott Nanda, Andrea Degasperi, Katie Urgo, Theodoros I Roumeliotis, Chukwuma A Agu, Cherif Badja, Sophie Momen, Jamie Young, Tauanne Dias Amarante, Lucy Side, Glen Brice, Vanesa Perez-Alonso, Daniel Rueda, Celine Gomez, Wendy Bushell, Rebecca Harris, Jyoti S Choudhary, Josef Jiricny, William C Skarnes, and Serena Nik-Zainal. A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. *Nature Cancer*, pages 1–15, April 2021.

[15] Nadezda V Volkova, Bettina Meier, Víctor González-Huici, Simone Bertolini, Santiago Gonzalez, Harald Vöhringer, Federico Abascal, Iñigo Martincorena, Peter J Campbell, Anton Gartner, and Moritz Gerstung. Mutational signatures are jointly shaped by DNA damage and repair. *Nat. Commun.*, 11(1):2169, May 2020.

[16] Harald Vöhringer, Arne Van Hoeck, Edwin Cuppen, and Moritz Gerstung. Learning mutational signatures and their multidimensional genomic properties with TensorSignatures. *Nat. Commun.*, 12(1):3628, June 2021.

[17] Serena Nik-Zainal, Helen Davies, Johan Staaf, Manasa Ramakrishna, Dominik Glodzik, Xueqing Zou, Inigo Martincorena, Ludmil B Alexandrov, Sancha Martin, David C Wedge, Peter Van Loo, Young Seok Ju, Marcel Smid, Arie B Brinkman, Sandro Morganella, Miriam R Aure, Ole Christian Lingjærde, Anita Langerød, Markus Ringnér, Sung-Min Ahn, Sandrine Boyault, Jane E Brock, Annegien Broeks, Adam Butler, Christine Desmedt, Luc Dirix, Serge Dronov, Aquila Fatima, John A Foekens, Moritz Gerstung, Gerrit K J Hooijer, Se Jin Jang, David R Jones, Hyung-Yong Kim, Tari A King, Savitri Krishnamurthy, Hee Jin Lee, Jeong-Yeon Lee, Yilong Li, Stuart McLaren, Andrew Menzies, Ville Mustonen, Sarah O'Meara, Iris Pauporté, Xavier Pivot, Colin A Purdie, Keiran Raine, Kamna Ramakrishnan, F Germán Rodríguez-González, Gilles Romieu,

Anieta M Sieuwerts, Peter T Simpson, Rebecca Shepherd, Lucy Stebbings, Olafur A Stefansson, Jon Teague, Stefania Tommasi, Isabelle Treilleux, Gert G Van den Eynden, Peter Vermeulen, Anne Vincent-Salomon, Lucy Yates, Carlos Caldas, Laura van't Veer, Andrew Tutt, Stian Knappskog, Benita Kiat Tee Tan, Jos Jonkers, Åke Borg, Naoto T Ueno, Christos Sotiriou, Alain Viari, P Andrew Futreal, Peter J Campbell, Paul N Span, Steven Van Laere, Sunil R Lakhani, Jorunn E Eyfjord, Alastair M Thompson, Ewan Birney, Hendrik G Stunnenberg, Marc J van de Vijver, John W M Martens, Anne-Lise Børresen-Dale, Andrea L Richardson, Gu Kong, Gilles Thomas, and Michael R Stratton. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, June 2016.

[18] Nicholas A Willis, Richard L Frock, Francesca Menghi, Erin E Duffey, Arvind Panday, Virginia Camacho, E Paul Hasty, Edison T Liu, Frederick W Alt, and Ralph Scully. Mechanism of tandem duplication formation in BRCA1-mutant cells. *Nature*, 551(7682):590–595, November 2017.

[19] Gad Getz, Holger Höfling, Jill P Mesirov, Todd R Golub, Matthew Meyerson, Robert Tibshirani, and Eric S Lander. Comment on "the consensus coding sequences of human breast and colorectal cancers". *Science*, 317(5844):1500, September 2007.

[20] Michael S. Lawrence, Petar Stojanov, Paz Polak, Gregory V. Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L. Carter, Chip Stewart, Craig H. Mermel, Steven A. Roberts, Adam Kiezun, Peter S. Hammerman, Aaron McKenna, Yotam Drier, Lihua Zou, Alex H. Ramos, Trevor J. Pugh, Nicolas Stransky, Elena Helman, Jaegil Kim, Carrie Sougnez, Lauren Ambrogio, Elizabeth Nickerson, Erica Shefler, Maria L. Cortés, Daniel Auclair, Gordon Saksena, Douglas Voet, Michael Noble, Daniel DiCara, Pei Lin, Lee Lichtenstein, David I. Heiman, Timothy Fennell, Marcin Imielinski, Bryan Hernandez, Eran Hodis, Sylvan Baca, Austin M. Dulak, Jens Lohr, Dan-Avi Landau, Catherine J. Wu, Jorge Melendez-Zajgla, Alfredo Hidalgo-Miranda, Amnon Koren, Steven A. McCarroll, Jaume Mora, Ryan S. Lee, Brian Crompton, Robert Onofrio, Melissa Parkin, Wendy Winckler, Kristin Ardlie, Stacey B. Gabriel, Charles W. M. Roberts, Jaclyn A. Biegel, Kimberly Stegmaier, Adam J. Bass, Levi A. Garraway, Matthew Meyerson, Todd R. Golub, Dmitry A. Gordenin, Shamil Sunyaev, Eric S. Lander, and Gad Getz. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499:214–218, 6 2013.

[21] Iñigo Martincorena, Keiran M Raine, Moritz Gerstung, Kevin J Dawson,

Kerstin Haase, Peter Van Loo, Helen Davies, Michael R Stratton, and Peter J Campbell. Universal patterns of selection in cancer and somatic tissues. *Cell*, October 2017.

[22] Ludmil B Alexandrov, Young Seok Ju, Kerstin Haase, Peter Van Loo, Iñigo Martincorena, Serena Nik-Zainal, Yasushi Totoki, Akihiro Fujimoto, Hidewaki Nakagawa, Tatsuhiro Shibata, Peter J Campbell, Paolo Vineis, David H Phillips, and Michael R Stratton. Mutational signatures associated with tobacco smoking in human cancer. *Science*, 354(6312):618–622, November 2016.

[23] Wei Jiao, Gurnit Atwal, Paz Polak, Rosa Karlic, Edwin Cuppen, PCAWG Tumor Subtypes and Clinical Translation Working Group, Alexandra Danyi, Jeroen de Ridder, Carla van Herpen, Martijn P Lolkema, Neeltje Steeghs, Gad Getz, Quaid Morris, Lincoln D Stein, and PCAWG Consortium. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat. Commun.*, 11(1):728, February 2020.

[24] Helen Davies, Dominik Glodzik, Sandro Morganella, Lucy R Yates, Johan Staaf, Xueqing Zou, Manasa Ramakrishna, Sancha Martin, Sandrine Boyault, Anieta M Sieuwerts, Peter T Simpson, Tari A King, Keiran Raine, Jorunn E Eyfjord, Gu Kong, Ake Borg, Ewan Birney, Hendrik G Stunnenberg, Marc J van de Vijver, Anne-Lise Børresen-Dale, John W M Martens, Paul N Span, Sunil R Lakhani, Anne Vincent-Salomon, Christos Sotiriou, Andrew Tutt, Alastair M Thompson, Steven Van Laere, Andrea L Richardson, Alain Viari, Peter J Campbell, Michael R Stratton, and Serena Nik-Zainal. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nature Medicine*, 23:nm.4292, 2017.

[25] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel A J R Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolo Bolli, Ake Borg, Anne-Lise Borresen-Dale, Sandrine Boyault, Birgit Burkhardt, Adam P Butler, Carlos Caldas, Helen R Davies, Christine Desmedt, Roland Eils, Jorunn Erla Eyfjord, John A Foekens, Mel Greaves, Fumie Hosoda, Barbara Hutter, Tomislav Ilicic, Sandrine Imbeaud, Marcin Imielinsk, Natalie Jager, David T W Jones, David Jones, Stian Knappskog, Marcel Kool, Sunil R Lakhani, Carlos Lopez-Otin, Sancha Martin, Nikhil C Munshi, Hiromi Nakamura, Paul A Northcott, Marina Pajic, Elli Papaemmanuil, Angelo Paradiso, John V Pearson, Xose S Puente, Keiran Raine, Manasa Ramakrishna, Andrea L Richardson, Julia Richter, Philip Rosenstiel, Matthias Schlesner, Ton N Schumacher, Paul N Span, Jon W Teague, Yasushi Totoki, Andrew N J Tutt, Rafael Valdes-Mas,

Marit M van Buuren, Laura van /'t Veer, Anne Vincent-Salomon, Nicola Waddell, Lucy R Yates, Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, Jessica Zucman-Rossi, P Andrew Futreal, Ultan McDermott, Peter Lichter, Matthew Meyerson, Sean M Grimmond, Reiner Siebert, Elias Campo, Tatsuhiro Shibata, Stefan M Pfister, Peter J Campbell, and Michael R Stratton. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, August 2013.

[26] Jungmin Choi, Aranzazu Manzano, Weilai Dong, Stefania Bellone, Elena Bonazzoli, Luca Zammataro, Xiaotong Yao, Aditya Deshpande, Samir Zaidi, Adele Guglielmi, Barbara Gnutti, Nupur Nagarkatti, Joan R Tymon-Rosario, Justin Harold, Dennis Mauricio, Burak Zeybek, Gulden Menderes, Gary Altwerger, Kyungjo Jeong, Siming Zhao, Natalia Buza, Pei Hui, Antonella Ravaggi, Eliana Bignotti, Chiara Romani, Paola Todeschini, Laura Zanotti, Franco Odicino, Sergio Pecorelli, Laura Ardighieri, Kaya Bilguvar, Charles M Quick, Dan-Arin Silasi, Gloria S Huang, Vaagn Andikyan, Mitchell Clark, Elena Ratner, Masoud Azodi, Marcin Imielinski, Peter E Schwartz, Ludmil B Alexandrov, Richard P Lifton, Joseph Schlessinger, and Alessandro D Santin. Integrated mutational landscape analysis of uterine leiomyosarcomas. *Proc. Natl. Acad. Sci. U. S. A.*, 118(15):e2025182118, April 2021.

[27] Ludmil B Alexandrov, Jaegil Kim, Nicholas J Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, Kyle R Covington, Dmitry A Gordenin, Erik N Bergstrom, S M Ashiqul Islam, Nuria Lopez-Bigas, Leszek J Klimczak, John R McPherson, Sandro Morganella, Radhakrishnan Sabarinathan, David A Wheeler, Ville Mustonen, PCAWG Mutational Signatures Working Group, Gad Getz, Steven G Rozen, Michael R Stratton, and PCAWG Consortium. The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101, February 2020.

[28] Neil T Umbreit, Cheng-Zhong Zhang, Luke D Lynch, Logan J Blaine, Anna M Cheng, Richard Tourdot, Lili Sun, Hannah F Almubarak, Kim Judge, Thomas J Mitchell, Alexander Spektor, and David Pellman. Mechanisms generating cancer genome complexity from a single cell division error. *Science*, 368(6488), April 2020.

[29] Ofer Shoshani, Simon F Brunner, Rona Yaeger, Peter Ly, Yael Nechemia-Arbely, Dong Hyun Kim, Rongxin Fang, Guillaume A Castillon, Miao Yu, Julia S Z Li, Ying Sun, Mark H Ellisman, Bing Ren, Peter J Campbell, and Don W Cleveland. Chromothripsis drives the evolution of gene amplification in cancer. *Nature*, December 2020.

[30] Peter Ly, Simon F Brunner, Ofer Shoshani, Dong Hyun Kim, Weijie Lan, Tatyana Pyntikova, Adrienne M Flanagan, Sam Behjati, David C Page, Peter J Campbell, and Don W Cleveland. Chromosome segregation errors generate a diverse spectrum of simple and complex genomic rearrangements. *Nat. Genet.*, March 2019.

[31] David C Wedge, Gunes Gundem, Thomas Mitchell, Dan J Woodcock, Inigo Martincorena, Mohammed Ghori, Jorge Zamora, Adam Butler, Hayley Whitaker, Zsofia Kote-Jarai, Ludmil B Alexandrov, Peter Loo, Charlie E Massie, Stefan Dentro, Anne Y Warren, Clare Verrill, Dan M Berney, Nening Dennis, Sue Merson, Steve Hawkins, William Howat, Yong-Jie Lu, Adam Lambert, Jonathan Kay, Barbara Kremeyer, Katalin Karaszi, Hayley Luxton, Niedzica Camacho, Luke Marsden, Sandra Edwards, Lucy Matthews, Valeria Bo, Daniel Leongamornlert, Stuart McLaren, Anthony Ng, Yongwei Yu, Hongwei Zhang, Tokhir Dadaev, Sarah Thomas, Douglas F Easton, Mahbubl Ahmed, Elizabeth Bancroft, Cyril Fisher, Naomi Livni, David Nicol, Simon Tavaré, Pelvender Gill, Christopher Greenman, Vincent Khoo, Nicholas As, Pardeep Kumar, Christopher Ogden, Declan Cahill, Alan Thompson, Erik Mayer, Edward Rowe, Tim Dudderidge, Vincent Gnanapragasam, Nimish C Shah, Keiran Raine, David Jones, Andrew Menzies, Lucy Stebbings, Jon Teague, Steven Hazell, Cathy Corbishley, Johann Bono, Gerhardt Attard, William Isaacs, Tapio Visakorpi, Michael Fraser, Paul C Boutros, Robert G Bristow, Paul Workman, Chris Sander, Freddie C Hamdy, Andrew Futreal, Ultan McDermott, Bissan Al-Lazikani, Andrew G Lynch, G Steven Bova, Christopher S Foster, Daniel S Brewer, David E Neal, Colin S Cooper, and Rosalind A Eeles. Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets. *Nat. Genet.*, page 1, April 2018.

[32] Peter Priestley, Jonathan Baber, Martijn Lolkema, Neeltje Steeghs, Ewart de Bruijn, Korneel Duyvesteyn, Susan Haidari, Arne van Hoeck, Wendy Onstenk, Paul Roepman, Charles Shale, Mircea Voda, Haiko Bloemendal, Vivianne Tjan-Heijnen, Carla van Herpen, Mariette Labots, Petronella Witteveen, Egbert Smit, Stefan Sleijfer, Emile Voest, and Edwin Cuppen. Pan-cancer whole genome analyses of metastatic solid tumors. *bioRxiv*, 2019. Published online August 12, 2019.

