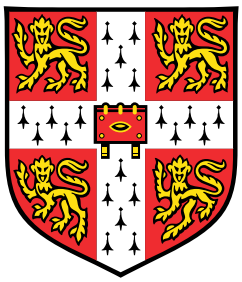
Variation-aware algorithms for cancer genome analysis



Eric T. Dawson

Supervisor: Prof. Richard Durbin
Dr. Stephen Chanock

Department of Genetics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Churchill College September 2019

For my grandfathers, "Pop" and George, and my great uncle Frank.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 60,000 words exclusive of tables, footnotes, bibliography, and appendices.

Eric T. Dawson
September 2019

Variation-aware algorithms for cancer genome analysis

Eric T. Dawson

In this thesis, I explore variation-aware algorithms for analyzing cancer genomes. The scientific community has extensively catalogued millions of mutations present in cancer cells. This information is rarely used during read alignment and variant calling because of a lack of algorithms for doing so. Rediscovering these variants wastes significant computational time and negatively impacts the sensitivity of detection, motivating the development of new solutions.

The variants in malignant cells can arise from a number of genetic and environmental sources. In the years after the Chernobyl nuclear disaster, thousands of individuals in areas where radionuclides were deposited developed thyroid cancer. The excess relative risk of thyroid cancer has been estimated to be between fifteen and thirty fold higher following ^{131}I exposure. In Chapter 2, I analyze more than 300 thyroid cancer cases from children and young adults exposed to ionizing radiation from increased levels of ^{131}I originating from Chernobyl. I characterize the mutational landscape of these tumors across the variant size spectrum and compare it to sporadic thyroid cancer cases. I investigate possible signs of radiation exposure in the genome, especially large balanced structural variants and small indels.

In Chapter 3, I develop methods for working with structural variants in variation graphs. Variation graphs have been shown to reduce reference bias and improve alignment to variant sites, though previous work has primarily focused on variants less than fifty base pairs in size. I describe the tradeoffs of various representations of large variants in the graph. I then develop several methods for genotyping and calling large variants in graphs. I construct graphs of both germline and somatic variation and describe how to work with these structures. I develop a method for fast structural variant genotyping using graph mappings and show that this significantly outperforms standard structural variant callers for certain types of variation. I describe how to locate structural variant mismapping signatures on variation graphs and how variation graphs can improve calling of structural variants.

Lastly in Chapter 4, I demonstrate a new application of an alignment-free algorithm for genome analysis. I describe a MinHash toolkit for viral coinfection analysis and its application to Human Papillomavirus (HPV) samples. This toolkit is able to classify

individual reads from multiple sequencing technologies and accurately detect clinically-relevant HPV coinfections. Finally, I dicuss how these approaches can be applied in other genomic analyses.

Acknowledgements

There are many people who have made this thesis possible, and to whom I am immeasurably grateful.

My parents, Mary, Scott, and Suma, always encouraged me to chase my passions, a privilege I am forever grateful for. Even before you were driving me to bicycle races and airports, you taught me how to assemble computers on the dining room table and make biscuits in the kitchen. My brother Dave, as much a second father as a brother at times, and the rest of our family — Andrea, Bella, and Mariana — have always encouraged me to fly higher even when it meant I couldn't always be home for the holidays.

I owe an immense deal of gratitude to my mentors, Dr. Stephen Chanock and Dr. Richard Durbin. The two of you have encouraged my curiousity and taught me to be diligent and rigorous in my academic pursuits. I had a great deal of fun these four years, which have flown by. You have each pushed me to go farther than I knew I could.

There are many mentors, colleagues, and friends in the Division of Cancer Epidemiology and Genetics who made this thesis possible. Thank you to Dr. Danielle Karyadi, Dr. Lindsay Morton, Tammy Perdikis, Dr. Mia Steinberg, Dr. Tim Myers, Dr. Lea Jessop, Dr. Mitch Machiela, Dr. Leandro Colli, and Dr. Jiyeon Choi for getting me through.

Dr. Gad Getz and Dr. Chip Stewart have been not just collaborators but teachers; thank you both for sharing your knowledge and allowing me to make mistakes.

I could not have started this PhD without the help of my mentors that came before it. Kim Will and Dr. Doug James taught me what mentorship looks like. Dr. Claus Wilke taught me to think scientifically. Dr. Matthew Vaughn taught me to code and trusted a first year biology student with a TOP500 supercomputer. Dr. Lincoln Stein, Dr. Guanming Wu, and Dr. Robin Haw taught me that science crosses borders and that collaboration triumphs over competition. Dr. Dariya Sydykova, another former member of the Wilke Lab, encouraged me from well before we were graduate students. I am grateful I got to share the journey with you.

Jacquelyn Pavilon and Rachel Yen have both been steadfast friends despite many time zones. Thank you both for listening, often over a meal, and for inspiring many adventures outside of the lab.

My many other Cambridge friends made this time so much richer. I am grateful that Sam Katz, Dr. Nick Ader, and Gianmarco Raddi were with me through the OxCam journey. I will cherish many memories of our travels together. Dr. Ananth Kumar was the first person I met in town and has been a great listener and source of plentiful inspiration. Brian Graves has been a constant presence from the beginning of the PhD, and I have always been able to rely on him as a source of both humor and knowledge. Dr. Roey Baror was another great source of humor and the host of many of the best Summer parties. Dr. Sasha Siegel listened patiently to many of my best (and worst) ideas and drove me to think deeply and critically.

Dr. Erik Garrison, Marcus Klarqvist, and Dr. Alexis Braun were with me throughout my time at Sanger and the Department of Genetics and have been some of my closest confidantes. I must also thank Erik for inviting me to join him on the graph genome adventure and for inspiring many scientific ideas over coffee. Markus and Alexis honed many of those ideas; more than that, all three have been steady sources of support. I have crossed paths with each of you in many countries and I hope to do so again. Thank you for making my time at this remarkable place so special.

Sarah Dowd, J.D., has been a steadfast light in the darkest times of this thesis. You are a constant source of guidance, encouragement, and joy. It is no exaggeration that without your support, and your proofreading, this thesis could not have happened.

Thank you all.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Cancer is a disease of the genome, and in comparison to normal human cells the cancer cell population harbors significant amounts of variation and heterogeneity. This variation remains difficult to analyze, in part because of our reliance on the linear reference genome. Structural variants remain particularly difficult to analyze because of their size. While other fields are transitioning to data structures that can encode population variation and the reference together as a pangenome, cancer genomics has largely lagged behind.

In this thesis, I describe new methods for improving our understanding of genomic variation in cancer genomes, especially large-scale genomic alterations, by adapting and extending data structures and algorithms from analogous problems in other fields. Chapter 2 of this thesis is devoted to the analysis of hundreds of papillary thyroid carcinoma samples from the Chernobyl Tissue Bank. It is known that radiation induces double-strand breaks in DNA and that radiation exposed tumors often carry large structural variants and other genomic markers of radiation exposure. I discuss methods for detecting structural variation using variation graphs in Chapter 3. Variation graphs provide a succinct encoding of both a reference genome and variation, which often comes from previous studies. In Chapter 4, I discuss a toolkit for detecting viral coinfections that uses a variation-aware exact kmer matching strategy.

In this chapter I provide background for my work. I cover the human genome and the many mutations found in cancer genomes. I will discuss structural variation and challenges in detecting large variants. I describe how variation-aware algorithms can improve genomic analyses and discuss the use of one type of variation-aware algorithm, the graph genome, in detail. I describe mixtures of genomes. Lastly, I discuss radiation-associated papillary thyroid carcinoma in the context of the Chernobyl nuclear disaster.

1.1 The Cancer Genome

The word cancer comes from the Greek word for crab, highlighting the crustacean-like appearance of a solid tumor invading surrounding tissue using finger-like projections. A state of malignancy is characterized by a cellular population having achieved sufficient evolutionary advantage to invade and outcompete normal tissues. Much like Darwinian evolution in organismal populations, somatic mutations are acquired randomly and then filtered via natural selection to increase the allele frequency of modifications conferring higher fitness [5]. Such changes arise in the genome and alter cellular processes by modifying RNA and protein expression and function. Malignancies may acquire changes that can increase the mutation rate (e.g., by increased genomic instability) or confer a selective advantage, such as increased metabolism. Changes may either be a gain-of-function, in which case the affected gene is termed an "oncogene," or loss-of-function, in which case the affected gene is termed a "tumor suppressor."

Cancer afflicts many eukaryotic organisms, although the remainder of this work will focus on cancer in *Homo sapiens*. Evidence of cancer in humans has been recorded in literature for at least 3000 years [6]. Our understanding of it as a genomic disease has only emerged in roughly the last four decades, though the idea that it originates in the genome was proposed over 100 years ago [7, 8].

Since the advent of short-read sequencing, cancer genomes have become common subjects of sequencing studies. While sequencing has largely remained within the research space, clinical sequencing has become a major priority for health organizations both for hereditary cancer predisposition screening and for targeted precision medicine. Such sequencing may well become routine within the next ten years.

When searching for disease-relevant mutations, short sequencing reads are mapped to the human reference genome. Differences between this reference and the reads are determined through a process known as variant calling. These variants are then assessed to determine whether they may be contributing to disease. This makes understanding the underlying genome crucial to inferences about how the variants within a tumor contribute to the cancerous state.

1.1.1 Reference genomes

Reference genomes fulfill several essential functions that enable contemporary genomic analysis. Most importantly, a reference provides a common coordinate system for genomes from the same species. In the same way, a linear haploid reference provides a sequence

context for variation. This implies a set of states that define variants, such as "reference" and "non-reference." Together, these two roles allow the description of intra- or inter-organismal variation in a concise and consistent manner using a single grammar. Next, a reference genome serves as a representative for its species of origin when comparing genomes of different species. This simplifies processes such as homolog comparison. Lastly, a reference genome provides a basis for further study of organisms within a species. Genome assembly of every individual has historically been prohibitively expensive. A reference genome provides a prior which may be used to scaffold assemblies, align reads, and call variants via resequencing rather than *de novo* assembly.

1.1.2 The Human Genome

The human genome is composed of approximately three billion basepairs organized into 23 pairs of chromosomes. This karyotipic structure is largely conserved among healthy humans. Each individual carries twenty-two pairs of autosomes and one pair of sex chromosomes. By the middle of the 19th century these structures were observable with microscopy, although only the gross structure was discernable. By the middle of the 20th century changes in chromosome structure were implicated in disease, such as the canonical Philadelphia chromosome (a reciprocal translocation t(9;22)(q34;q11)) observed in chronic myelogenous leukemia [9] and the complete or near-complete duplication of chromosome 21 in Down Syndrome [10].

Techniques such as chromosome painting, multicolor flourescence in-situ hybridization (mFISH), and spectral karyotyping (SKY) later provided increased cytogenetic resolution, allowing the detection of events smaller than chromosome arms. Flourescence *in situ* hybridization followed from early experiments which tagged DNA with radioactive labels. The technique allowed tagging of specific sequences within a chromosome with a flourescent tag visible by microscopy. mFISH expanded this technique by using multiple flourescent color tags. The use of multiple tags enables the detection of intrachromosomal translocations or inversions when multiple tags fall on the same chromosome and interchromosomal events when they fall on separate chromosomes. This approach was combined with a defined band multiplexing scheme in spectral karyotyping, which allows detection of multiple kinds of cytogenetics events. However, these methods cannot detect changes much smaller than several hundred kilobases. Such a coarse-grained picture hides a vast array of genomic modifications that vary dramatically in both size and effect.

The development of sequencing technologies (specifically the Maxam-Gilbert [11] and Sanger [12] methods) enabled examination of genomes at basepair resolution over

longer stretches of DNA. Maxam-Gilbert relied on base-specific cleavage of the DNA. The Sanger method used termination of DNA chain extension by incorporation of ddNTPs; the Sanger method would become the dominant technique over the next twenty years largely because it was easier to scale and required fewer hazardous chemicals. These early techniques were automated and commoditized heavily by the mid 1990s. Adaptations in the chemical processes, particularly the invention of removable base terminators and pyrosequencing, led to the development of high throughput short-read shotgun sequencing. Shotgun sequencing would be used to complete the majority of the human genome.

The first draft of the human reference took thirteen years and three billion US dollars to complete [13]. A "complete" version containing 99% of euchromatic sequence (but with many gaps) was announced in 2003. It was decided early on in the study that the reference would be a composite of several individuals so that no individual would be the "gold-standard" human being. In total, sequence from many donors was used to create the reference, though nearly 70% came from a single individual [13]. It was also decided that the reference should be haploid, representing only one-half of the chromosomal material in a typical human cell. This made the reference a standard unit of comparison with only one allele per site and has made it easier to write algorithms against a simpler data structure [14]. However, it also means the reference represents neither the full genome of a single individual nor an average of humanity.

1.1.3 Human genome resequencing studies

The second generation of human genome studies leveraged the human reference to help characterize the diversity present within and between human populations, predominantly focusing on single-base substitution variants. The International HapMap Project began in 2002 with the goal of characterizing small variants in genomes of individuals from diverse populations and coordinating the deposit of these into a public database [15]. In total, the project reported over one million single-nucleotide polymorphisms (SNPs) from 269 people of Yoruban, European, Han Chinese, or Japanese descent. HapMap specifically targeted SNPs with a minor allele frequency $> 5\%$ in specific populations as well as those in linkage disequilibrium (LD), where specific alleles of SNPs are linked into short blocks between recombination breakpoints. These blocks, called haplotypes, are composed of runs of SNPs in LD and are often population-specific. SNPs in LD tend to carry a predictable set of alleles; knowing one or a few marker SNP(s) within the haplotype block allows one to impute the remaining alleles. With the confirmation of linkage disequilibrium in the human genome and a catalogue of haplotypes and SNPs,

the HapMap project enabled researchers to begin assessing which SNPs were associated with traits and disease phenotypes.

Genome-wide association studies became popular in the mid-2000s as a way to correlate genomic regions and specific SNPs to phenotype. Although HapMap relied on the sequenced human reference genome, the individuals in the study were not fully sequenced. Instead, microarrays were designed that assessed genotypes at specific common SNPs at regular intervals in the genome. In the case of Phase 1, this interval was 5kb; Phase 2 dramatically expanded the array to genotype roughly 4.6 million SNPs [16].

HapMap provided a useful if incomplete picture of human genetic diversity, successfully categorizing the majority of SNPs at $> 5\%$ allele frequency [17]. The study could only assess specific alleles at specific sites because it relied on arrays rather than sequencing. While sites in-between could be imputed, this process was not as accurate as sequencing every base. In addition, the HapMap dataset came from only 269 individuals from four populations. While ambitious for its time, the arrival of affordable high-throughput sequencing and a need for more diversity precipitated the need for further studies.