[33] Philip J. Stephens, Chris D. Greenman, Beiyuan Fu, Fengtang Yang, Graham R. Bignell, Laura J. Mudie, Erin D. Pleasance, King Wai Lau, David Beare, Lucy A. Stebbings, Stuart McLaren, Meng-Lay Lin, David J. McBride, Ignacio Varela, Serena Nik-Zainal, Catherine Leroy, Mingming Jia, Andrew Menzies, Adam P. Butler, Jon W. Teague, Michael A. Quail, John Burton, Harold Swerdlow, Nigel P. Carter, Laura A. Morsberger,

Christine Iacobuzio-Donahue, George A. Follows, Anthony R. Green, Adrienne M. Flanagan, Michael R. Stratton, P. Andrew Futreal, and Peter J. Campbell. Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell.*, 144:27–40, 2011.

[34] Isidro Cortés-Ciriano, Jake June-Koo Lee, Ruibin Xi, Dhawal Jain, Youngsook L Jung, Lixing Yang, Dmitry Gordenin, Leszek J Klimczak, Cheng-Zhong Zhang, David S Pellman, PCAWG Structural Variation Working Group, Peter J Park, and PCAWG Consortium. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat. Genet.*, February 2020.

[35] Sylvan C. Baca, Davide Prandi, Michael S. Lawrence, Juan Miguel Mosquera, Alessandro Romanel, Yotam Drier, Kyung Park, Naoki Kitabayashi, Theresa Y. MacDonald, Mahmoud Ghandi, Eliezer Van Allen, Gregory V. Kryukov, Andrea Sboner, Jean-Philippe Theurillat, T. David Soong, Elizabeth Nickerson, Daniel Auclair, Ashutosh Tewari, Himisha Beltran, Robert C. Onofrio, Gunther Boysen, Candace Guiducci, Christopher E. Barbieri, Kristian Cibulskis, Andrey Sivachenko, Scott L. Carter, Gordon Saksena, Douglas Voet, Alex H. Ramos, Wendy Winckler, Michelle Cipicchio, Kristin Ardlie, Philip W. Kantoff, Michael F. Berger, Stacey B. Gabriel, Todd R. Golub, Matthew Meyerson, Eric S. Lander, Olivier Elemento, Gad Getz, Francesca Demichelis, Mark A. Rubin, and Levi A. Garraway. Punctuated Evolution of Prostate Cancer Genomes. *Cell.*, 153:666–677, 2013.

[36] Michael C Haffner, Martin J Aryee, Antoun Toubaji, David M Esopi, Roula Albadine, Bora Gurel, William B Isaacs, G Steven Bova, Wennuan Liu, Jianfeng Xu, Alan K Meeker, George Netto, Angelo M De Marzo, William G Nelson, and Srinivasan Yegnasubramanian. Androgen-induced TOP2B-mediated double-strand breaks and prostate cancer gene rearrangements. *Nat. Genet.*, 42(8):668–675, August 2010.

[37] Konstantin Helmsauer, Maria E Valieva, Salaheddine Ali, Rocío Chamorro González, Robert Schöpflin, Claudia Röefzaad, Yi Bei, Heathcliff Dorado Garcia, Elias Rodriguez-Fos, Montserrat Puiggròs, Katharina Kasack, Kerstin Haase, Csilla Keskeny, Celine Y Chen, Luis P Kuschel, Philipp Euskirchen, Verena Heinrich, Michael I Robson, Carolina Rosswog, Joern Toedling, Annabell Szymansky, Falk Hertwig, Matthias Fischer, David Torrents, Angelika Eggert, Johannes H Schulte, Stefan Mundlos, Anton G Henssen, and Richard P Koche. Enhancer hijacking determines extrachromosomal circular MYCN amplicon architecture in neuroblastoma. *Nat. Commun.*, 11(1):1–12, November 2020.

[38] Sihan Wu, Kristen M Turner, Nam Nguyen, Ramya Raviram, Marcella Erb, Jennifer Santini, Jens Luebeck, Utkrisht Rajkumar, Yarui Diao, Bin Li, Wenjing Zhang, Nathan Jameson, M Ryan Corces, Jeffrey M Granja, Xingqi Chen, Ceyda Coruh, Armen Abnousi, Jack Houston, Zhen Ye, Rong Hu, Miao Yu, Hoon Kim, Julie A Law, Roel G W Verhaak, Ming Hu, Frank B Furnari, Howard Y Chang, Bing Ren, Vineet Bafna, and Paul S Mischel. Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature*, November 2019.

[39] Hoon Kim, Nam-Phuong Nguyen, Kristen Turner, Sihan Wu, Amit D Gujar, Jens Luebeck, Jihe Liu, Viraj Deshpande, Utkrisht Rajkumar, Sandeep Namburi, Samirkumar B Amin, Eunhee Yi, Francesca Menghi, Johannes H Schulte, Anton G Henssen, Howard Y Chang, Christine R Beck, Paul S Mischel, Vineet Bafna, and Roel G W Verhaak. Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nat. Genet.*, August 2020.

[40] J W Gaubatz. Extrachromosomal circular DNAs and genomic sequence plasticity in eukaryotic cells. *Mutat. Res.*, 237(5-6):271–292, September 1990.

[41] Yilong Li, Nicola D Roberts, Jeremiah A Wala, Ofer Shapira, Steven E Schumacher, Kiran Kumar, Ekta Khurana, Sebastian Waszak, Jan O Korbel, James E Haber, Marcin Imielinski, PCAWG Structural Variation Working Group, Joachim Weischenfeldt, Rameen Beroukhim, Peter J Campbell, and PCAWG Consortium. Patterns of somatic structural variation in human cancer genomes. *Nature*, 578(7793):112–121, February 2020.

[42] Serena Nik-Zainal, Ludmil B Alexandrov, David C Wedge, Peter Van Loo, Christopher D Greenman, Keiran Raine, David Jones, Jonathan Hinton, John Marshall, Lucy A Stebbings, Andrew Menzies, Sancha Martin, Kenric Leung, Lina Chen, Catherine Leroy, Manasa Ramakrishna, Richard Rance, King Wai Lau, Laura J Mudie, Ignacio Varela, David J McBride, Graham R Bignell, Susanna L Cooke, Adam Shlien, John Gamble, Ian Whitmore, Mark Maddison, Patrick S Tarpey, Helen R Davies, Elli Papaemmanuil, Philip J Stephens, Stuart McLaren, Adam P Butler, Jon W Teague, Göran Jönsson, Judy E Garber, Daniel Silver, Penelope Miron, Aquila Fatima, Sandrine Boyault, Anita Langerød, Andrew Tutt, John W M Martens, Samuel A J R Aparicio, Åke Borg, Anne Vincent Salomon, Gilles Thomas, Anne-Lise Børresen-Dale, Andrea L Richardson, Michael S Neuberger, P Andrew Futreal, Peter J Campbell, Michael R Stratton, and Breast Cancer Working Group of the International Cancer Genome Con-

sortium. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993, May 2012.

[43] Francesca Menghi, Floris P Barthel, Vinod Yadav, Ming Tang, Bo Ji, Zhonghui Tang, Gregory W Carter, Yijun Ruan, Ralph Scully, Roel G W Verhaak, Jos Jonkers, and Edison T Liu. The tandem duplicator phenotype is a prevalent Genome-Wide cancer configuration driven by distinct gene mutations. *Cancer Cell*, 34(2):197–210.e5, August 2018.

[44] Julie M Behr, Xiaotong Yao, Kevin Hadi, Huasong Tian, Aditya Deshpande, Joel Rosiene, Titia de Lange, and Marcin Imieliński. Loose ends in cancer genome structure. *bioRxiv*, page 2021.05.26.445837, May 2021.

[45] Roel G. W. Verhaak, Vineet Bafna, and Paul S. Mischel. Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. *Nature Reviews Cancer*, 19:283–288, 2019.

[46] Jeremiah A Wala, Pratiti Bandopadhayay, Noah Greenwald, Ryan O'Rourke, Ted Sharpe, Chip Stewart, Steve Schumacher, Yilong Li, Joachim Weischenfeldt, Xiaotong Yao, Chad Nusbaum, Peter Campbell, Gad Getz, Matthew Meyerson, Cheng-Zhong Zhang, Marcin Imielinski, and Rameen Beroukhim. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome research*, 28:581–591, 2018.

[47] Daniel L Cameron, Jan Schröder, Jocelyn Sietsma Penington, Hongdo Do, Ramyar Molania, Alexander Dobrovic, Terence P Speed, and Anthony T Papenfuss. GRIDSS: sensitive and specific genomic rearrangement detection using positional de bruijn graph assembly. *Genome Res.*, 27(12):2050–2060, December 2017.

[48] Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M Stütz, Vladimir Benes, and Jan O Korbel. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, September 2012.

[49] Jianmin Wang, Charles G Mullighan, John Easton, Stefan Roberts, Sue L Heatley, Jing Ma, Michael C Rusch, Ken Chen, Christopher C Harris, Li Ding, Linda Holmfeldt, Debbie Payne-Turner, Xian Fan, Lei Wei, David Zhao, John C Obenauer, Clayton Naeve, Elaine R Mardis, Richard K Wilson, James R Downing, and Jinghui Zhang. CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods*, 8(8):652–654, June 2011.

[50] Ryan M Layer, Colby Chiang, Aaron R Quinlan, and Ira M Hall. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.*, 15(6):R84, June 2014.

[51] Gavin Ha, Andrew Roth, Jaswinder Khattra, Julie Ho, Damian Yap, Leah M Prentice, Nataliya Melnyk, Andrew McPherson, Ali Bashashati, Emma Laks, Justina Biele, Jiarui Ding, Alan Le, Jamie Rosner, Karey Shumansky, Marco A Marra, C Blake Gilks, David G Huntsman, Jessica N McAlpine, Samuel Aparicio, and Sohrab P Shah. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.*, 24(11):1881–1893, November 2014.

[52] Ruibin Xi, Angela G Hadjipanayis, Lovelace J Luquette, Tae-Min Kim, Eunjung Lee, Jianhua Zhang, Mark D Johnson, Donna M Muzny, David A Wheeler, Richard A Gibbs, Raju Kucherlapati, and Peter J Park. Copy number variation detection in whole-genome sequencing data using the bayesian information criterion. *Proc. Natl. Acad. Sci. U. S. A.*, 108(46):E1128–36, November 2011.

[53] Scott L Carter, Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W Laird, Robert C Onofrio, Wendy Winckler, Barbara A Weir, Rameen Beroukhim, David Pellman, Douglas A Levine, Eric S Lander, Matthew Meyerson, and Gad Getz. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.*, 30(5):413–421, May 2012.

[54] F Favero, T Joshi, A M Marquard, N J Birkbak, M Krzystanek, Q Li, Z Szallasi, and A C Eklund. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.*, 26(1):64–70, January 2015.

[55] Peter Van Loo, Silje H Nordgard, Ole Christian Lingjærde, Hege G Russnes, Inga H Rye, Wei Sun, Victor J Weigman, Peter Marynen, Anders Zetterberg, Bjørn Naume, Charles M Perou, Anne-Lise Børresen-Dale, and Vessela N Kristensen. Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U. S. A.*, 107(39):16910–16915, September 2010.

[56] Ronglai Shen and Venkatraman E Seshan. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.*, 44(16):e131, September 2016.

[57] Yuval Benjamini and Terence P. Speed. Summarizing and correcting the

gc content bias in high-throughput sequencing. *Nucleic acids research*, 40(10):e72, May 2012.

[58] Hayan Lee and Michael C. Schatz. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics*, 28(16):2097–2105, Aug 2012.

[59] Xiang Chen, Pankaj Gupta, Jianmin Wang, Joy Nakitandwe, Kathryn Roberts, James D Dalton, Matthew Parker, Samir Patel, Linda Holmfeldt, Debbie Payne, John Easton, Jing Ma, Michael Rusch, Gang Wu, Aman Patel, Suzanne J Baker, Michael A Dyer, Sheila Shurtleff, Stephen Espy, Stanley Pounds, James R Downing, David W Ellison, Charles G Mullighan, and Jinghui Zhang. CONSERTING: integrating copy-number analysis with structural-variation detection. *Nat. Methods*, 12(6):527–530, June 2015.

[60] Brent S Pedersen and Aaron R Quinlan. Duphold: scalable, depth-based annotation and curation of high-confidence structural variant calls. *Gigascience*, 8(4), April 2019.

[61] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, Gilean McVean, Richard Durbin, and 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, August 2011.

[62] Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18(11):1851–1858, November 2008.

[63] H Li, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, R Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[64] Ryan L Collins, Harrison Brand, Konrad J Karczewski, Xuefang Zhao, Jessica Alföldi, Laurent C Francioli, Amit V Khera, Chelsea Lowther, Laura D Gauthier, Harold Wang, Nicholas A Watts, Matthew Solomonson, Anne O'Donnell-Luria, Alexander Baumann, Ruchi Munshi, Mark Walker, Christopher W Whelan, Yongqing Huang, Ted Brookings, Ted Sharpe, Matthew R Stone, Elise Valkanas, Jack Fu, Grace Tiao, Kristen M Laricchia, Valentin Ruano-Rubio, Christine Stevens, Namrata Gupta, Caroline Cusick, Lauren Margolin, Genome Aggregation Database Produc-

tion Team, Genome Aggregation Database Consortium, Kent D Taylor, Henry J Lin, Stephen S Rich, Wendy S Post, Yii-Der Ida Chen, Jerome I Rotter, Chad Nusbaum, Anthony Philippakis, Eric Lander, Stacey Gabriel, Benjamin M Neale, Sekar Kathiresan, Mark J Daly, Eric Banks, Daniel G MacArthur, and Michael E Talkowski. A structural variation reference for medical and population genetics. *Nature*, 581(7809):444–451, May 2020.

[65] Claudia M B Carvalho, Melissa B Ramocki, Davut Pehlivan, Luis M Franco, Claudia Gonzaga-Jauregui, Ping Fang, Alanna McCall, Eniko Karman Pivnick, Stacy Hines-Dowell, Laurie H Seaver, Linda Friehling, Sansan Lee, Rosemarie Smith, Daniela Del Gaudio, Marjorie Withers, Pengfei Liu, Sau Wai Cheung, John W Belmont, Huda Y Zoghbi, P J Hastings, and James R Lupski. Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nat. Genet.*, 43(11):1074–1081, October 2011.