The 1000 Genomes Project used low-coverage sequencing of thousands of individuals to catalogue a much greater number of variants than HapMap [18]. The use of sequencing rather than arrays allowed the assessment of more complex variation. In addition to SNPs the project characterized multi-base insertion-deletion (indel) variants, multi-nucleotide polymorphisms, and structural variants. The initial release of the 1000 Genomes data in 2012 reported 38 million SNPs, 1.4 million indel variants, and more than 14,000 long deletions from 1,092 individuals. A later release in 2015 greatly increased this number, providing a public database of 84 million SNPs, 3.6 million indels, and almost 70,000 structural variants [19, 20].

Multiple projects have extended the pioneering work of HapMap and 1000 Genomes in surveying human genetic diversity. The Simons Genome Diversity Project sequenced 300 individuals from 142 ethnic groups to find unique variants within these populations [21]. gnOMAD analyzed data from almost 20,000 whole genomes and 200,000 exomes at high depth by combining data from many studies and processing it in a single workflow [22]. This produced calls for millions of small variants and hundreds of thousands of structural variants that are polymorphic in human populations [22, 23]. Sequencing studies of hundreds of thousands of individuals are already underway. The 100,000 Genomes Project has sequenced 100,000 genomes from 85,000 individuals in the UK. This study focused on sequencing probands with rare diseases and their families. The UK Biobank study catalogued genotyping arrays from 500,000 individuals along with rich health information.

Over the course of the next five years, the National Institutes of Health All of Us program aims to sequence 1 million individuals from diverse backgrounds.

Since its initial release, over 300 million single-nucleotide variants [24], tens of millions of small insertions and deletions, and around half a million larger structural variants [23] that differ between individuals and the haploid human reference have been characterized. We have also determined that significant portions of an individual's DNA may be absent from the reference [25, 21] and that our understanding of human genetic variation, while much improved, remains incomplete. The differences in our DNA explain much of the variation in human phenotypic traits as well as the underlying cause of many diseases. Today, sequencing human-scale genomes has become routine, taking less than a day and costing roughly $1000. The dramatic drop in the cost of sequencing has led to exponential growth in the number of genomes sequenced worldwide. This, in turn, has dramatically improved our understanding of human genetic variation. Despite the reference genome portraying it as a simplistic, fixed model, our genome is a highly mutable object at both the individual base level and in its large-scale structure.

1.1.4 Cancer sequencing studies

Cancer resequencing studies began shortly after the completion of the human reference genome in 2003 [26]. By 2008, multiple groups had published results from whole-genome sequencing of individual tumors and matched normal tissue [27, 28]. Techniques for purifying and amplifying the coding DNA of the genome (i.e., the exome) and improvements in sequencing technology greatly reduced the cost of performing such analyses for the portions of the genome thought to be most important in cancer. Cancer sequencing studies also sought to obtain molecular data types not used in population studies. This included information about methylation, mRNA, miRNA, and protein abundance.

Major resequencing and multi-omics studies of cancer began around 2008 [29, 30]. The Cancer Genome Atlas sequenced over 11,000 patients during its ten-year timeframe. The International Cancer Genome Consortium combined this dataset with that of the UK's Cancer Genome Project and dozens of other research nodes to generate a set of over 25,000 exome-sequenced somatic pairs. The next generation of studies would focus on analyzing tumors from across many cancer types, often using whole genome sequencing. The Pancancer Analysis of Whole Genomes is the largest of these to date, combining more than 2,800 whole-genome sequenced samples processed on standardized analysis pipelines [31].

The cancer sequencing projects of the last twenty years provided valuable insights into the genomic and molecular signatures of many cancer types. Results from these studies were collected into databases. The Catalog of Somatic Mutations in Cancer provides an overview of individual somatic mutations seen across thousands of samples [32]. The National Cancer Institute's Genomic Data Commons provides mutation information as well as raw sequencing data for thousands of samples including those of The Cancer Genome Atlas project, while the International Cancer Genome Consosrtium maintains its own data portal with tens of thousands of samples. Together, these studies have provided invaluable insight into the spectrum of mutations present in the cancer genome, from single-base substitutions to complex structural variants.

1.2 Mutations in cancer genomes

The mutations observed in cancer cells highlight the genetic plasticity of the human genome as well as that plasticity's role in oncogenesis. Germline mutations are present in all cells of an individual and are passed on between generations. In addition, mutations in the germline occur in humans at a rate of about $1 - 2 \times 10^{-8}$ mutations per basepair per generation, resulting in tens to hundreds of mutations between the genomes of offspring and their parents [33].

Cells in the body also acquire mutations relative to the genetic background of the individual over time [34, 35]. Such mutations are termed "somatic" as they are present in only a subpopulation of cells in the individual and are not passed on to future generations via the germline. Both germline and somatic mutations may contribute to cancer development and span a broad spectrum of size and complexity.

1.2.1 Germline mutations in cancer

The role of germline mutations in cancer was established using pedigree studies. Li and Fraumeni first described a pattern of inherited cancer predisposition in 1969. Li and Fraumeni observed four families each with pairs of offspring that had developed cancer among a collection of nearly 700 medical records. Affected families appeared to show autosomal dominant patterns of transmission for various cancers of soft tissues. Such familial cancer patterns strongly supported the notion that cancer is a disease of the genome as DNA is the molecule of heredity. Later, this inherited cancer syndrome would be attributed to mutations in $P53$, a potent tumor suppressor gene. Mutations in

$P53$ allow proliferation of cells with acquired genomic instability; over time, these cells may become neoplastic. The autosomal dominant inheritance pattern of Li-Fraumeni Syndrome can be attributed to the fact that even individuals heterozygous for loss of $P53$ are at increased risk for development of cancer.

Hereditary cancer predisposition need not travel in an autosomal dominant inheritance pattern, though this is the most common form of inheritance. Xeroderma pigmentsosum is caused by autosomal recessive mutations in the genes XPC, $ERCC2$, $POLH$; the disease is characterized by extreme sensitivity to UV light and a significantly increased risk of certain skin cancers. Fanconi anemia provides another example of autosomally-inherited cancer predisposition where mutations in several genes of the FA DNA repair pathway lead to increased risk of several types of cancer over patients' lifetimes.

Mutations in almost 200 genes have since been implicated in hereditary cancer syndromes [36, 37]. While the vast majority of these are tumor suppressors, there are proto-oncogenes that may be mutated in hereditary cancer syndromes (e.g. *ret* in MEN2 syndrome) [38]. Genes may harbor many such predisposition mutations. Of roughly 25,000 mutations cataloged in the BRCA1 and BRCA2 genes, over 4,000 are associated with a significantly increased lifetime risk of breast or ovarian cancer development [39]. These mutations are frequent targets of genetic screening because of their high contribution to increased risk.

Hereditary cancers due to a highly-penetrant germline mutation account for 5-10% of all cancer cases. Familial clustering without a known predisposition mutation is observed in as many as 10-15% of other cases [40, 41]. These familiar patterns may not adhere to strict Mendelian inheritance patterns. Such patterns may be the product of mutations with low penetrance, interactions caused by similar lifestyles and exposures, or simply chance clusterings of cancer occurrences.

1.2.2 The interaction between germline and somatic mutation

A model for the interplay between germline and somatic mutations in the genesis of cancer was proposed by Knudson in 1971 [42]. Knudson observed distinct patterns of disease in retinoblastoma patients. Retinoblastoma occurrence could skip generations, suggesting an autosomal recessive model of inheritance. Patients with a family history of retinoblastoma also presented earlier and often with bilateral disease, but patients with no family history never displayed bilateral tumors. Knudson assumed that retinoblastoma was due to the inactivation of a single recessive gene and would require loss of both copies for retinoblastoma development. In Knudson's data, bilateral disease fit a model in which

only a single mutation needed to occur for a patient to acquire retinoblastoma. Unilateral disease best fit a model requiring two mutations to occur. This prompted Knudson to propose what became known as the two-hit hypothesis, wherein a second mutation is acquired somatically and the first may be either somatically acquired or inherited via the germline. The two-hit hypothesis provided the mathematical foundation for models of cancer predisposition inheritance via recessive mutation of tumor suppressor genes followed by somatic inactivation of the remaining functional gene copy.

A set of eight "Hallmarks of Cancer" have been used to describe when a population of cells has become malignant [43]. The basis of all of these hallmarks lies in genomic instability (i.e. mutation) of the underlying cellular population. Hanahan and Weinberg proposed an initial six characteristics that helped provide a framework for understanding the neoplastic state. Sustained proliferative signaling allows cancer cells to grow unrestricted. At the same time, evasion of growth inhibitors prevents suppression of this unsustained growth. Cancer cells can acquire resistance to cell death signals, meaning that individual cells may persist longer in the neoplastic state. Cells may exhibit replicative immortality, producing daughter cells for many generations beyond typical programmed senescence. Lastly, in a search for resources, cancer cells may invade surrounding tissues and induce the creation of new blood vessels.

The hallmarks of cancer provide a cell biology framework that facilitates a more clear-cut definition of neoplastic state. In addition, the hallmarks can often be linked to specific genomic alterations that promote or inhibit relevant processes. Mutations which actively contribute to the cancer phenotype are often called "drivers" of the neoplasm. Much as the driver of a car determines the direction of travel, driver mutations determine the evolutionary trajectory of the cancer cell population by participating in one or several of the hallmarks directly.

Cancer cells may harbor hundreds of thousands of mutations compared to the germline background from which they originate [44]. Mutation rates differ by orders of magnitude across cancer types and between tumors of the same organ. Pediatric tumors may harbor only a single clone of cells with less than a dozen mutations [45], while certain melanomas and breast cancers may harbor multiple subclones with dozens of mutations per megabase. Some of these populations will harbor multiple driver mutations either in a single clonal population or across multiple subclones within the tumor [46]. The mutation rate of a tumor is not an effective predictor of its growth; however, a higher mutation rate does increase the risk of treatment resistance and cancer recurrence [47].

Only a small minority of mutations in cancer cells confer selective advantage. Though the number varies across tissue of origin, on average approximately five mutations within a tumor are under positive selection [48, 49]. An even smaller proportion (less than one per tumor, on average) are under negative selection while the remainder are selected neutrally. All of these mutations arise randomly due to genomic instability, errors in repair, or exposure to mutagens [50]. In principle, mutations may occur at any site in the genome, though mutations at some sites (e.g. those which disrupt essential genes) are not tolerated by cells. As only 1-2% of the genome encodes protein sequences [51], mutations will most often fall in non-coding regions.

Mutations which do not confer selective advantage either on their own or in concert with other mutations have been termed "passengers" to distinguish them from "driver" mutations which contribute to the cancer phenotype. The total number of passenger mutations in a given cancer mutation catalog is often two or three orders of magnitude greater than the number of drivers [49]. It is important to note that under this definition a mutation need not be sufficient or necessary for tumor formation to be considered a driver. Because neutral passenger mutations may be present in clones with positively-selected drivers, such mutations may rise in frequency within a given tumor without conferring a selective advantage. The probability of seeing the same neutrally-selected mutation at high frequency across tumors from different patients is very small, however, and so observing the same mutation multiple times across a number of tumors is often considered evidence in favor of that mutation being a driver.

1.2.3 Single nucleotide variants

Single nucleotide variants (SNVs) are common cancer driver mutations and are the best characterized of the mutational classes. Most single nucleotide driver mutations reported in previous studies are in coding regions. Single nucleotide substitutions may be activating or inactivating. The $BRAF^{V600E}$ mutation is a single-nucleotide change that is an established driver across many cancer types. $BRAF$ is a protein kinase that is important in the $MAPK$ signaling pathway, which controls a broad array of cellular functions involving growth and proliferation. A substitution of glutamic acid for valine at codon 600 in the protein causes constitutive activation, leading to constitutive signaling in the downstream $MAPK$ pathway. $BRAF^{V600E}$ is a primary driver of thyroid cancer, malignant melanoma, and carcinomas of the lung and colon [52].

There are examples of non-coding SNVs that drive tumorigenesis. Mutations in the promoter of the $TERT$ telomere reverse transcriptase gene are found across cancer

types but are especially present in hepatocellular carcninomas. These mutations are the most frequent genetic alteration in these neoplasms and confer a significant immortality advantage by preventing telomere shortening. In thyroid cancer, promoter mutations in the $TERT$ gene are associated with more aggressive tumors but are insufficient to generate a malignant state [53]. Such mutations, which are insufficient to drive tumorigenesis on their own but which contribute to more aggressive disease, are sometimes referred to as "backseat drivers" to distinguish them from the primary driver of a tumor (Gad Getz, private correspondence).

1.2.4 Indels and multinucleotide variants

Small variants affecting more than one nucleotide are also common in cancer. Indel variants are common somatic mutations and can act as drivers. Most tumors will harbor approximately 1 somatic indel for every 10 somatic SNVs [44]. However, tumors with mutations leading to microsatellite instability and those with other mutations affecting DNA repair can have higher ratios of indels to SNVs [54, 55]. Indels in short tandem repeats have been implicated as drivers when they occur in specific genes [56].

Indel variants are more difficult to detect than SNVs. This is partially because alignnment algorithms often penalize gaps more than mismatches, though parameter tuning can help alleviate this effect. This is further exacerbated by the fact that indels are most often located in repetitive regions. Long reads improve indel calling since they can span the entire repetitive region, though homopolymer runs still present challenges.

Multinucleotide variants (MNVs), in which multiple adjacent bases are substituted but not inserted or deleted, also occur. Often, one of these variants will be protein-altering and the second mutation will revert the codon to produce the original amino acid. MNVs have been shown to occur at roughly 1-3% of the SNV rate in the germline [57]. Somatic MNV rates vary across cancer types, with lung and skin cancers having five times the rate of other cancers due to the effects of tobacco- and UV-induced DNA adducts [58]. MNVs are often deleterious in genes associated with developmental disorders [59] and can form driver mutations at well known cancer hotspots such as $BRAF^{V600E}$ [58].

1.2.5 Mutational signatures

Mutations in cancer may arise from a number of mutagenic processes, both exogenous and endogenous. Some mutagens act through distinct mechanisms which leave defined patterns of mutation in the DNA. Chemicals such as those in tobacco smoke are well-

established mutagens. So too are energetic particles such as UV light or ionizing radiation. DNA repair processes, whether acting normally in response to stimulus or when aberrated, may induce mutations in the cell endogenously.