[66] Tyler Funnell, Allen W Zhang, Diljot Grewal, Steven McKinney, Ali Bashashati, Yi Kan Wang, and Sohrab P Shah. Integrated structural variation and point mutation signatures in cancer genomes using correlated topic models. *PLoS Comput. Biol.*, 15(2):e1006799, February 2019.

[67] John Maciejowski and Marcin Imielinski. Modeling cancer rearrangement landscapes: from pattern to mechanism, and back. *Current Opinion in Systems Biology*, 2016.

[68] Daniel R Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Res.*, 18(5):821–829, May 2008.

[69] Jared T Simpson and Richard Durbin. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics*, 26(12):i367–73, June 2010.

[70] H Li. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly, 2012.

[71] Erik Garrison, Jouni Sirén, Adam M Novak, Glenn Hickey, Jordan M Eizenga, Eric T Dawson, William Jones, Shilpa Garg, Charles Markello, Michael F Lin, Benedict Paten, and Richard Durbin. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.*, August 2018.

[72] Glenn Hickey, David Heller, Jean Monlong, Jonas A Sibbesen, Jouni Sirén,

Jordan Eizenga, Eric T Dawson, Erik Garrison, Adam M Novak, and Benedict Paten. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.*, 21(1):35, February 2020.

[73] Jouni Sirèn, Jean Monlong, Xian Chang, Adam M Novak, Jordan M Eizenga, Charles Markello, Jonas A Sibbesen, Glenn Hickey, Pi-Chuan Chang, Andrew Carroll, David Haussler, Erik Garrison, and Benedict Paten. Genotyping common, large structural variations in 5,202 genomes using pangenomes, the giraffe mapper, and the vg toolkit. *Cold Spring Harbor Laboratory*, page 2020.12.04.412486, December 2020.

[74] Heng Li, Xiaowen Feng, and Chong Chu. The design and construction of reference pangenome graphs with minigraph. *Genome Biol.*, 21(1):265, October 2020.

[75] Max A Alekseyev and Pavel A Pevzner. Breakpoint graphs and ancestral genome reconstructions. *Genome Res.*, 19(5):943–957, May 2009.

[76] Pavel Avdeyev, Shuai Jiang, Sergey Aganezov, Fei Hu, and Max A Alekseyev. Reconstruction of ancestral genomes in presence of gene gain and loss. *J. Comput. Biol.*, 23(3):150–164, March 2016.

[77] C D Greenman, E D Pleasance, S Newman, F Yang, B Fu, S Nik-Zainal, D Jones, K W Lau, N Carter, P A W Edwards, P A Futreal, M R Stratton, and P J Campbell. Estimation of rearrangement phylogeny for cancer genomes. *Genome Research*, 22:346–361, 2012.

[78] Layla Oesper, Anna Ritz, Sarah J Aerni, Ryan Drebin, and Benjamin J Raphael. Reconstructing cancer genomes from paired-end sequencing data. *BMC Bioinformatics*, 13(6):S10, 2012.

[79] Misko Dzamba, Arun K Ramani, Pawel Buczkowicz, Yue Jiang, Man Yu, Cynthia Hawkins, and Michael Brudno. Identification of complex genomic rearrangements in cancers using CouGaR. *Genome Res.*, 27(1):107–117, January 2017.

[80] Viraj Deshpande, Jens Luebeck, Nam-Phuong D Nguyen, Mehrdad Bakhtiari, Kristen M Turner, Richard Schwab, Hannah Carter, Paul S Mischel, and Vineet Bafna. Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. *Nat. Commun.*, 10(1):392, January 2019.

[81] Sergey Aganezov, Ilya Zban, Vitaly Aksenov, Nikita Alexeev, and

Michael C Schatz. Recovering rearranged cancer chromosomes from karyotype graphs. *BMC Bioinformatics*, 20(Suppl 20):641, December 2019.

[82] Yeonghun Lee and Hyunju Lee. Integrative reconstruction of cancer genome karyotypes using InfoGenomeR. *Nat. Commun.*, 12(1):2467, April 2021.

[83] Kevin Hadi, Xiaotong Yao, Julie M Behr, Aditya Deshpande, Charalampos Xanthopoulakis, Huasong Tian, Sarah Kudman, Joel Rosiene, Madison Darmofal, Joseph DeRose, Rick Mortensen, Emily M Adney, Alon Shaiber, Zoran Gajic, Michael Sigouros, Kenneth Eng, Jeremiah A Wala, Kazimierz O Wrzeszczyński, Kanika Arora, Minita Shah, Anne-Katrin Emde, Vanessa Felice, Mayu O Frank, Robert B Darnell, Mahmoud Ghandi, Franklin Huang, Sally Dewhurst, John Maciejowski, Titia de Lange, Jeremy Setton, Nadeem Riaz, Jorge S Reis-Filho, Simon Powell, David A Knowles, Ed Reznik, Bud Mishra, Rameen Beroukhim, Michael C Zody, Nicolas Robine, Kenji M Oman, Carissa A Sanchez, Mary K Kuhner, Lucian P Smith, Patricia C Galipeau, Thomas G Paulson, Brian J Reid, Xiaohong Li, David Wilkes, Andrea Sboner, Juan Miguel Mosquera, Olivier Elemento, and Marcin Imielinski. Distinct classes of complex structural variation uncovered across thousands of cancer genome graphs. *Cell*, 183(1):197–210.e32, October 2020.

[84] Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T Morgan, and Vincent J Carey. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, 9(8):e1003118, August 2013.

[85] James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. Integrative genomics viewer. *Nat. Biotechnol.*, 29(1):24–26, January 2011.

[86] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at UCSC. *Genome Res.*, 12(6):996–1006, June 2002.

[87] Jonathan R Belyeu, Murad Chowdhury, Joseph Brown, Brent S Pedersen, Michael J Cormier, Aaron R Quinlan, and Ryan M Layer. Samplot: a platform for structural variant visual validation and automated filtering. *Genome Biol.*, 22(1):161, May 2021.

[88] Wolfgang Beyer, Adam M Novak, Glenn Hickey, Jeffrey Chan, Vanessa Tan, Benedict Paten, and Daniel R Zerbino. Sequence tube maps: making

graph genomes intuitive to commuters. *Bioinformatics*, 35(24):5318–5320, December 2019.

[89] Toshiyuki T Yokoyama, Yoshitaka Sakamoto, Masahide Seki, Yutaka Suzuki, and Masahiro Kasahara. MoMI-G: modular multi-scale integrated genome graph browser. *BMC Bioinformatics*, 20(1):548, November 2019.

[90] Maria Nattestad, Chen-Shan Chin, and Michael C Schatz. Ribbon: Visualizing complex genome alignments and structural variation. *bioRxiv*, page 082123, October 2016.

[91] Maria Nattestad, Sara Goodwin, Karen Ng, Timour Baslan, Fritz J Sedlazeck, Philipp Rescheneder, Tyler Garvin, Han Fang, James Gurtowski, Elizabeth Hutton, Elizabeth Tseng, Chen-Shan Chin, Timothy Beck, Yogi Sundaravadanam, Melissa Kramer, Eric Antoniou, John D McPherson, James Hicks, W Richard McCombie, and Michael C Schatz. Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res.*, 28(8):1126–1135, August 2018.

[92] Andrew W McPherson, Andrew Roth, Gavin Ha, Cedric Chauve, Adi Steif, Camila P E de Souza, Peter Eirew, Alexandre Bouchard-Côté, Sam Aparicio, S Cenk Sahinalp, and Sohrab P Shah. ReMixT: clone-specific genomic structure estimation in cancer. *Genome Biol.*, 18(1):140, July 2017.

[93] Yang Li, Shiguo Zhou, David C Schwartz, and Jian Ma. Allele-Specific quantification of structural variations in cancer genomes. *Cell Syst*, 3(1):21–34, July 2016.

[94] Paul Medvedev, Marc Fiume, Misko Dzamba, Tim Smith, and Michael Brudno. Detecting copy number variation with mated short reads. *Genome Research*, 20:1613–1622, 2010.

[95] Dale W. Garsed, Owen J. Marshall, Vincent D.A. Corbin, Arthur Hsu, Leon Di Stefano, Jan Schröder, Jason Li, Zhi-Ping Feng, Bo W. Kim, Mark Kowarsky, Ben Lansdell, Ross Brookwell, Ola Myklebost, Leonardo Meza-Zepeda, Andrew J. Holloway, Florence Pedeutour, K.H. Andy Choo, Michael A. Damore, Andrew J. Deans, Anthony T. Papenfuss, and David M. Thomas. The Architecture and Evolution of Cancer Neochromosomes. *Cancer Cell*, 26:653–667, 2014.

[96] Andrew V Goldberg and Alexander V Karzanov. Path problems in skew-symmetric graphs. *Combinatorica*, 16(3):353–382, September 1996.

[97] Jeremiah A Wala, Pratiti Bandopadhayay, Noah Greenwald, Ryan O'Rourke, Ted Sharpe, Chip Stewart, Steve Schumacher, Yilong Li, Joachim Weischenfeldt, Xiaotong Yao, Chad Nusbaum, Peter Campbell, Gad Getz, Matthew Meyerson, Cheng-Zhong Zhang, Marcin Imielinski, and Rameen Beroukhim. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.*, March 2018.

[98] Timo R. Maarleveld, Ruchir A. Khandelwal, Brett G. Olivier, Bas Teusink, and Frank J. Bruggeman. Basic concepts and principles of stoichiometric modeling of metabolic networks. *Biotechnology Journal*, 8(9):997–1008, 2013.

[99] G Ha, A Roth, D Lai, A Bashashati, J Ding, R Goya, R Giuliany, J Rosner, A Oloumi, K Shumansky, S F Chin, G Turashvili, M Hirst, C Caldas, M A Marra, S Aparicio, and S P Shah. Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Research*, 22(10):1995–2007, 2012.

[100] Adam B. Olshen, E. S. Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5:557–572, 2004.

[101] Yilong Li, Nicola Roberts, Joachim Weischenfeldt, Jeremiah Anthony Wala, Ofer Shapira, Steven Schumacher, Ekta Khurana, Jan O Korbel, Marcin Imielinski, Rameen Beroukhim, and Peter Campbell. Patterns of structural variation in human cancer. *bioRxiv*, page 181339, 2017. Published online August 27, 2017.

[102] Sergey Aganezov and Benjamin J Raphael. Reconstruction of clone- and haplotype-specific cancer genome karyotypes from bulk tumor samples. *Genome Res.*, 30(9):1274–1290, September 2020.

[103] Zechen Chong, Jue Ruan, Min Gao, Wanding Zhou, Tenghui Chen, Xian Fan, Li Ding, Anna Y Lee, Paul Boutros, Junjie Chen, and Ken Chen. novobreak: local assembly for breakpoint detection in cancer genomes. *Nat. Methods*, 14(1):65–67, January 2017.

[104] Jordan M Eizenga, Adam M Novak, Emily Kobayashi, Flavia Villani, Cecilia Cisar, Simon Heumos, Glenn Hickey, Vincenza Colonna, Benedict

Paten, and Erik Garrison. Efficient dynamic variation graphs. *Bioinformatics*, July 2020.

[105] Daniel Butler, Christopher Mozsary, Cem Meydan, Jonathan Foox, Joel Rosiene, Alon Shaiber, David Danko, Ebrahim Afshinnekoo, Matthew MacKay, Fritz J Sedlazeck, Nikolay A Ivanov, Maria Sierra, Diana Pohle, Michael Zietz, Undina Gisladottir, Vijendra Ramlall, Evan T Sholle, Edward J Schenck, Craig D Westover, Ciaran Hassan, Krista Ryon, Benjamin Young, Chandrima Bhattacharya, Dianna L Ng, Andrea C Granados, Yale A Santos, Venice Servellita, Scot Federman, Phyllis Ruggiero, Arkarachai Fungtammasan, Chen-Shan Chin, Nathaniel M Pearson, Bradley W Langhorst, Nathan A Tanner, Youngmi Kim, Jason W Reeves, Tyler D Hether, Sarah E Warren, Michael Bailey, Justyna Gawrys, Dmitry Meleshko, Dong Xu, Mara Couto-Rodriguez, Dorottya Nagy-Szakal, Joseph Barrows, Heather Wells, Niamh B O'Hara, Jeffrey A Rosenfeld, Ying Chen, Peter A D Steel, Amos J Shemesh, Jenny Xiang, Jean Thierry-Mieg, Danielle Thierry-Mieg, Angelika Iftner, Daniela Bezdan, Elizabeth Sanchez, Thomas R Campion, Jr, John Sipley, Lin Cong, Arryn Craney, Priya Velu, Ari M Melnick, Sagi Shapira, Iman Hajirasouliha, Alain Borczuk, Thomas Iftner, Mirella Salvatore, Massimo Loda, Lars F Westblade, Melissa Cushing, Shixiu Wu, Shawn Levy, Charles Chiu, Robert E Schwartz, Nicholas Tatonetti, Hanna Rennert, Marcin Imielinski, and Christopher E Mason. Shotgun transcriptome, spatial omics, and isothermal profiling of SARS-CoV-2 infection reveals unique host responses, viral diversification, and drug interactions. *Nat. Commun.*, 12(1):1660, March 2021.

[106] Thomas Paulson, Patricia Galipeau, Kenji Oman, Carissa Sanchez, Mary Kuhner, Lucian Smith, Kevin Hadi, Minita Shah, Kanika Arora, Jennifer Shelton, Andre Corvelo, Molly Johnson, Anne-Kathrin Emde, Carlo Maley, Xiaotong Yao, Rashesh Sanghvi, Elisa Venturini, Benjamin Hubert, Marcin Imielinski, Nicolas Robine, Brian Reid, and Xiaohong Li. Somatic whole genome dynamics of precancer in barrett's esophagus. *Research Square*, March 2021.

[107] Ekta Khurana, Yao Fu, Vincenza Colonna, Xinmeng Jasmine Mu, Hyun Min Kang, Tuuli Lappalainen, Andrea Sboner, Lucas Lochovsky, Jieming Chen, Arif Harmanci, Jishnu Das, Alexej Abyzov, Suganthi Balasubramanian, Kathryn Beal, Dimple Chakravarty, Daniel Challis, Yuan Chen, Declan Clarke, Laura Clarke, Fiona Cunningham, Uday S Evani, Paul Flicek, Robert Fragoza, Erik Garrison, Richard Gibbs, Zeynep H Gümüş, Javier Herrero, Naoki Kitabayashi, Yong Kong, Kasper Lage, Vaja Liluashvili, Steven M Lipkin, Daniel G MacArthur, Gabor Marth, Donna

Muzny, Tune H Pers, Graham R S Ritchie, Jeffrey A Rosenfeld, Cristina Sisu, Xiaomu Wei, Michael Wilson, Yali Xue, Fuli Yu, 1000 Genomes Project Consortium, Emmanouil T Dermitzakis, Haiyuan Yu, Mark A Rubin, Chris Tyler-Smith, and Mark Gerstein. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science*, 342(6154):1235587, October 2013.

[108] Federica Panebianco, Lindsey M Kelly, Pengyuan Liu, Shan Zhong, Sanja Dacic, Xiaosong Wang, Aatur D Singhi, Rajiv Dhir, Simion I Chiosea, Shih-Fan Kuan, Rohit Bhargava, David Dabbs, Sumita Trivedi, Manoj Gandhi, Rachel Diaz, Abigail I Wald, Sally E Carty, Robert L Ferris, Adrian V Lee, Marina N Nikiforova, and Yuri E Nikiforov. THADA fusion is a mechanism of IGF2BP3 activation and IGF1R signaling in thyroid cancer. *Proc. Natl. Acad. Sci. U. S. A.*, 114(9):2307–2312, February 2017.

[109] Esther Rheinbay, Morten Muhlig Nielsen, Federico Abascal, Jeremiah A Wala, Ofer Shapira, Grace Tiao, Henrik Hornshøj, Julian M Hess, Randi Istrup Juul, Ziao Lin, Lars Feuerbach, Radhakrishnan Sabarinathan, Tobias Madsen, Jaegil Kim, Loris Mularoni, Shimin Shuai, Andrés Lanzós, Carl Herrmann, Yosef E Maruvka, Ciyue Shen, Samirkumar B Amin, Pratiti Bandopadhayay, Johanna Bertl, Keith A Boroevich, John Busanovich, Joana Carlevaro-Fita, Dimple Chakravarty, Calvin Wing Yiu Chan, David Craft, Priyanka Dhingra, Klev Diamanti, Nuno A Fonseca, Abel Gonzalez-Perez, Qianyun Guo, Mark P Hamilton, Nicholas J Haradhvala, Chen Hong, Keren Isaev, Todd A Johnson, Malene Juul, Andre Kahles, Abdullah Kahraman, Youngwook Kim, Jan Komorowski, Kiran Kumar, Sushant Kumar, Donghoon Lee, Kjong-Van Lehmann, Yilong Li, Eric Minwei Liu, Lucas Lochovsky, Keunchil Park, Oriol Pich, Nicola D Roberts, Gordon Saksena, Steven E Schumacher, Nikos Sidiropoulos, Lina Sieverling, Nasa Sinnott-Armstrong, Chip Stewart, David Tamborero, Jose M C Tubio, Husen M Umer, Liis Uusküla-Reimand, Claes Wadelius, Lina Wadi, Xiaotong Yao, Cheng-Zhong Zhang, Jing Zhang, James E Haber, Asger Hobolth, Marcin Imielinski, Manolis Kellis, Michael S Lawrence, Christian von Mering, Hidewaki Nakagawa, Benjamin J Raphael, Mark A Rubin, Chris Sander, Lincoln D Stein, Joshua M Stuart, Tatsuhiko Tsunoda, David A Wheeler, Rory Johnson, Jüri Reimand, Mark Gerstein, Ekta Khurana, Peter J Campbell, Núria López-Bigas, Joachim Weischenfeldt, Rameen Beroukhim, Iñigo Martincorena, Jakob Skou Pedersen, and Gad Getz. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*, 578(7793):102–111, February 2020.

[110] Thomas B K Watkins, Emilia L Lim, Marina Petkovic, Sergi Elizalde, Nicolai J Birkbak, Gareth A Wilson, David A Moore, Eva Grönroos, Andrew

Rowan, Sally M Dewhurst, Jonas Demeulemeester, Stefan C Dentro, Stuart Horswell, Lewis Au, Kerstin Haase, Mickael Escudero, Rachel Rosenthal, Maise Al Bakir, Hang Xu, Kevin Litchfield, Wei Ting Lu, Thanos P Mourikis, Michelle Dietzen, Lavinia Spain, George D Cresswell, Dhruva Biswas, Philippe Lamy, Iver Nordentoft, Katja Harbst, Francesc Castro-Giner, Lucy R Yates, Franco Caramia, Fanny Jaulin, Cécile Vicier, Ian P M Tomlinson, Priscilla K Brastianos, Raymond J Cho, Boris C Bastian, Lars Dyrskjøt, Göran B Jönsson, Peter Savas, Sherene Loi, Peter J Campbell, Fabrice Andre, Nicholas M Luscombe, Neeltje Steeghs, Vivianne C G Tjan-Heijnen, Zoltan Szallasi, Samra Turajlic, Mariam Jamal-Hanjani, Peter Van Loo, Samuel F Bakhoum, Roland F Schwarz, Nicholas McGranahan, and Charles Swanton. Pervasive chromosomal instability and karyotype order in tumour evolution. *Nature*, September 2020.

[111] Sally M Dewhurst, Xiaotong Yao, Joel Rosiene, Huasong Tian, Julie Behr, Nazario Bosco, Kaori K Takai, Titia de Lange, and Marcin Imieliński. Structural variant evolution after telomere crisis. *Nat. Commun.*, 12(1):2093, April 2021.

[112] Hans Zahn, Adi Steif, Emma Laks, Peter Eirew, Michael VanInsberghe, Sohrab P Shah, Samuel Aparicio, and Carl L Hansen. Scalable whole-genome single-cell library preparation without preamplification. *Nat. Methods*, 14(2):167–173, February 2017.

[113] Emma Laks, Andrew McPherson, Hans Zahn, Daniel Lai, Adi Steif, Jazmine Brimhall, Justina Biele, Beixi Wang, Tehmina Masud, Jerome Ting, Diljot Grewal, Cydney Nielsen, Samantha Leung, Viktoria Bojilova, Maia Smith, Oleg Golovko, Steven Poon, Peter Eirew, Farhia Kabeer, Teresa Ruiz de Algara, So Ra Lee, M. Jafar Taghiyar, Curtis Huebner, Jessica Ngo, Tim Chan, Spencer Vatrt-Watts, Pascale Walters, Nafis Abrar, Sophia Chan, Matt Wiens, Lauren Martin, R. Wilder Scott, T. Michael Underhill, Elizabeth Chavez, Christian Steidl, Daniel Da Costa, Yussanne Ma, Robin J.N. Coope, Richard Corbett, Stephen Pleasance, Richard Moore, Andrew J. Mungall, Colin Mar, Fergus Cafferty, Karen Gelmon, Stephen Chia, The CRUK IMAXT Grand Challenge Team, Gregory J. Hannon, Giorgia Battistoni, Dario Bressan, Ian Cannell, Hannah Casbolt, Cristina Jauset, Tatjana Kovačević, Claire Mulvey, Fiona Nugent, Marta Paez Ribes, Isabella Pearsall, Fatime Qosaj, Kirsty Sawicka, Sophia Wild, Elena Williams, Samuel Aparicio, Emma Laks, Yangguang Li, Ciara O'Flanagan, Austin Smith, Teresa Ruiz, Shankar Balasubramanian, Maximillian Lee, Bernd Bodenmiller, Marcel Burger, Laura Kuett, Sandra Tietscher, Jonas Windager, Edward Boyden, Shahar Alon, Yi Cui, Amauche Emenari, Dan Goodwin, Emmanouil Karagiannis, Anubhav Sinha, As-

mamaw T. Wassie, Carlos Caldas, Alejandra Bruna, Maurizio Callari, Wendy Greenwood, Giulia Lerda, Yaniv Lubling, Alastair Marti, Oscar Rueda, Abigail Shea, Owen Harris, Robby Becker, Flaminia Grimaldi, Suvi Harris, Sara Vogl, Johanna A. Joyce, Jean Hausser, Spencer Watson, Sorhab Shah, Andrew McPherson, Ignacio Vázquez-García, Simon Tavaré, Khanh Dinh, Eyal Fisher, Russell Kunes, Nicolas A. Walton, Mohammad Al Sa'd, Nick Chornay, Ali Dariush, Eduardo Gonzales Solares, Carlos Gonzalez-Fernandez, Aybuke Kupcu Yoldas, Neil Millar, Xiaowei Zhuang, Jean Fan, Hsuan Lee, Leonardo Sepulveda Duran, Chenglong Xia, Pu Zheng, Marco A. Marra, Carl Hansen, Sohrab P. Shah, and Samuel Aparicio. Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing. *Cell*, 179(5):1207–1221.e22, 2019.

[114] Sohrab Salehi, Farhia Kabeer, Nicholas Ceglia, Mirela Andronescu, Marc J Williams, Kieran R Campbell, Tehmina Masud, Beixi Wang, Justina Biele, Jazmine Brimhall, David Gee, Hakwoo Lee, Jerome Ting, Allen W Zhang, Hoa Tran, Ciara O'Flanagan, Fatemeh Dorri, Nicole Rusk, Teresa Ruiz de Algara, So Ra Lee, Brian Yu Chieh Cheng, Peter Eirew, Takako Kono, Jenifer Pham, Diljot Grewal, Daniel Lai, Richard Moore, Andrew J Mungall, Marco A Marra, Andrew McPherson, Alexandre Bouchard-Côté, Samuel Aparicio, and Sohrab P Shah. Clonal fitness inferred from time-series modelling of single-cell cancer genomes. *Nature*, pages 1–6, June 2021.

[115] Mahmoud Ghandi, Franklin W Huang, Judit Jané-Valbuena, Gregory V Kryukov, Christopher C Lo, E Robert McDonald, 3rd, Jordi Barretina, Ellen T Gelfand, Craig M Bielski, Haoxin Li, Kevin Hu, Alexander Y Andreev-Drakhlin, Jaegil Kim, Julian M Hess, Brian J Haas, François Aguet, Barbara A Weir, Michael V Rothberg, Brenton R Paolella, Michael S Lawrence, Rehan Akbani, Yiling Lu, Hong L Tiv, Prafulla C Gokhale, Antoine de Weck, Ali Amin Mansour, Coyin Oh, Juliann Shih, Kevin Hadi, Yanay Rosen, Jonathan Bistline, Kavitha Venkatesan, Anupama Reddy, Dmitriy Sonkin, Manway Liu, Joseph Lehar, Joshua M Korn, Dale A Porter, Michael D Jones, Javad Golji, Giordano Caponigro, Jordan E Taylor, Caitlin M Dunning, Amanda L Creech, Allison C Warren, James M McFarland, Mahdi Zamanighomi, Audrey Kauffmann, Nicolas Stransky, Marcin Imielinski, Yosef E Maruvka, Andrew D Cherniack, Aviad Tsherniak, Francisca Vazquez, Jacob D Jaffe, Andrew A Lane, David M Weinstock, Cory M Johannessen, Michael P Morrissey, Frank Stegmeier, Robert Schlegel, William C Hahn, Gad Getz, Gordon B Mills, Jesse S Boehm, Todd R Golub, Levi A Garraway, and William R Sellers. Next-generation characterization of the cancer cell line encyclopedia. *Nature*, 569(7757):503–508, May 2019.

[116] Alexander M. Frankell, SriGanesh Jammula, Xiaodun Li, Gianmarco Contino, Sarah Killcoyne, Sujath Abbas, Juliane Perner, Lawrence Bower, Ginny Devonshire, Emma Ococks, Nicola Grehan, James Mok, Maria O'Donovan, Shona MacRae, Matthew D. Eldridge, Simon Tavaré, and Rebecca C. Fitzgerald. The landscape of selection in 551 esophageal adenocarcinomas defines genomic biomarkers for the clinic. *Nature Genetics*, 51:506–516, 2019.

[117] Jake June-Koo Lee, Seongyeol Park, Hansol Park, Sehui Kim, Jongkeun Lee, Junehawk Lee, Jeonghwan Youk, Kijong Yi, Yohan An, In Kyu Park, Chang Hyun Kang, Doo Hyun Chung, Tae Min Kim, Yoon Kyung Jeon, Dongwan Hong, Peter J Park, Young Seok Ju, and Young Tae Kim. Tracing oncogene rearrangements in the mutational history of lung adenocarcinoma. *Cell*, 0(0), May 2019.

[118] Julia Richter, Matthias Schlesner, Steve Hoffmann, Markus Kreuz, Ellen Leich, Birgit Burkhardt, Maciej Rosolowski, Ole Ammerpohl, Rabea Wagener, Stephan H Bernhart, Dido Lenze, Monika Szczepanowski, Maren Paulsen, Simone Lipinski, Robert B Russell, Sabine Adam-Klages, Gordana Apic, Alexander Claviez, Dirk Hasenclever, Volker Hovestadt, Nadine Hornig, Jan O Korbel, Dieter Kube, David Langenberger, Chris Lawerenz, Jasmin Lisfeld, Katharina Meyer, Simone Picelli, Jordan Pischimarov, Bernhard Radlwimmer, Tobias Rausch, Marius Rohde, Markus Schilhabel, René Scholtysik, Rainer Spang, Heiko Trautmann, Thorsten Zenz, Arndt Borkhardt, Hans G Drexler, Peter Möller, Roderick A F MacLeod, Christiane Pott, Stefan Schreiber, Lorenz Trümper, Markus Loeffler, Peter F Stadler, Peter Lichter, Roland Eils, Ralf Küppers, Michael Hummel, Wolfram Klapper, Philip Rosenstiel, Andreas Rosenwald, Benedikt Brors, Reiner Siebert, and ICGC MMML-Seq Project. Recurrent mutation of the ID3 gene in burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. *Nat. Genet.*, 44(12):1316–1320, December 2012.

[119] Nicholas K. Hayward, James S. Wilmott, Nicola Waddell, Peter A. Johansson, Matthew A. Field, Katia Nones, Ann-Marie Patch, Hojabr Kakavand, Ludmil B. Alexandrov, Hazel Burke, Valerie Jakrot, Stephen Kazakoff, Oliver Holmes, Conrad Leonard, Radhakrishnan Sabarinathan, Loris Mularoni, Scott Wood, Qinying Xu, Nick Waddell, Varsha Tembe, Gulietta M. Pupo, Ricardo De Paoli-Iseppi, Ricardo E. Vilain, Ping Shang, Loretta M. S. Lau, Rebecca A. Dagg, Sarah-Jane Schramm, Antonia Pritchard, Ken Dutton-Regester, Felicity Newell, Anna Fitzgerald, Catherine A. Shang, Sean M. Grimmond, Hilda A. Pickett, Jean Y. Yang, Jonathan R. Stretch, Andreas Behren, Richard F. Kefford, Peter Hersey, Georgina V.