Many of these processes produce specific mutational profiles, or signatures, in the genome. Understanding the signature produced by a mutagen can help inform clinicians about any exposures that may have led to oncogenesis. Mutational signatures may also be indicators for specific clinical treatments. While more than eighty signatures have been deciphered many of these have not yet been attributed to an exposure or endogenous process [44]. There has been significant interest in new methods for detecting signatures and attributing them to specific mutagens in the last ten years.

Mutational signature extraction relies on dimensionality reduction techniques such as non-negative matrix factorization [60]. These algorithms are used to generate vectors that approximately model the mutational patterns of a given exposure (the signature) and the amount of that signature seen in the genome (the amplitude or dosage). Most signatures examined so far have been on the 96 trinucleotide contexts of the genome. Many mutagens leave signatures in these contexts, and they are easy to interpret while limiting the number of features in the model.

Signatures have been found in higher-order (e.g. pentanucleotide) contexts as well as by examining features based on microhomology, strandedness, and length of indels and characteristics of structural variant breakpoints. Today, several dozen signatures are commonly observed across a range of cancers, many of which can be attributed to a particular exposure [44]. In addition, an increasing amount of evidence points to signatures or combinations of signatures caused by non-linear interactions between mutagens [61]. Signatures for exposure to particular environmental agents have now been established in human cell lines [62], and as our ability to create single-exposure systems through gene editing improves we will likely gain even more insight into how mutational signatures manifest in the genome. Mathematical advances, such as the introduction of denoising and sparsity-enforcement techniques from the field of signal processing should also provide improved methods for detection.

1.3 Structural variation

Structural variants are typically defined as any changes in the genome larger than 50 basepairs in size. This definition encompasses one- or multi-breakpoint events as well as chromosomal abnormalities such as chromothripsis. On average, structural variants

affect a much larger number of basepairs within a given individual than do small variants [19, 63, 64]. Although large structural variants are easily visible by microscopy and were described early in genomic history, they have not been as well characterized as SNVs and indels in the modern era. This is largely due to their size relative to sequencing reads, forcing them to be called heuristically by mismapping signals rather than directly as small variants are. Reads overlapping either breakpoint of a variant are either softclipped or split-mapped, and often have lower mapping quality versus reads mapped in invariant regions. Structural variants also tend to be present in repetitive regions, further frustrating read mapping and discovery.

1.3.1 Simple structural variants

Simple structural variants (i.e. those that are formed from relatively few breakpoints) can be divided into several classes. Balanced events do not produce a change in copy number. Breakpoints may be reciprocal (i.e., DNA strands on either side of the breakpoint are evenly exchanged) or portions of the genome may be lost during the event. Unbalanced structural variants produce a change in copy number. Unbalanced structural variant types include deletions, foreign sequence insertions, and duplications. These vary greatly in size and aetiology. Foreign sequence insertions such as ALU element movements are highly polymorphic in human populations.

Balanced events include inversions and translocations. Inversions, while rare in the general population, occur at a much higher rate in somatic variation [65, 66, 54]. Translocations may place distant parts of the same chromosome together to form new subchromosomal structures. They may also place portions of sequence from two or more different chromosomes together. These events are again quite rare in the general population but have been identified frequently in cancer and other diseases.

1.3.2 Chromoanagenesis and multibreakpoint events

Beyond the simple structural variant forms, many tumors exhibit large and complex structural changes. These may involve one or many chromosomes and may be accompanied by a high local substitution rate. Such dramatic changes in the chromosomal structure have collectively been labeled chromoanagenesis [67].

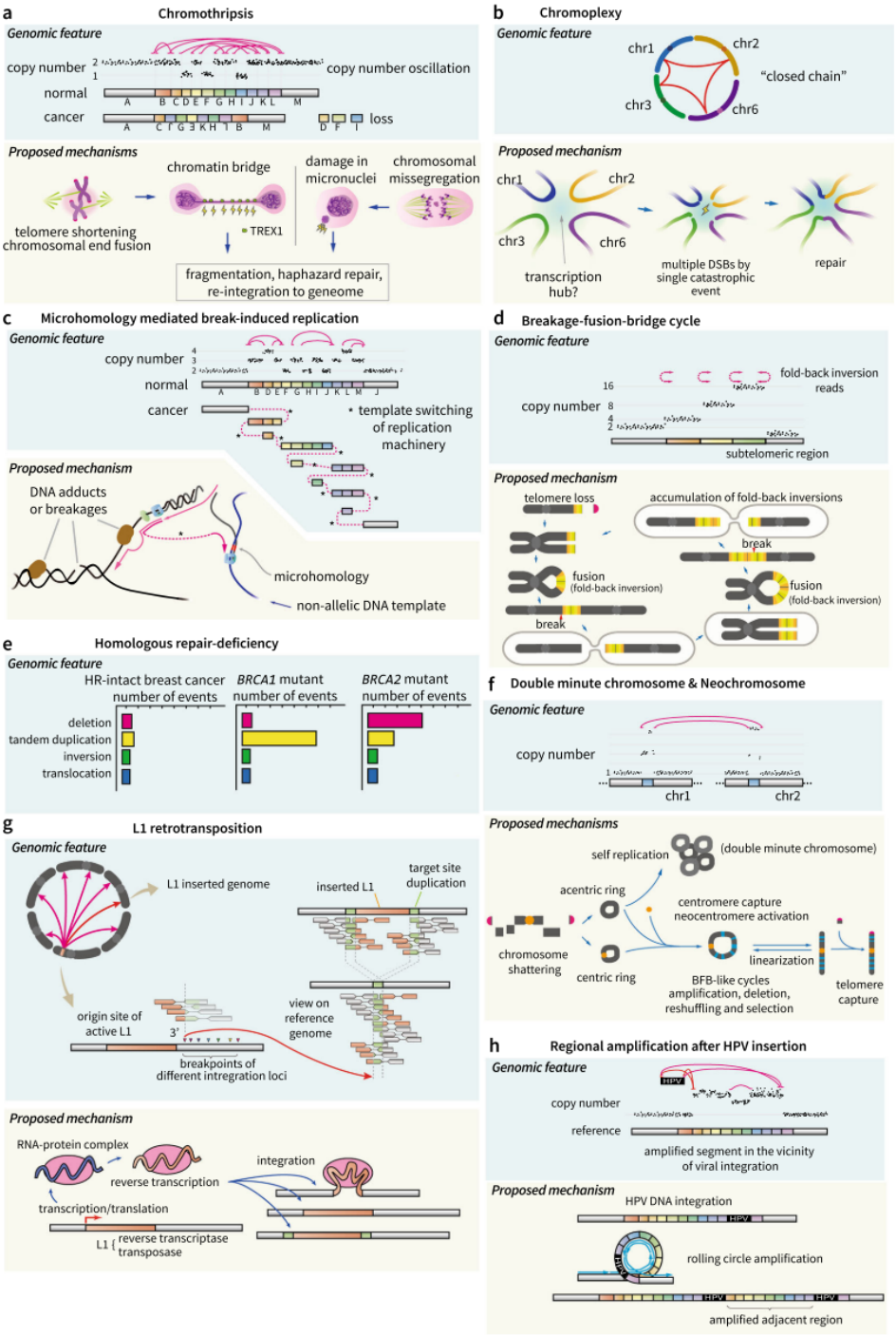


Fig. 1.1 Types of complex structural variant events, reproduced directly from Figure 3 of reference [1].

Several distinct patterns of chromoanagenesis have been described in cancer. Reference [1] describes these in detail and an overview of common patterns is reproduced from their paper in Figure 1.1.

Localized SNV hypermutation, or kataegis, may occur within chromothripsis regions. While not considered a structural variant event, kataegis is a common pan-cancer pattern and can be an important marker for genomic instability [68]. Mutations in $TP53$ are associated with kataegis. Hotspots often reflect the activity of the $APOBEC$ family of cytidine deaminases as well. While common across cancers, bladder carcinomas display a much higher rate of kataegis than other cancer types [68].

Chromothripsis was first described in 2011 and is characterized by apparent chromosomal shattering and repair. This generates a series of what may be more than 100 clustered events on one or a few chromosome arms. All types of non-insertion simple event (i.e., deletions, translocations, and inversions) can be present and are often present at similar ratios. A defining characteristic of chromothripsis is locally oscillating copy number over the event's span.

Chromoplexy was first observed in prostate tumors by Baca *et al.* [69]. Chromoplexy is characterized by the involvement of multiple chromosome arms in a series of interchromosomal translocations [1]. These are most often balanced and copy-neutral. The total number of composite events is usually much less than in chromothripsis, and events most often form a complete cycle (or "closed-chain") among the breakpoints of affected chromosomes [69]. It has been proposed that chromoplexy is caused by erroneous repair of double-strand breaks in the DNA.

Other well-defined patterns of hypermutation have been described. Fold-back inversions are present in roughly one-sixth of pancreatic tumors [70]. These events are defined by a duplicated DNA portion with everted (i.e. outward-facing) read pairs at the breakpoint. Fold-back inversions are hypothesized to be a product of breakage-fusion-bridge cycles, first described by McClintock in 1941 [71]. Mutations in genes involved in the DNA homologous repair pathway (such as $BRCA1$ and $BRCA2$) can generate high rates of copy-number altering events at the structural variant scale as well as in small indels. L1 retrotransposition can lead to insertion of the L1 retrotransposon sequence at novel locations in the genome. Lastly, viruses such as Human Papillomavirus can insert into the genome and induce localized duplications near the integration site.

1.3.3 Structural variants in cancer

Structural variants are well established as important contributors to disease. Structural variants have been implicated as disease drivers in schizophrenia [72, 73], Down Syndrome [10], epilepsy, hemophilia, pancreatitis, and numerous disorders of the nervous system [72, 74–76]. Many structural variants, acting as either passengers or drivers, have been

characterized in cancer. Because of their complexity and size, structural variants may impact the cancer phenotype in diverse ways.

Structural variants were perhaps the first genomic abnormality to be described in cancer. Theodore Boveri first hypothesized that chromosome abnormalities visible by light microscopy might lead to cancer development in 1914 [8]. One of the earliest descriptions of a structural variant as a driver was made by Nowell and Hungerford in the 1960s. A specific chromosomal abnormality was identified by microscopy in patients with chronic myelogenous leukemia. This abnormally small, recurrent chromosome was termed the Philadelphia Chromosome [9]. As cytogenetic techniques improved, the genomic portions involved were located to chromosomes 9 and 22 and it was determined that a reciprocal translocation generated the Philadelphia Chromosome. Later, the mechanism of action was discovered to be generation of a gene fusion product, $BCR-ABL$, which leads to constutive activation of the $JAK/STAT$ and $MAPK$ pathways and unchecked cellular expansion.

Structural variants may impact the copy number of genomic regions. Unbalanced variants which alter the copy number of genomic portions may directly affect gene dosage. Loss of tumor suppressor genes such as $P53$ in the genome allows cells to proliferate uncontrollably. Genomic loss in the short arm of chromosome 22 is an apparent pan-cancer backseat driver mutation, although specific genes that contribute to this phenotype have not been described. Loss of heterozygosity, which may be caused by imperfect replication or partial chromosome loss, can result in loss of functioning gene products in the affected region. Duplication of oncogenic regions can increase gene dosage. This mechanism is commonly seen among genes in the RAF-RAS signaling pathway, including MYC, RAF, and RAS [77].

Structural variants may remove or join functional domains of proteins by altering the exonic sequence of the DNA. Large deletions of inhibitory domains can cause affected genes to become constitutively active. A deletion of six exons in BRAF between exons 2 and 8 has been associated with increased resistance to treatment [78]. This mutation causes deletion of the RAS-binding domain and constitutive activation of the gene product, leading to constant signaling in the $RAF-RAS$ pathway. Structural variants may also form activating fusions involving multiple gene partners. Gene fusion drivers are common across cancers. In thyroid cancer, $RET-PTC$ fusions occur between the RET receptor tyrosine kinase and one of several gene partners in roughly 15% of sporadic thyroid cancers. The most common $RET-PTC$ fusion ($RET-CCDC6$) is driven by a multi-megabase inversion on chromosome 10, while other fusion partners form after

interchromosomal translocation events. In lung cancer, the $ALK - STRN$ fusion drives a significant portion of tumors. This specific fusion is also targetable by therapy, however, and treatment to deactivate the fusion protein proves effective for many patients.

1.3.4 Difficulties in structural variant discovery

Large, simple structural variants are easily detected by microscopy. Large inversions and translocations are often genotyped using flourescence in-situ hybridization techniques. The original FISH protocol was invented by Rudkin and Stollar in 1977 [79]. Today, multicolor FISH and spectral karyotyping techniques provide effective methods for analyzing variants. In these methods, flourescent genomic probes are hybridized to targeted DNA sequences. Differences in expected banding patterns can be used to discern normal from abnormal chromosomes. Spectral karyotyping combines the use of multicolor FISH probes with a defined chromosomal color banding scheme to look at events across the genome. However, the minimum event size for intrachromosomal events in either of these approaches is still roughly 1 mbp [66]. Interchromosomal exchanges are easier to detect as they are not size-dependent. These methods represent the current state of the art for assaying radiation-associated structural events [80, 81]. Structural variants, particularly inversions and translocations, have been noted as stable, long-term markers of radiation exposure. Smaller events have also been noted in individuals exposed to large amounts of medical radiation [54].

Structural variant detection with short next-generation sequencing reads relies on identifying imperfect read mappings to the reference and translating these signals into structural variant calls. The earliest short read structural variant callers relied on discordant paired-end fragment length and soft clipping of reads to detect where they overlapped structural variant breakpoints [82–84]. Over time tools would improve detection of all event classes by incorporating these signals together [85, 86]. Other tools utilized assembly-based approaches to detect the full spectrum of events, with a particular advantage in detecting insertions [87].

Several new technologies have led to renewed interest and ability in assaying structural variation over the past decade. Oxford Nanopore and Pacific Biosciences (PacBio) have both released sequencing platforms capable of producing reads as long as one megabase. Such remarkably long read lengths allow the capture of whole structural variant alleles in a single read. Mapping these reads required the development of new read mapping algorithms and structural variant callers [88]. The cost of such length is that both technologies have per-base error rates as high as 12%, though these are continually

improving [89] and, in general, read length rather than basepair accuracy is more important for structural variant detection. Algorithmic improvements to read mapping and variant calling were required to make analyzing such data feasible, though it now outperforms short read data for detecting most classes of events [88]. Linked-read technologies, such as those from 10X Genomics and Moleculo, have also enabled better resolution by providing information about the larger molecules of DNA from which sequencing reads originate [90]. Lastly, Hi-C, optical mapping, and other technologies have allowed phasing large chunks of chromosomes including when structural variants are present, especially when combined with long reads [91]. These advances in technology have highlighted that structural variation plays key roles in both healthy individuals and disease.