Long, Jonathan Cebon, Mark Shackleton, Andrew J. Spillane, Robyn P. M. Saw, Núria López-Bigas, John V. Pearson, John F. Thompson, Richard A. Scolyer, and Graham J. Mann. Whole-genome landscapes of major melanoma subtypes. *Nature*, 545:175, 2017.

[120] Marcin Imielinski, Alice H Berger, Peter S Hammerman, Bryan Hernandez, Trevor J Pugh, Eran Hodis, Jeonghee Cho, James Suh, Marzia Capelletti, Andrey Sivachenko, Carrie Sougnez, Daniel Auclair, Michael S Lawrence, Petar Stojanov, Kristian Cibulskis, Kyusam Choi, Luc de Waal, Tanaz Sharifnia, Angela Brooks, Heidi Greulich, Shantanu Banerji, Thomas Zander, Danila Seidel, Frauke Leenders, Sascha Ansén, Corinna Ludwig, Walburga Engel-Riedel, Erich Stoelben, Jürgen Wolf, Chandra Goparju, Kristin Thompson, Wendy Winckler, David Kwiatkowski, Bruce E Johnson, Pasi A Jänne, Vincent A Miller, William Pao, William D Travis, Harvey I Pass, Stacey B Gabriel, Eric S Lander, Roman K Thomas, Levi A Garraway, Gad Getz, and Matthew Meyerson. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*, 150(6):1107–1120, September 2012.

[121] Li Ding, Matthew H Bailey, Eduard Porta-Pardo, Vesteinn Thorsson, Antonio Colaprico, Denis Bertrand, David L Gibbs, Amila Weerasinghe, Kuan-Lin Huang, Collin Tokheim, Isidro Cortés-Ciriano, Reyka Jayasinghe, Feng Chen, Lihua Yu, Sam Sun, Catharina Olsen, Jaegil Kim, Alison M Taylor, Andrew D Cherniack, Rehan Akbani, Chayaporn Suphavilai, Niranjan Nagarajan, Joshua M Stuart, Gordon B Mills, Matthew A Wyczalkowski, Benjamin G Vincent, Carolyn M Hutter, Jean Claude Zenklusen, Katherine A Hoadley, Michael C Wendl, Llya Shmulevich, Alexander J Lazar, David A Wheeler, Gad Getz, and Cancer Genome Atlas Research Network. Perspective on oncogenic processes at the end of the beginning of cancer genomics. *Cell*, 173(2):305–320.e10, April 2018.

[122] Shay Zakov, Marcus Kinsella, and Vineet Bafna. An algorithmic approach for breakage-fusion-bridge detection in tumor genomes. *Proc. Natl. Acad. Sci. U. S. A.*, 110:5546–5551, 2013.

[123] B McClintock. The behavior in successive nuclear divisions of a chromosome broken at meiosis. *Proc. Natl. Acad. Sci. U. S. A.*, 25(8):405–416, August 1939.

[124] Travis I Zack, Steven E Schumacher, Scott L Carter, Andrew D Cherniack, Gordon Saksena, Barbara Tabak, Michael S Lawrence, Cheng-Zhong Zhang, Jeremiah Wala, Craig H Mermel, Carrie Sougnez, Stacey B Gabriel, Bryan Hernandez, Hui Shen, Peter W Laird, Gad Getz, Matthew Meyer-

son, and Rameen Beroukhim. Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, 45:1134–1140, 2013.

[125] Julie D R Reimann and Christopher D M Fletcher. Chapter 37 - Soft-Tissue sarcomas. In John Mendelsohn, Peter M Howley, Mark A Israel, Joe W Gray, and Craig B Thompson, editors, *The Molecular Basis of Cancer (Third Edition)*, pages 471–477. W.B. Saunders, Philadelphia, January 2008.

[126] Alexander N Shoushtari, Rodrigo R Munhoz, Deborah Kuk, Patrick A Ott, Douglas B Johnson, Katy K Tsai, Suthee Rapisuwon, Zeynep Eroglu, Ryan J Sullivan, Jason J Luke, Tara C Gangadhar, April K S Salama, Varina Clark, Clare Burias, Igor Puzanov, Michael B Atkins, Alain P Algazi, Antoni Ribas, Jedd D Wolchok, and Michael A Postow. The efficacy of anti-PD-1 agents in acral and mucosal melanoma. *Cancer*, 122(21):3354–3362, November 2016.

[127] Irina V Kovtun, Stephen J Murphy, Sarah H Johnson, John C Cheville, and George Vasmatzis. Chromosomal catastrophe is a frequent event in clinically insignificant prostate cancer. *Oncotarget*, 6(30):29087–29096, October 2015.

[128] John M Furgason, Robert F Koncar, Sharon K Michelhaugh, Fazlul H Sarkar, Sandeep Mittal, Andrew E Sloan, Jill S Barnholtz-Sloan, and El Mustapha Bahassi. Whole genome sequence analysis links chromothripsis to EGFR, MDM2, MDM4, and CDK4 amplification in glioblastoma. *Oncoscience*, 2(7):618–628, July 2015.

[129] Joshua T Lange, Celine Y Chen, Yuriy Pichugin, Frank Xie, Jun Tang, King L Hung, Kathryn E Yost, Quanming Shi, Marcella L Erb, Utkrisht Rajkumar, Sihan Wu, Charles Swanton, Zhe Liu, Weini Huang, Howard Y Chang, Vineet Bafna, Anton G Henssen, Benjamin Werner, and Paul S Mischel. Principles of ecDNA random inheritance drive rapid genome change and therapy resistance in human cancers. *bioRxiv*, page 2021.06.11.447968, June 2021.

[130] Craig M Bielski and Barry S Taylor. Homing in on genomic instability as a therapeutic target in cancer. *Nat. Commun.*, 12(1):3663, June 2021.

[131] T de Lange, L Shiue, R M Myers, D R Cox, S L Naylor, A M Killery, and H E Varmus. Structure and variability of human chromosome ends. *Mol. Cell. Biol.*, 10(2):518–527, February 1990.

[132] Titia de Lange. Shelterin: the protein complex that shapes and safeguards human telomeres. *Genes Dev.*, 19(18):2100–2110, September 2005.

[133] S E Artandi, S Chang, S L Lee, S Alson, G J Gottlieb, L Chin, and R A DePinho. Telomere dysfunction promotes non-reciprocal translocations and epithelial cancers in mice. *Nature*, 406(6796):641–645, August 2000.

[134] John Maciejowski and Titia de Lange. Telomeres in cancer: tumour suppression and genome instability. *Nat. Rev. Mol. Cell Biol.*, 18(3):175–186, March 2017.

[135] Jerry W Shay and Woodring E Wright. Senescence and immortalization: role of telomeres and telomerase. *Carcinogenesis*, 26(5):867–874, May 2005.

[136] D Gisselsson, T Jonson, A Petersén, B Strömbeck, P Dal Cin, M Höglund, F Mitelman, F Mertens, and N Mandahl. Telomere dysfunction triggers extensive DNA fragmentation and evolution of complex chromosome abnormalities in human malignant tumors. *Proc. Natl. Acad. Sci. U. S. A.*, 98(22):12683–12688, October 2001.

[137] Heng Li. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30(20):2843–2851, October 2014.

[138] Rónán C O'Hagan, Sandy Chang, Richard S Maser, Ramya Mohan, Steven E Artandi, Lynda Chin, and Ronald A DePinho. Telomere dysfunction provokes regional amplification and deletion in cancer genomes. *Cancer Cell*, 2(2):149–155, August 2002.

[139] Zhihu Ding, Chang-Jiun Wu, Mariela Jaskelioff, Elena Ivanova, Maria Kost-Alimova, Alexei Protopopov, Gerald C Chu, Guocan Wang, Xin Lu, Emma S Labrot, Jian Hu, Wei Wang, Yonghong Xiao, Hailei Zhang, Jianhua Zhang, Jingfang Zhang, Boyi Gan, Samuel R Perry, Shan Jiang, Liren Li, James W Horner, Y Alan Wang, Lynda Chin, and Ronald A DePinho. Telomerase reactivation following telomere dysfunction yields murine prostate tumors with bone metastases. *Cell*, 148(5):896–907, March 2012.

[140] Kate Liddiard, Brian Ruis, Taylor Takasugi, Adam Harvey, Kevin E Ashelford, Eric A Hendrickson, and Duncan M Baird. Sister chromatid telomere fusions, but not NHEJ-mediated inter-chromosomal telomere fusions, occur independently of DNA ligases 3 and 4. *Genome Res.*, 26(5):588–600, May 2016.

[141] John Maciejowski, Yilong Li, Nazario Bosco, Peter J. Campbell, and Titia de Lange. Chromothripsis and Kataegis Induced by Telomere Crisis. *Cell*, 163, 2015.

[142] John Maciejowski, Aikaterini Chatzipli, Alexandra Dananberg, Kevan Chu, Eleonore Toufektchan, Leszek J Klimczak, Dmitry A Gordenin, Peter J Campbell, and Titia de Lange. APOBEC3-dependent kataegis and TREX1-driven chromothripsis during telomere crisis. *Nat. Genet.*, July 2020.

[143] Kez Cleal, Rhiannon E Jones, Julia W Grimstead, Eric A Hendrickson, and Duncan M Baird. Chromothripsis during telomere crisis is independent of NHEJ, and consistent with a replicative origin. *Genome Res.*, 29(5):737–749, May 2019.

[144] C M Counter, A A Avilion, C E LeFeuvre, N G Stewart, C W Greider, C B Harley, and S Bacchetti. Telomere shortening associated with chromosome instability is arrested in immortal cells which express telomerase activity. *EMBO J.*, 11(5):1921–1929, May 1992.

[145] J W Shay and W E Wright. Quantitation of the frequency of immortalization of normal human diploid fibroblasts by SV40 large t-antigen. *Exp. Cell Res.*, 184(1):109–118, September 1989.

[146] T M Bryan, A Englezou, J Gupta, S Bacchetti, and R R Reddel. Telomere elongation in immortal human cells without detectable telomerase activity. *EMBO J.*, 14(17):4240–4248, September 1995.

[147] Daniel L Cameron, Jonathan Baber, Charles Shale, Jose Espejo Valle-Inclan, Nicolle Besselink, Arne van Hoeck, Roel Janssen, Edwin Cuppen, Peter Priestley, and Anthony T Papenfuss. GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. *Genome Biol.*, 22(1):202, July 2021.

[148] Kunitoshi Chiba, Franziska K Lorbeer, A Hunter Shain, David T McSwiggen, Eva Schruf, Areum Oh, Jekwan Ryu, Xavier Darzacq, Boris C Bastian, and Dirk Hockemeyer. Mutations in the promoter of the telomerase gene TERT contribute to tumorigenesis by a two-step mechanism. *Science*, August 2017.

[149] Duncan M Baird, Jan Rowson, David Wynford-Thomas, and David Kipling. Extensive allelic variation and ultrashort telomeres in senescent human cells. *Nat. Genet.*, 33(2):203–207, February 2003.

[150] Jeremiah Wala and Rameen Beroukhim. SeqLib: a C ++ API for rapid BAM manipulation, sequence alignment and sequence assembly. *Bioinformatics*, 33(5):751–753, March 2017.

[151] Mia Petljak, Ludmil B Alexandrov, Jonathan S Brammeld, Stacey Price, David C Wedge, Sebastian Grossmann, Kevin J Dawson, Young Seok Ju, Francesco Iorio, Jose M C Tubio, Ching Chiek Koh, Ilias Georgakopoulos-Soares, Bernardo Rodríguez-Martín, Burçak Otlu, Sarah O'Meara, Adam P Butler, Andrew Menzies, Shriram G Bhosle, Keiran Raine, David R Jones, Jon W Teague, Kathryn Beal, Calli Latimer, Laura O'Neill, Jorge Zamora, Elizabeth Anderson, Nikita Patel, Mark Maddison, Bee Ling Ng, Jennifer Graham, Mathew J Garnett, Ultan McDermott, Serena Nik-Zainal, Peter J Campbell, and Michael R Stratton. Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell*, 176(6):1282–1294.e20, March 2019.

[152] Jian Carrot-Zhang, Nyasha Chambwe, Jeffrey S Damrauer, Theo A Knijnenburg, A Gordon Robertson, Christina Yau, Wanding Zhou, Ashton C Berger, Kuan-Lin Huang, Justin Y Newberg, R Jay Mashl, Alessandro Romanel, Rosalyn W Sayaman, Francesca Demichelis, Ina Felau, Garrett M Frampton, Seunghun Han, Katherine A Hoadley, Anab Kemal, Peter W Laird, Alexander J Lazar, Xiuning Le, Ninad Oak, Hui Shen, Christopher K Wong, Jean C Zenklusen, Elad Ziv, Cancer Genome Atlas Analysis Network, Andrew D Cherniack, and Rameen Beroukhim. Comprehensive analysis of genetic ancestry and its molecular correlates in cancer. *Cancer Cell*, 37(5):639–654.e6, May 2020.

[153] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2019. *CA Cancer J. Clin.*, 69(1):7–34, January 2019.

[154] Tushar J Desai, Douglas G Brownfield, and Mark A Krasnow. Alveolar progenitor and stem cells in lung development, renewal and cancer. *Nature*, 507(7491):190–194, March 2014.

[155] Joshua D Campbell, Anton Alexandrov, Jaegil Kim, Jeremiah Wala, Alice H Berger, Chandra Sekhar Pedamallu, Sachet A Shukla, Guangwu Guo, Angela N Brooks, Bradley A Murray, Marcin Imielinski, Xin Hu, Shiyun Ling, Rehan Akbani, Mara Rosenberg, Carrie Cibulskis, Aruna Ramachandran, Eric A Collisson, David J Kwiatkowski, Michael S Lawrence, John N Weinstein, Roel G W Verhaak, Catherine J Wu, Peter S Hammerman, Andrew D Cherniack, Gad Getz, Cancer Genome Atlas Research Network, Maxim N Artyomov, Robert Schreiber, Ramaswamy Govindan, and Matthew Meyerson. Distinct patterns of somatic genome alterations

in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.*, 48(6):607–616, June 2016.