Structural variants present unique challenges in resequencing and assembly compared to small variants, mostly due to their size relative to sequencing reads. Small variants are often recurrent and can be easily defined by a single position in the genome. Procedures for producting consistent representations of individual small variants are well-defined [92].

Normalizing variant representation for large variants remains an unsolved problem in the field. While structural variants often generate similar genomic structures or transcript products it is difficult to normalize variant calls in a consistent manner. The same structural variant may vary in its breakpoints by thousands of basepairs. Variants may also overlap, be composed of multiple breakpoints, or be fully nested within other variants. The $RET-CCDC6$ (PTC) fusion is a good example of a structural variant that produces a recurrent, well-characterized gene fusion product with relatively consistent breakpoints. The individual breakpoints of the fusion still vary over several hundred basepairs in the introns of $RET - CCDC6$. Non-coding variation is more difficult to classify as recurrent or non-recurrent given that there is not necessarily a consistent genomic product and breakpoints may therefore not be under functional selection.

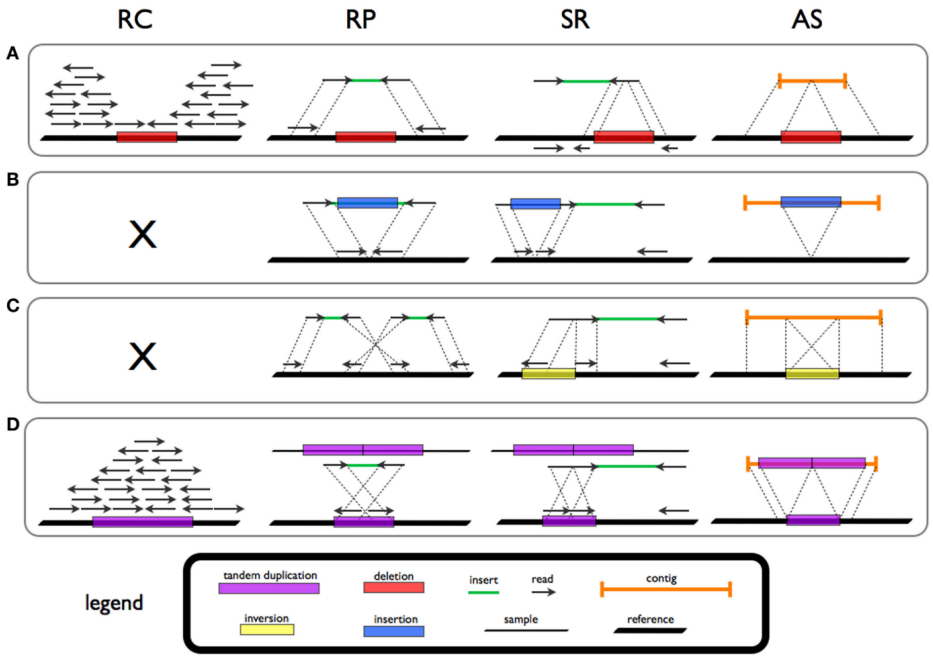


Fig. 1.2 An overview of discordant read signatures induced by structural variants. Duplicated directly from figure one from [2]. Columns represent different types of short-read evidence for a structural variant based on read count (RC), read pair (RP), split read (SR), and assembly (AS) methods for (A) a deletion, (B) a novel sequence insertion, (C) an inversion, and (D) a tandem duplication.

Short-read technologies have dominated the era of resequencing due to their cost advantages. The relative size of these reads compared to structural variants means that individual reads (and read pairs) often fail to span both sides of a variant's breakpoints. When a read spans a single breakpoint, it can still be difficult to map across the break as the portions on either side will be small; this is further confounded by the fact that structural variants are often present in repetitive regions or in regions of microhomology at the breakpoints.

In addition, small-variant calling algorithms are not well suited for locating structural variants, and the two classes of variation have traditionally been analyzed separately even though they often occur together. Structural variant calling pipelines often use a consensus step to harmonize calls among many structural variant callers and produce a higher-specificity callset. Consensus pipelines have been shown to improve both sensitivity and specificity over running individual tools [93]. Running multiple algorithms and merging structural variant calls adds significantly to the computational and analytical burden presented by such data. Such pipelines can be prohibitively expensive and difficult to manage for smaller studies.

Long read technologies produce reads that are often of sufficient length to fully span many smaller structural variants and provide anchoring sequence to breakpoints. Linked reads provide many of the same advantages and can yield useful information about molecule of origin. However, both of these technologies introduce computational challenges. Linked reads rely on special software to incorporate Unique Molecular Identifier information for mapping, variant calling, and phasing; without this, they are essentially just standard short Illumina reads. Long reads require new algorithms for effective mapping due to their length and error rate, though this is improving with greater read fidelity. Nonetheless, the process of aligning long reads remains slow and expensive.

For long reads, read error rates confound structural variant calling. Homopolymer runs have proven especially hard to sequence, meaning repeat expansions/contractions are difficult to genotype. Sequencing errors also frustrate read alignment and small variant calling, though the random or pseudo-random nature of long-read errors (outside of homopolymers) means that this error can be effectively scrubbed given sufficient depth. This still adds to the complexity and expense associated with calling such variation.

Further confounding structural variant comparison is the lack of a common format for variants. While the Variant Call Format has special syntax designed for structural variation calls, many callers produce values for the SVLEN, SVTYPE, or END fields that deviate from the description in the specification. Complex events in "breakend" format are often not convertable back to the more human-readable SVTAG format. The required information fields are not always clearly defined in the specification. There is also no notion of left-alignment and trimming — as there is for small variants — that minimizes the effects of variable positon and imprecise representation. Representation and naming of complex events is especially challenging given that they may span many VCF records.

1.4 The rediscovery problem

The standard practice for analyzing new genomes is known as resequencing, in which the genome of interest is shotgun-sequenced, its reads are mapped to the reference genome of its species, and variant locations are called against the reference. Whether the variant has been seen in the population before is ignored at this stage; recurrent variants in the population are "rediscovered" in the individual as if they were novel. A common next step in analysis is to annotate variants from resequencing with allele frequencies and impacts from population and functional genetics studies.

This circular process wastes considerable computational effort, though it is still significantly more cost-effective than performing whole-genome assembly with current practices in sequencing and computing. While every new genome will differ by 3-5 million SNVs and indels as well as millions of bases affected by structural variation, many of these variants are likely to have been previously observed. In diseases such as cancer, only a very small proportion of variants, perhaps even a single one, drive the disease phenotype. Such variants are highly recurrent and have in general been well-cataloged.

Modern genomic pipelines were largely designed in the mid 2000s and fail to make use of prior knowledge of population variation before the annotation stages. Sample sizes, however, have grown exponentially over the past decade. Large institutes now regularly sequence more samples in a month than were sequenced in the first 15 years of the field. At the same time, the data point for each individual sequenced grows richer with longer reads and new technologies. Even at the largest research centers it has become difficult to maintain sufficient computing infrastructure to process the ever-increasing amounts of data; because of this, traditional high-performance compute systems are being phased out and replaced by commodity cloud systems. While computing power has increased in exponential fashion according to Moore's Law, the cost of analysis has not dropped concurrently with the cost of sequencing.