[156] Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–550, July 2014.

[157] Lei Zhong, Yueshan Li, Liang Xiong, Wenjing Wang, Ming Wu, Ting Yuan, Wei Yang, Chenyu Tian, Zhuang Miao, Tianqi Wang, and Shengyong Yang. Small molecules in targeted cancer therapy: advances, challenges, and future perspectives. *Signal Transduct. Target. Ther.*, 6(1):201, May 2021.

[158] Jude Canon, Karen Rex, Anne Y Saiki, Christopher Mohr, Keegan Cooke, Dhanashri Bagal, Kevin Gaida, Tyler Holt, Charles G Knutson, Neelima Koppada, Brian A Lanman, Jonathan Werner, Aaron S Rapaport, Tisha San Miguel, Roberto Ortiz, Tao Osgood, Ji-Rong Sun, Xiaochun Zhu, John D McCarter, Laurie P Volak, Brett E Houk, Marwan G Fakih, Bert H O'Neil, Timothy J Price, Gerald S Falchook, Jayesh Desai, James Kuo, Ramaswamy Govindan, David S Hong, Wenjun Ouyang, Haby Henary, Tara Arvedson, Victor J Cee, and J Russell Lipford. The clinical KRAS(G12C) inhibitor AMG 510 drives anti-tumour immunity, 2019.

[159] Jonathan M Ostrem, Ulf Peters, Martin L Sos, James A Wells, and Kevan M Shokat. K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. *Nature*, 503(7477):548–551, November 2013.

[160] Michael J Clark, Rui Chen, Hugo Y K Lam, Konrad J Karczewski, Rong Chen, Ghia Euskirchen, Atul J Butte, and Michael Snyder. Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.*, 29(10):908–914, September 2011.

[161] Lynnette Fernandez-Cuesta, Dennis Plenker, Hirotaka Osada, Ruping Sun, Roopika Menon, Frauke Leenders, Sandra Ortiz-Cuaran, Martin Peifer, Marc Bos, Juliane Daßler, Florian Malchers, Jakob Schöttle, Wenzel Vogel, Ilona Dahmen, Mirjam Koker, Roland T Ullrich, Gavin M Wright, Prudence A Russell, Zoe Wainer, Benjamin Solomon, Elisabeth Brambilla, Hélène Nagy-Mignotte, Denis Moro-Sibilot, Christian G Brambilla, Sylvie Lantuejoul, Janine Altmüller, Christian Becker, Peter Nürnberg, Johannes M Heuckmann, Erich Stoelben, Iver Petersen, Joachim H Clement, Jörg Sänger, Lucia A Muscarella, Annamaria la Torre, Vito M Fazio, Idoya Lahortiga, Timothy Perera, Souichi Ogata, Marc Parade, Dirk Brehmer, Martin Vingron, Lukas C Heukamp, Reinhard Buettner, Thomas Zander, Jürgen Wolf, Sven Perner, Sascha Ansén, Stefan A Haas, Yasushi Yatabe,

and Roman K Thomas. CD74–NRG1 fusions in lung adenocarcinoma. *Cancer Discov.*, 4(4):415–422, April 2014.

[162] Marcin Imielinski, Heidi Greulich, Bethany Kaplan, Luiz Araujo, Joseph Amann, Leora Horn, Joan Schiller, Miguel A Villalona-Calero, Matthew Meyerson, and David P Carbone. Oncogenic and sorafenib-sensitive ARAF mutations in lung adenocarcinoma, 2014.

[163] Diana Cai, Peter S Choi, Maya Gelbard, and Matthew Meyerson. Identification and characterization of oncogenic SOS1 mutations in lung adenocarcinoma. *Mol. Cancer Res.*, 17(4):1002–1012, April 2019.

[164] Marcin Imielinski, Guangwu Guo, and Matthew Meyerson. Insertions and deletions target Lineage-Defining genes in human cancers. *Cell*, 168(3):460–472.e14, January 2017.

[165] Craig H Mermel, Steven E Schumacher, Barbara Hill, Matthew L Meyerson, Rameen Beroukhim, and Gad Getz. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.*, 12(4):R41, April 2011.

[166] Ferdinandos Skoulidis, Michael E Goldberg, Danielle M Greenawalt, Matthew D Hellmann, Mark M Awad, Justin F Gainor, Alexa B Schrock, Ryan J Hartmaier, Sally E Trabucco, Laurie Gay, Siraj M Ali, Julia A Elvin, Gaurav Singal, Jeffrey S Ross, David Fabrizio, Peter M Szabo, Han Chang, Ariella Sasson, Sujaya Srinivasan, Stefan Kirov, Joseph Szustakowski, Patrik Vitazka, Robin Edwards, Jose A Bufill, Neelesh Sharma, Sai-Hong I Ou, Nir Peled, David R Spigel, Hira Rizvi, Elizabeth Jimenez Aguilar, Brett W Carter, Jeremy Erasmus, Darragh F Halpenny, Andrew J Plodkowski, Niamh M Long, Mizuki Nishino, Warren L Denning, Ana Galan-Cobo, Haifa Hamdi, Taghreed Hirz, Pan Tong, Jing Wang, Jaime Rodriguez-Canales, Pamela A Villalobos, Edwin R Parra, Neda Kalhor, Lynette M Sholl, Jennifer L Sauter, Achim A Jungbluth, Mari Mino-Kenudson, Roxana Azimi, Yasir Y Elamin, Jianjun Zhang, Giulia C Leonardi, Fei Jiang, Kwok-Kin Wong, J Jack Lee, Vassiliki A Papadimitrakopoulou, Ignacio I Wistuba, Vincent A Miller, Garrett M Frampton, Jedd D Wolchok, Alice T Shaw, Pasi A Jänne, Philip J Stephens, Charles M Rudin, William J Geese, Lee A Albacker, and John V Heymach. STK11/LKB1 mutations and PD-1 inhibitor resistance in KRAS-Mutant lung adenocarcinoma. *Cancer Discov.*, 8(7):822–835, July 2018.

[167] M Ryan Corces, Jeffrey M Granja, Shadi Shams, Bryan H Louie, Jose A Seoane, Wanding Zhou, Tiago C Silva, Clarice Groeneveld, Christopher K

Wong, Seung Woo Cho, Ansuman T Satpathy, Maxwell R Mumbach, Katherine A Hoadley, A Gordon Robertson, Nathan C Sheffield, Ina Felau, Mauro A A Castro, Benjamin P Berman, Louis M Staudt, Jean C Zenklusen, Peter W Laird, Christina Curtis, Cancer Genome Atlas Analysis Network, William J Greenleaf, and Howard Y Chang. The chromatin accessibility landscape of primary human cancers. *Science*, 362(6413), October 2018.

[168] Yao Fu, Zhu Liu, Shaoke Lou, Jason Bedford, Xinmeng Jasmine Mu, Kevin Y Yip, Ekta Khurana, and Mark Gerstein. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.*, 15(10):480, 2014.

[169] Ekta Khurana, Yao Fu, Dimple Chakravarty, Francesca Demichelis, Mark A Rubin, and Mark Gerstein. Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.*, 17(2):93–108, February 2016.

[170] Matteo Marchesini, Yamini Ogoti, Elena Fiorini, Anil Aktas Samur, Luigi Nezi, Marianna D'Anca, Paola Storti, Mehmet Kemal Samur, Irene Ganan-Gomez, Maria Teresa Fulciniti, Nipun Mistry, Shan Jiang, Naran Bao, Valentina Marchica, Antonino Neri, Carlos Bueso-Ramos, Chang-Jiun Wu, Li Zhang, Han Liang, Xinxin Peng, Nicola Giuliani, Giulio Draetta, Karen Clise-Dwyer, Hagop Kantarjian, Nikhil Munshi, Robert Orlowski, Guillermo Garcia-Manero, Ronald A DePinho, and Simona Colla. ILF2 is a regulator of RNA splicing and DNA damage response in 1q21-amplified multiple myeloma. *Cancer Cell*, 32(1):88–100.e6, July 2017.

[171] Charles Swanton and Ramaswamy Govindan. Clinical implications of genomic discoveries in lung cancer. *N. Engl. J. Med.*, 374(19):1864–1873, May 2016.

[172] Aaron M Goodman, Shumei Kato, Lyudmila Bazhenova, Sandip P Patel, Garrett M Frampton, Vincent Miller, Philip J Stephens, Gregory A Daniels, and Razelle Kurzrock. Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. *Mol. Cancer Ther.*, 16(11):2598–2608, November 2017.

[173] Ahmet Zehir, Ryma Benayed, Ronak H Shah, Aijazuddin Syed, Sumit Middha, Hyunjae R Kim, Preethi Srinivasan, Jianjiong Gao, Debyani Chakravarty, Sean M Devlin, Matthew D Hellmann, David A Barron, Alison M Schram, Meera Hameed, Snjezana Dogan, Dara S Ross, Jaclyn F Hechtman, Deborah F DeLair, Jinjuan Yao, Diana L Mandelker,

Donavan T Cheng, Raghu Chandramohan, Abhinita S Mohanty, Ryan N Ptashkin, Gowtham Jayakumaran, Meera Prasad, Mustafa H Syed, Anoop Balakrishnan Rema, Zhen Y Liu, Khedoudja Nafa, Laetitia Borsu, Justyna Sadowska, Jacklyn Casanova, Ruben Bacares, Iwona J Kiecka, Anna Razumova, Julie B Son, Lisa Stewart, Tessara Baldi, Kerry A Mullaney, Hikmat Al-Ahmadie, Efsevia Vakiani, Adam A Abeshouse, Alexander V Penson, Philip Jonsson, Niedzica Camacho, Matthew T Chang, Helen H Won, Benjamin E Gross, Ritika Kundra, Zachary J Heins, Hsiao-Wei Chen, Sarah Phillips, Hongxin Zhang, Jiaojiao Wang, Angelica Ochoa, Jonathan Wills, Michael Eubank, Stacy B Thomas, Stuart M Gardos, Dalicia N Reales, Jesse Galle, Robert Durany, Roy Cambria, Wassim Abida, Andrea Cercek, Darren R Feldman, Mrinal M Gounder, A Ari Hakimi, James J Harding, Gopa Iyer, Yelena Y Janjigian, Emmet J Jordan, Ciara M Kelly, Maeve A Lowery, Luc G T Morris, Antonio M Omuro, Nitya Raj, Pedram Razavi, Alexander N Shoushtari, Neerav Shukla, Tara E Soumerai, Anna M Varghese, Rona Yaeger, Jonathan Coleman, Bernard Bochner, Gregory J Riely, Leonard B Saltz, Howard I Scher, Paul J Sabbatini, Mark E Robson, David S Klimstra, Barry S Taylor, Jose Baselga, Nikolaus Schultz, David M Hyman, Maria E Arcila, David B Solit, Marc Ladanyi, and Michael F Berger. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.*, 23(6):703–713, June 2017.

[174] Kathryn C Arbour, Emmett Jordan, Hyunjae Ryan Kim, Jordan Dienstag, Helena A Yu, Francisco Sanchez-Vega, Piro Lito, Michael Berger, David B Solit, Matthew Hellmann, Mark G Kris, Charles M Rudin, Ai Ni, Maria Arcila, Marc Ladanyi, and Gregory J Riely. Effects of co-occurring genomic alterations on outcomes in patients with KRAS-Mutant Non-Small cell lung cancer. *Clin. Cancer Res.*, 24(2):334–340, January 2018.

[175] R Sears, F Nuckolls, E Haura, Y Taya, K Tamai, and J R Nevins. Multiple ras-dependent phosphorylation pathways regulate myc protein stability. *Genes Dev.*, 14(19):2501–2514, October 2000.

[176] Haiquan Chen, Jian Carrot-Zhang, Yue Zhao, Haichuan Hu, Samuel S Freeman, Su Yu, Gavin Ha, Alison M Taylor, Ashton C Berger, Lindsay Westlake, Yuanting Zheng, Jiyang Zhang, Aruna Ramachandran, Qiang Zheng, Yunjian Pan, Difan Zheng, Shanbo Zheng, Chao Cheng, Muyu Kuang, Xiaoyan Zhou, Yang Zhang, Hang Li, Ting Ye, Yuan Ma, Zhendong Gao, Xiaoting Tao, Han Han, Jun Shang, Ying Yu, Ding Bao, Yechao Huang, Xiangnan Li, Yawei Zhang, Jiaqing Xiang, Yihua Sun, Yuan Li, Andrew D Cherniack, Joshua D Campbell, Leming Shi, and Matthew

Meyerson. Genomic and immune profiling of pre-invasive lung adenocarcinoma. *Nat. Commun.*, 10(1):5472, November 2019.

[177] Matthias Drosten, Eleanor Y M Sum, Carmen G Lechuga, Lucía Simón-Carrasco, Harrys K C Jacob, Raquel García-Medina, Sidong Huang, Roderick L Beijersbergen, Rene Bernards, and Mariano Barbacid. Loss of p53 induces cell proliferation via ras-independent activation of the Raf/Mek/Erk signaling pathway. *Proc. Natl. Acad. Sci. U. S. A.*, 111(42):15155–15160, October 2014.

[178] Lorena Salgueiro, Christopher Buccitelli, Konstantina Rowald, Kalman Somogyi, Sridhar Kandala, Jan O Korbel, and Rocio Sotillo. Acquisition of chromosome instability is a mechanism to evade oncogene addiction, 2020.

[179] Carlos Oliver, Dexiong Chen, Vincent Mallet, Pericles Philippopoulos, and Karsten Borgwardt. Approximate network motif mining via graph learning. June 2022.

[180] Shuo Yu, Yufan Feng, Da Zhang, Hayat Dino Bedru, Bo Xu, and Feng Xia. Motif discovery in networks: A survey. *Computer Science Review*, 37:100267, August 2020.

[181] D Conte, P Foggia, C Sansone, and M Vento. THIRTY YEARS OF GRAPH MATCHING IN PATTERN RECOGNITION. *Int. J. Pattern Recognit Artif Intell.*, 18(03):265–298, May 2004.

[182] Niusvel Acosta-Mendoza, Andrés Gago-Alonso, and José E Medina-Pagola. Frequent approximate subgraphs as features for graph-based image classification. *Knowledge-Based Systems*, 27:381–392, March 2012.

[183] Shijie Zhang and Jiong Yang. RAM: Randomized approximate graph mining. In *Scientific and Statistical Database Management*, pages 187–203. Springer Berlin Heidelberg, 2008.

[184] Andrea Degasperi, Xueqing Zou, Tauanne Dias Amarante, Andrea Martinez-Martinez, Gene Ching Chiek Koh, João M L Dias, Laura Heskin, Lucia Chmelova, Giuseppe Rinaldi, Valerie Ya Wen Wang, Arjun S Nanda, Aaron Bernstein, Sophie E Momen, Jamie Young, Daniel Perez-Gil, Yasin Memari, Cherif Badja, Scott Shooter, Jan Czarnecki, Matthew A Brown, Helen R Davies, Null Null, Serena Nik-Zainal, J C Ambrose, P Arumugam, R Bevers, M Bleda, F Boardman-Pretty, C R Boustred, H Brit-

tain, M J Caulfield, G C Chan, T Fowler, A Giess, A Hamblin, S Henderson, T J P Hubbard, R Jackson, L J Jones, D Kasperaviciute, M Kayikci, A Kousathanas, L Lahnstein, S E A Leigh, I U S Leong, F J Lopez, F Maleady-Crowe, M McEntagart, F Minneci, L Moutsianas, M Mueller, N Murugaesu, A C Need, P O'Donovan, C A Odhams, C Patch, D Perez-Gil, M B Pereira, J Pullinger, T Rahim, A Rendon, T Rogers, K Savage, K Sawant, R H Scott, A Siddiq, A Sieghart, S C Smith, A Sosinsky, A Stuckey, M Tanguy, A L Taylor Tavares, E R A Thomas, S R Thompson, A Tucci, M J Welland, E Williams, K Witkowska, and S M Wood. Substitution mutational signatures in whole-genome–sequenced cancers in the UK population. *Science*, 376(6591):abl9283, 2022.

[185] Peter Priestley, Jonathan Baber, Martijn P Lolkema, Neeltje Steeghs, Ewart de Bruijn, Charles Shale, Korneel Duyvesteyn, Susan Haidari, Arne van Hoeck, Wendy Onstenk, Paul Roepman, Mircea Voda, Haiko J Bloemendal, Vivianne C G Tjan-Heijnen, Carla M L van Herpen, Mariette Labots, Petronella O Witteveen, Egbert F Smit, Stefan Sleijfer, Emile E Voest, and Edwin Cuppen. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature*, 575(7781):210–216, November 2019.

[186] Srinivas R Viswanathan, Gavin Ha, Andreas M Hoff, Jeremiah A Wala, Jian Carrot-Zhang, Christopher W Whelan, Nicholas J Haradhvala, Samuel S Freeman, Sarah C Reed, Justin Rhoades, Paz Polak, Michelle Cipicchio, Stephanie A Wankowicz, Alicia Wong, Tushar Kamath, Zhenwei Zhang, Gregory J Gydush, Denisse Rotem, J Christopher Love, Gad Getz, Stacey Gabriel, heng Zhong, Scott M Dehm, Peter S Nelson, Eliezer M Van Allen, Atish D Choudhury, Viktor A Adalsteinsson, Rameen Beroukhim, Mary-Ellen Taplin, and Matthew Meyerson. Structural Alterations Driving Castration-Resistant Prostate Cancer Revealed by Linked-Read Genome Sequencing. *Cell*, 0, 2018.

[187] Lisanne F van Dessel, Job van Riet, Minke Smits, Yanyun Zhu, Paul Hamberg, Michiel S van der Heijden, Andries M Bergman, Inge M van Oort, Ronald de Wit, Emile E Voest, Neeltje Steeghs, Takafumi N Yamaguchi, Julie Livingstone, Paul C Boutros, John W M Martens, Stefan Sleijfer, Edwin Cuppen, Wilbert Zwart, Harmen J G van de Werken, Niven Mehra, and Martijn P Lolkema. The genomic landscape of metastatic castration-resistant prostate cancers reveals multiple distinct genotypes with potential clinical impact. *Nat. Commun.*, 10(1):5251, November 2019.

[188] Susanne N Gröbner, Barbara C Worst, Joachim Weischenfeldt, Ivo Buchhalter, Kortine Kleinheinz, Vasilisa A Rudneva, Pascal D Johann, Gnana Prakash Balasubramanian, Maia Segura-Wang, Sebastian Brabetz,

Sebastian Bender, Barbara Hutter, Dominik Sturm, Elke Pfaff, Daniel Hübschmann, Gideon Zipprich, Michael Heinold, Jürgen Eils, Christian Lawerenz, Serap Erkek, Sander Lambo, Sebastian Waszak, Claudia Blattmann, Arndt Borkhardt, Michaela Kuhlen, Angelika Eggert, Simone Fulda, Manfred Gessler, Jenny Wegert, Roland Kappler, Daniel Baumhoer, Stefan Burdach, Renate Kirschner-Schwabe, Udo Kontny, Andreas E Kulozik, Dietmar Lohmann, Simone Hettmer, Cornelia Eckert, Stefan Bielack, Michaela Nathrath, Charlotte Niemeyer, Günther H Richter, Johannes Schulte, Reiner Siebert, Frank Westermann, Jan J Molenaar, Gilles Vassal, Hendrik Witt, ICGC PedBrain-Seq Project, ICGC MMML-Seq Project, Birgit Burkhardt, Christian P Kratz, Olaf Witt, Cornelis M van Tilburg, Christof M Kramm, Gudrun Fleischhack, Uta Dirksen, Stefan Rutkowski, Michael Frühwald, Katja von Hoff, Stephan Wolf, Thomas Klingebiel, Ewa Koscielniak, Pablo Landgraf, Jan Koster, Adam C Resnick, Jinghui Zhang, Yanling Liu, Xin Zhou, Angela J Waanders, Danny A Zwijnenburg, Pichai Raman, Benedikt Brors, Ursula D Weber, Paul A Northcott, Kristian W Pajtler, Marcel Kool, Rosario M Piro, Jan O Korbel, Matthias Schlesner, Roland Eils, David T W Jones, Peter Lichter, Lukas Chavez, Marc Zapatka, and Stefan M Pfister. The landscape of genomic alterations across childhood cancers. *Nature*, 555(7696):321–327, March 2018.

[189] Paul A Northcott, Ivo Buchhalter, A Sorana Morrissy, Volker Hovestadt, Joachim Weischenfeldt, Tobias Ehrenberger, Susanne Gröbner, Maia Segura-Wang, Thomas Zichner, Vasilisa A Rudneva, Hans-Jörg Warnatz, Nikos Sidiropoulos, Aaron H Phillips, Steven Schumacher, Kortine Kleinheinz, Sebastian M Waszak, Serap Erkek, David T W Jones, Barbara C Worst, Marcel Kool, Marc Zapatka, Natalie Jäger, Lukas Chavez, Barbara Hutter, Matthias Bieg, Nagarajan Paramasivam, Michael Heinold, Zuguang Gu, Naveed Ishaque, Christina Jäger-Schmidt, Charles D Imbusch, Alke Jugold, Daniel Hübschmann, Thomas Risch, Vyacheslav Amstislavskiy, Francisco German Rodriguez Gonzalez, Ursula D Weber, Stephan Wolf, Giles W Robinson, Xin Zhou, Gang Wu, David Finkelstein, Yanling Liu, Florence M G Cavalli, Betty Luu, Vijay Ramaswamy, Xiaochong Wu, Jan Koster, Marina Ryzhova, Yoon-Jae Cho, Scott L Pomeroy, Christel Herold-Mende, Martin Schuhmann, Martin Ebinger, Linda M Liau, Jaume Mora, Roger E McLendon, Nada Jabado, Toshihiro Kumabe, Eric Chuah, Yussanne Ma, Richard A Moore, Andrew J Mungall, Karen L Mungall, Nina Thiessen, Kane Tse, Tina Wong, Steven J M Jones, Olaf Witt, Till Milde, Andreas Von Deimling, David Capper, Andrey Korshunov, Marie-Laure Yaspo, Richard Kriwacki, Amar Gajjar, Jinghui Zhang, Rameen Beroukhim, Ernest Fraenkel, Jan O Korbel, Benedikt Brors, Matthias Schlesner, Roland Eils, Marco A Marra, Stefan M Pfis-

ter, Michael D Taylor, and Peter Lichter. The whole-genome landscape of medulloblastoma subtypes. *Nature*, 547(7663):311–317, July 2017.

[190] Ryan L Collins, Harrison Brand, Claire E Redin, Carrie Hanscom, Caroline Antolik, Matthew R Stone, Joseph T Glessner, Tamara Mason, Giulia Pregno, Naghmeh Dorrani, Giorgia Mandrile, Daniela Giachino, Danielle Perrin, Cole Walsh, Michelle Cipicchio, Maura Costello, Alexei Stortchevoi, Joon-Yong An, Benjamin B Currall, Catarina M Seabra, Ashok Ragavendran, Lauren Margolin, Julian A Martinez-Agosto, Diane Lucente, Brynn Levy, Stephan J Sanders, Ronald J Wapner, Fabiola Quintero-Rivera, Wigard Kloosterman, and Michael E Talkowski. Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome Biol.*, 18(1):36, March 2017.

[191] Inigo Martincorena, Joanna C Fowler, Agnieszka Wabik, Andrew R J Lawson, Federico Abascal, Michael W J Hall, Alex Cagan, Kasumi Murai, Krishnaa Mahbubani, Michael R Stratton, Rebecca C Fitzgerald, Penny A Handford, Peter J Campbell, Kourosh Saeb-Parsy, and Philip H Jones. Somatic mutant clones colonize the human esophagus with age. *Science*, 57:eaau3879, 2018.

[192] Shobana Sekar, Livia Tomasini, Christos Proukakis, Taejeong Bae, Logan Manlove, Yeongjun Jang, Soraya Scuderi, Bo Zhou, Maria Kalyva, Anahita Amiri, Jessica Mariani, Fritz J Sedlazeck, Alexander E Urban, Flora M Vaccarino, and Alexej Abyzov. Complex mosaic structural variations in human fetal brains. *Genome Res.*, 30(12):1695–1704, December 2020.

[193] Xuefang Zhao, Ryan L Collins, Wan-Ping Lee, Alexandra M Weber, Yukyung Jun, Qihui Zhu, Ben Weisburd, Yongqing Huang, Peter A Audano, Harold Wang, Mark Walker, Chelsea Lowther, Jack Fu, Human Genome Structural Variation Consortium, Mark B Gerstein, Scott E Devine, Tobias Marschall, Jan O Korbel, Evan E Eichler, Mark J P Chaisson, Charles Lee, Ryan E Mills, Harrison Brand, and Michael E Talkowski. Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *Am. J. Hum. Genet.*, March 2021.

[194] Jens Luebeck, Ceyda Coruh, Siavash R Dehkordi, Joshua T Lange, Kristen M Turner, Viraj Deshpande, Dave A Pai, Chao Zhang, Utkrisht Rajkumar, Julie A Law, Paul S Mischel, and Vineet Bafna. AmpliconReconstruc-

tor integrates NGS and optical mapping to resolve the complex structures of focal amplifications. *Nat. Commun.*, 11(1):1–14, September 2020.

[195] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V Bzikadze, Alla Mikheenko, Mitchell R Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J Hoyt, Mark Diekhans, Glennis A Logsdon, Michael Alonge, Stylianos E Antonarakis, Matthew Borchers, Gerard G Bouffard, Shelise Y Brooks, Gina V Caldas, Haoyu Cheng, Chen-Shan Chin, William Chow, Leonardo G de Lima, Philip C Dishuck, Richard Durbin, Tatiana Dvorkina, Ian T Fiddes, Giulio Formenti, Robert S Fulton, Arkarachai Fungtammasan, Erik Garrison, Patrick G S Grady, Tina A Graves-Lindsay, Ira M Hall, Nancy F Hansen, Gabrielle A Hartley, Marina Haukness, Kerstin Howe, Michael W Hunkapiller, Chirag Jain, Miten Jain, Erich D Jarvis, Peter Kerpedjiev, Melanie Kirsche, Mikhail Kolmogorov, Jonas Korlach, Milinn Kremitzki, Heng Li, Valerie V Maduro, Tobias Marschall, Ann M McCartney, Jennifer McDaniel, Danny E Miller, James C Mullikin, Eugene W Myers, Nathan D Olson, Benedict Paten, Paul Peluso, Pavel A Pevzner, David Porubsky, Tamara Potapova, Evgeny I Rogaev, Jeffrey A Rosenfeld, Steven L Salzberg, Valerie A Schneider, Fritz J Sedlazeck, Kishwar Shafin, Colin J Shew, Alaina Shumate, Yumi Sims, Arian F A Smit, Daniela C Soto, Ivan Sović, Jessica M Storer, Aaron Streets, Beth A Sullivan, Françoise Thibaud-Nissen, James Torrance, Justin Wagner, Brian P Walenz, Aaron Wenger, Jonathan M D Wood, Chunlin Xiao, Stephanie M Yan, Alice C Young, Samantha Zarate, Urvashi Surti, Rajiv C McCoy, Megan Y Dennis, Ivan A Alexandrov, Jennifer L Gerton, Rachel J O'Neill, Winston Timp, Justin M Zook, Michael C Schatz, Evan E Eichler, Karen H Miga, and Adam M Phillippy. The complete sequence of a human genome. *bioRxiv*, page 2021.05.26.445798, May 2021.

[196] Peter Ebert, Peter A Audano, Qihui Zhu, Bernardo Rodriguez-Martin, David Porubsky, Marc Jan Bonder, Arvis Sulovari, Jana Ebler, Weichen Zhou, Rebecca Serra Mari, Feyza Yilmaz, Xuefang Zhao, Pinghsun Hsieh, Joyce Lee, Sushant Kumar, Jiadong Lin, Tobias Rausch, Yu Chen, Jingwen Ren, Martin Santamarina, Wolfram Höps, Hufsah Ashraf, Nelson T Chuang, Xiaofei Yang, Katherine M Munson, Alexandra P Lewis, Susan Fairley, Luke J Tallon, Wayne E Clarke, Anna O Basile, Marta Byrska-Bishop, André Corvelo, Uday S Evani, Tsung-Yu Lu, Mark J P Chaisson, Junjie Chen, Chong Li, Harrison Brand, Aaron M Wenger, Maryam Ghareghani, William T Harvey, Benjamin Raeder, Patrick Hasenfeld, Allison A Regier, Haley J Abel, Ira M Hall, Paul Flicek, Oliver Stegle, Mark B Gerstein, Jose M C Tubio, Zepeng Mu, Yang I Li, Xinghua Shi, Alex R Hastie, Kai Ye, Zechen Chong, Ashley D Sanders, Michael C Zody,

Michael E Talkowski, Ryan E Mills, Scott E Devine, Charles Lee, Jan O Korbel, Tobias Marschall, and Evan E Eichler. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, February 2021.

[197] Haoyu Cheng, Gregory T Concepcion, Xiaowen Feng, Haowen Zhang, and Heng Li. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods*, February 2021.

[198] Timour Baslan, Sam Kovaka, Fritz J Sedlazeck, Yanming Zhang, Robert Wappel, Scott W Lowe, Sara Goodwin, and Michael C Schatz. High resolution copy number inference in cancer using short-molecule nanopore sequencing. *bioRxiv*, page 2020.12.28.424602, December 2020.

[199] Rishvanth K Prabakar, Liya Xu, James Hicks, and Andrew D Smith. SMURF-seq: efficient copy number profiling on long-read sequencers. *Genome Biol.*, 20(1):134, July 2019.

[200] Marek Cmero, Ke Yuan, Cheng Soon Ong, Jan Schröder, PCAWG Evolution and Heterogeneity Working Group, Niall M Corcoran, Tony Papenfuss, Christopher M Hovens, Florian Markowetz, Geoff Macintyre, and PCAWG Consortium. Inferring structural variant cancer cell fraction. *Nat. Commun.*, 11(1):730, February 2020.

[201] Simone Zaccaria and Benjamin J Raphael. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. *Nat. Biotechnol.*, September 2020.

[202] C D Greenman, S L Cooke, J Marshall, M R Stratton, and P J Campbell. Modeling the evolution space of breakage fusion bridge cycles with a stochastic folding process. *J. Math. Biol.*, 72(1-2):47–86, January 2016.

[203] Manja Meggendorfer, Vaidehi Jobanputra, Kazimierz O Wrzeszczynski, Paul Roepman, Ewart de Bruijn, Edwin Cuppen, Reinhard Buttner, Carlos Caldas, Sean Grimmond, Charles G Mullighan, Olivier Elemento, Richard Rosenquist, Anna Schuh, and Torsten Haferlach. Analytical demands to use whole-genome sequencing in precision oncology. *Semin. Cancer Biol.*, June 2021.

[204] Asaf Zviran, Rafael C Schulman, Minita Shah, Steven T K Hill, Sunil Deochand, Cole C Khamnei, Dillon Maloney, Kristofer Patel, Will Liao, Adam J Widman, Phillip Wong, Margaret K Callahan, Gavin Ha, Sarah

Reed, Denisse Rotem, Dennie Frederick, Tatyana Sharova, Benchun Miao, Tommy Kim, Greg Gydush, Justin Rhoades, Kevin Y Huang, Nathaniel D Omans, Patrick O Bolan, Andrew H Lipsky, Chelston Ang, Murtaza Malbari, Catherine F Spinelli, Selena Kazancioglu, Alexi M Runnels, Samantha Fennessey, Christian Stolte, Federico Gaiti, Giorgio G Inghirami, Viktor Adalsteinsson, Brian Houck-Loomis, Jennifer Ishii, Jedd D Wolchok, Genevieve Boland, Nicolas Robine, Nasser K Altorki, and Dan A Landau. Genome-wide cell-free DNA mutational integration enables ultrasensitive cancer monitoring. *Nat. Med.*, June 2020.

[205] Adam J Widman, Minita Shah, Nadia Øgaard, Cole C Khamnei, Amanda Frydendahl, Aditya Deshpande, Anushri Arora, Mingxuan Zhang, Daniel Halmos, Jake Bass, Theophile Langanay, Srinivas Rajagopalan, Zoe Steinsnyder, Will Liao, Mads Heilskov Rasmussen, Sarah Østrup Jensen, Jesper Nors, Christina Therkildsen, Jesus Sotelo, Ryan Brand, Ronak H Shah, Alexandre Pellan Cheng, Colleen Maher, Lavinia Spain, Kate Krause, Dennie T Frederick, Murtaza S Malbari, Melissa Marton, Dina Manaa, Lara Winterkorn, Margaret K Callahan, Genevieve Boland, Jedd D Wolchok, Ashish Saxena, Samra Turajlic, Marcin Imielinski, Michael F Berger, Nasser K Altorki, Michael A Postow, Nicolas Robine, Claus Lindbjerg Andersen, and Dan A Landau. Machine learning guided signal enrichment for ultrasensitive plasma tumor burden monitoring. January 2022.

[206] Cameron Herberts, Matti Annala, Joonatan Sipola, Sarah W S Ng, Xinyi E Chen, Anssi Nurminen, Olga V Korhonen, Aslı D Munzur, Kevin Beja, Elena Schönlau, Cecily Q Bernales, Elie Ritch, Jack V W Bacon, Nathan A Lack, Matti Nykter, Rahul Aggarwal, Eric J Small, Martin E Gleave, David A Quigley, Felix Y Feng, Kim N Chi, Alexander W Wyatt, and SU2C/PCF West Coast Prostate Cancer Dream Team. Deep whole-genome ctDNA chronology of treatment-resistant prostate cancer. *Nature*, pages 1–10, July 2022.

[207] Xiangyu Zhang, Zheng Wang, Wanxiangfu Tang, Xinyu Wang, Rui Liu, Hua Bao, Xin Chen, Yulin Wei, Shuyu Wu, Hairong Bao, Xue Wu, Yang Shao, Jia Fan, and Jian Zhou. Ultrasensitive and affordable assay for early detection of primary liver cancer using plasma cell-free DNA fragmentomics. *Hepatology*, 76(2):317–329, August 2022.

[208] Jonathan C M Wan, Dennis Stephens, Lingqi Luo, James R White, Caitlin M Stewart, Benoît Rousseau, Dana W Y Tsui, and Luis A Diaz. Genome-wide mutational signatures in low-coverage whole genome sequencing of cell-free DNA. *Nat. Commun.*, 13(1):1–12, August 2022.

[209] Weichun Huang, Leping Li, Jason R Myers, and Gabor T Marth. ART: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, February 2012.

[210] Jan O. Korbel and Peter J. Campbell. Criteria for Inference of Chromothripsis in Cancer Genomes. *Cell*, 152:1226–1236, 2013.

[211] F Favero, T Joshi, A M Marquard, N J Birkbak, M Krzystanek, Q Li, Z Szallasi, and A C Eklund. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.*, 26(1):64–70, January 2015.

[212] Gavin Ha, Andrew Roth, Jaswinder Khattra, Julie Ho, Damian Yap, Leah M. Prentice, Nataliya Melnyk, Andrew McPherson, Ali Bashashati, Emma Laks, Justina Biele, Jiarui Ding, Alan Le, Jamie Rosner, Karey Shumansky, Marco A. Marra, C. Blake Gilks, David G. Huntsman, Jessica N. McAlpine, Samuel Aparicio, and Sohrab P. Shah. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Research*, 24:1881–1893, 2014.

[213] Sangtae Kim, Konrad Scheffler, Aaron L. Halpern, Mitchell A. Bekritsky, Eunho Noh, Morten Källberg, Xiaoyu Chen, Yeonbin Kim, Doruk Beyter, Peter Krusche, and Christopher T. Saunders. Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods*, 15:591–594, 2018.

[214] Swapan Mallick, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, Niru Chennagiri, Susanne Nordenfelt, Arti Tandon, Pontus Skoglund, Iosif Lazaridis, Sriram Sankararaman, Qiaomei Fu, Nadin Rohland, Gabriel Renaud, Yaniv Erlich, Thomas Willems, Carla Gallo, Jeffrey P. Spence, Yun S. Song, Giovanni Poletti, Francois Balloux, George van Driem, Peter de Knijff, Irene Gallego Romero, Aashish R. Jha, Doron M. Behar, Claudio M. Bravi, Cristian Capelli, Tor Hervig, Andres Moreno-Estrada, Olga L. Posukh, Elena Balanovska, Oleg Balanovsky, Sena Karachanak-Yankova, Hovhannes Sahakyan, Draga Toncheva, Levon Yepiskoposyan, Chris Tyler-Smith, Yali Xue, M. Syafiq Abdullah, Andres Ruiz-Linares, Cynthia M. Beall, Anna Di Rienzo, Choongwon Jeong, Elena B. Starikovskaya, Ene Metspalu, Jüri Parik, Richard Villems, Brenna M. Henn, Ugur Hodoglugil, Robert Mahley, Antti Sajantila, George Stamatoyannopoulos, Joseph T. S. Wee, Rita Khusainova, Elza Khusnutdinova, Sergey Litvinov, George Ayodo, David Comas, Michael F. Hammer, Toomas Kivisild, William Klitz, Cheryl A. Winkler, Damian Labuda, Michael Bamshad, Lynn B. Jorde,

Sarah A. Tishkoff, W. Scott Watkins, Mait Metspalu, Stanislav Dryomov, Rem Sukernik, Lalji Singh, Kumarasamy Thangaraj, Svante Pääbo, Janet Kelso, Nick Patterson, and David Reich. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538:201, 2016.

[215] Roozbeh Dehghannasiri, Donald E Freeman, Milos Jordanski, Gillian L Hsieh, Ana Damljanovic, Erik Lehnert, and Julia Salzman. Improved detection of gene fusions by applying statistical methods reveals oncogenic RNA cancer drivers. *Proc. Natl. Acad. Sci. U. S. A.*, 116(31):15524–15533, July 2019.

[216] H Li and R Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform, 2009.

[217] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A DePristo. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20(9):1297–1303, September 2010.

[218] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples, 2013.

[219] Christopher T Saunders, Wendy S W Wong, Sajani Swamy, Jennifer Becq, Lisa J Murray, and R Keira Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, 28(14):1811–1817, July 2012.

[220] Andreas Wilm, Pauline Poh Kim Aw, Denis Bertrand, Grace Hui Ting Yeo, Swee Hoe Ong, Chang Hua Wong, Chiea Chuen Khor, Rosemary Petric, Martin Lloyd Hibberd, and Niranjan Nagarajan. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets, 2012.

[221] Kai Ye, Marcel H Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–2871, November 2009.

[222] Giuseppe Narzisi, Jason A O'Rawe, Ivan Iossifov, Han Fang, Yoon-Ha

Lee, Zihua Wang, Yiyang Wu, Gholson J Lyon, Michael Wigler, and Michael C Schatz. Accurate de novo and transmitted indel detection in exome-capture data using microassembly. *Nat. Methods*, 11(10):1033–1036, October 2014.

[223] Pablo Cingolani, Viral M Patel, Melissa Coon, Tung Nguyen, Susan J Land, Douglas M Ruden, and Xiangyi Lu. Using drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift, 2012.

[224] Zbyslaw Sondka, Sally Bamford, Charlotte G Cole, Sari A Ward, Ian Dunham, and Simon A Forbes. The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer*, 18(11):696–705, October 2018.

[225] The 1000 Genomes Project Consortium and The 1000 Genomes Project Consortium. A global reference for human genetic variation, 2015.

[226] Konrad J Karczewski, Ben Weisburd, Brett Thomas, Matthew Solomonson, Douglas M Ruderfer, David Kavanagh, Tymor Hamamsy, Monkol Lek, Kaitlin E Samocha, Beryl B Cummings, Daniel Birnbaum, Mark J Daly, Daniel G MacArthur, and The Exome Aggregation Consortium. The ExAC browser: displaying reference data information from over 60 000 exomes, 2017.

[227] Malachi Griffith, Nicholas C Spies, Kilannin Krysiak, Joshua F McMichael, Adam C Coffman, Arpad M Danos, Benjamin J Ainscough, Cody A Ramirez, Damian T Rieke, Lynzey Kujan, Erica K Barnell, Alex H Wagner, Zachary L Skidmore, Amber Wollam, Connor J Liu, Martin R Jones, Rachel L Bilski, Robert Lesurf, Yan-Yang Feng, Nakul M Shah, Melika Bonakdar, Lee Trani, Matthew Matlock, Avinash Ramu, Katie M Campbell, Gregory C Spies, Aaron P Graubert, Karthik Gangavarapu, James M Eldred, David E Larson, Jason R Walker, Benjamin M Good, Chunlei Wu, Andrew I Su, Rodrigo Dienstmann, Adam A Margolin, David Tamborero, Nuria Lopez-Bigas, Steven J M Jones, Ron Bose, David H Spencer, Lukas D Wartman, Richard K Wilson, Elaine R Mardis, and Obi L Griffith. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat. Genet.*, 49(2):170–174, January 2017.

[228] The Uniprot Consortium and The UniProt Consortium. UniProt: a worldwide hub of protein knowledge, 2019.

[229] Jaegil Kim, Kent W Mouw, Paz Polak, Lior Z Braunstein, Atanas Kamburov, David J Kwiatkowski, Jonathan E Rosenberg, Eliezer M Van Allen, Alan D'Andrea, and Gad Getz. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.*, 48(6):600–606, June 2016.

[230] Jian Carrot-Zhang and Jacek Majewski. LoLoPicker: detecting low allelic-fraction variants from low-quality cancer samples, 2017.

[231] Adam B Olshen, E S Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, October 2004.

[232] Zuguang Gu, Roland Eils, and Matthias Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849, September 2016.

[233] Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, Viren Amin, John W Whitaker, Matthew D Schultz, Lucas D Ward, Abhishek Sarkar, Gerald Quon, Richard S Sandstrom, Matthew L Eaton, Yi-Chieh Wu, Andreas R Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R Alan Harris, Noam Shoresh, Charles B Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K Canfield, R Scott Hansen, Rajinder Kaul, Peter J Sabo, Mukul S Bansal, Annaick Carles, Jesse R Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, Ginell Elliott, Tim R Mercer, Shane J Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C Sallari, Kyle T Siebenthall, Nicholas A Sinnott-Armstrong, Michael Stevens, Robert E Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E Beaudet, Laurie A Boyer, Philip L De Jager, Peggy J Farnham, Susan J Fisher, David Haussler, Steven J M Jones, Wei Li, Marco A Marra, Michael T McManus, Shamil Sunyaev, James A Thomson, Thea D Tlsty, Li-Huei Tsai, Wei Wang, Robert A Waterland, Michael Q Zhang, Lisa H Chadwick, Bradley E Bernstein, Joseph F Costello, Joseph R Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A Stamatoyannopoulos, Ting Wang, and Manolis Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, February 2015.