1.4.1 Costs and limitations of computing at scale

Genomic data sizes have grown dramatically over the last twenty years. Initially, computational power and cost of storage kept pace with sequencing rates. In 2007 this trend reversed, with sequencing costs dropping roughly 2-3 times as fast as storage cost due to the introduction of massively-parallel shotgun sequencing approaches. A typical 30X short-read human genome now costs roughly $1000. Sequencing studies with tens to hundreds of thousands of participants are common, and studies with over a million individuals are underway.

Shortly after the introduction of the 454 and Illumina-Solexa sequencing machines it became apparent that smaller genomics centers would not be able to provision sufficient IT infrastructure fast enough to keep up with the amount of data being produced. This led some in the community to advocate for moving genomics data to the cloud, rather than local high-performance computing (HPC) environments which up until this point had been the dominant location for genomic analysis [94]. However, it would take many years for this transition to take place because of the regulatory and technological challenges such an undertaking presented. Many major studies, including the Pancancer Analysis of

Whole Genomes, have now been processed in the cloud, often seeing significant savings in cost and time [31].

For power users, HPC has been the preferred option for genomics data processing. Traditional HPC environments provide local storage and fast, on-demand compute running on bare-metal servers. Typically, job management is performed using a scheduling program, and user authentication is handled by SSH, sometimes accompanied by a firewall or VPN to prevent unwanted incursion to the server and its data. When performing genomic analyses, data is downloaded from collaborators or central repositories. Processing can then be done using any number of programming languages or licensed commercial software instances. Most academic centers maintain local compute clusters with teraflops of compute and several petabytes of storage.

While the cloud offers nearly-limitless compute and storage it brings with it a number of notable differences from the familiar architectures of HPC. Compute in the cloud is performed on virtual machines rather than bare metal servers. These must be provisioned and destroyed by users, and there is often no scheduler to handle such provisioning automatically. Installation of custom scripts on the cloud often falls to the user, whereas system administrators have typically performed this role in on-premise computing centers. Billing from the cloud is often much more transparent than from HPC centers; this can lead to sticker shock when users see the per-unit price without considering that local HPC is often heavily subsidized and its costs amortized over many years of operation.

New technologies offer some solutions for managing these challenges. Containerization, where software is packaged inside a runtime and can be distributed independently of a physical server, allows users to share code without sharing access to servers or data. In doing so, the new paradigm has become movement of code to data, rather than the copying and transfer of data to code on local compute. This dramatically improves reproducibility and lowers the barriers to entry for computing on the cloud. It also reduces data duplication across large studies. Most cloud providers support containers, and their performance is comparable to cloud virtual machines. In some use cases containerization is being replaced by managed services, where cloud computing companies provide access to highly-abstracted analysis pipelines through externally exposed APIs. These services reduce the development burden on users at the cost of transparency and flexibility.

An alternative approach for reducing the gap between compute capability and sequencing data influx is to develop algorithms that are more efficient. Algorithms designed to operate on subsets of data or which use heuristics to avoid excess work can significantly improve throughput. Examples of sublinear algorithms include the MinHash

implementation in [95] and [96]. Succinct data structures [97] may also be used to reduce the amount of memory needed to process data; these have been applied in several recent tools [98, 99]. Reference-based compression functions similarly and significantly reduces file sizes for aligned NGS data [100]. More efficient data structures and algorithms enable the use of smaller units of computing for analysis. This can make such analyses accessible to groups which might not be able to afford them otherwise. It also opens up new areas for remote research, such as real-time analysis at point-of-care or in environments that are not amenable to internet connection or compute infrastructure.

1.4.2 Variation-aware algorithms

A more efficient method for resequencing would be to align reads in a manner that incorporated prior population genetic information and genotype variation directly. Such alignment algorithms are "variation-aware," in that they are able to incorporate prior knowledge about genomic variation. Variation-aware algorithms also include kmer-matching algorithms which genotype known variants [96]. By incorporating prior information, variation-aware algorithms can significantly reduce runtime and provide better sensitivity. However, because they are only as good as their prior, variation-aware algorithms do not provide any benefit over other approaches when analyzing novel variation.

Significant interest has arisen around research into methods that can decrease the amount of reference bias in resequencing studies. Modified references containing alternative alleles have been used to improve population-specific studies [101]. Others have explored replacing all variant alleles in the reference with the major alleles in an effort to reduce population bias in the reference [102]. Disease-specific and personal references have been shown to improve read alignment and downstream analyses [103]. Efforts such as these seek to undo some of the decisions made for the sake of simplicity during the generation of the human reference and to create reference structures that are more representative of human populations.

Several methods exist for mapping reads in a variation-aware manner, though none have achieved widespread acceptance. HiSat2 [104] uses a set of hierarchical graph FM indexes to align reads to tiling graphs of small genomic regions containing variation. This has been shown to improve RNA-seq analysis as well as HLA typing and analysis of DNA fingerprinting [105]. The variation graph toolkit implements similar algorithms and has been shown to reduce reference bias when aligning to population variation [99]. In theory it should be possible to align reads to known variants, perform calling of known

and novel variants, and retrieve allele frequency and other annotations in a single process. Though no implementation of such a pipeline exists, it is reasonable to assume it could save significant amounts of time and money compared to current alignment, calling and annotation pipelines.

The common coordinate system provided by the reference genome is a major advantage of resequencing over genome assembly. The ability to refer to a variant by reference version, contig and position enables a consistent language for referring to variants and was a primary goal of the Human Genome Project. This enabled the generation of a massive body of genome annotations that is persistent and consistent across individuals, saving significant amounts of computational effort over personalized genome assembly and annotation.

Alternative data structures have also been proposed and are generating significant interest as possible replacements for the linear reference. Graph-based representations of genomes have become the most-common of these, and have been shown to reduce reference bias while still permitting usage of the reference-based coordinate system [106–109]. Graphs have been used in RNASeq analysis, small variant detection and structural variant typing and have consistently been shown to improve read mappings overall, though further analysis of variation to include in the graph is required.

1.5 Graph representations of genomes and pangenomes

Graph genomes are a type of variation-aware data structure that are under active development within the genomics community. The linear reference has multiple issues of representation, as discussed in subsection 1.1.1 and section 1.4. It is desirable to instead have a reference that represents collections of genomes. Such a collection is referred to as a "pangenome." A pangenome may contain genomes from the cells of a single individual, many individuals from a single species, or a metagenome from a collection of many species together. Pangenomes have been previously implemented and have provided novel insights in a variety of fields. Metagenome indexes are perhaps the most explored implementation of pangenomes, though read collections and MinHash sketch databases are other examples of such structures [110–112].

Genome graphs provide an alternative to the linear reference for representing the reference sequence and reference relative variation in a single common data structure [113]. A graph $G_{V,E}$ is defined by a set of nodes V and a set of edges E. In most graphs, nodes define objects and edges define relationships between nodes. Graph data structures

are common in network analysis, and research into graph data structures has increased with the rise of social networks in the 2010s.

Fig. 1.3 A variation graph containing five substitution variants. The model used in `vg` automatically generates edges for all possible haplotypes when incorporating a variant.

Many types of graph have been used to represent genomes. Assembly algorithms have historically used de Bruijn graphs or overlap graphs to represent the contiguous relationships between short reads [114–116]. The string graph has been used as a more generic model for genome assembly and variant calling [117]. A general model, the sequence graph, is used by [99] and [108] to represent population variation called against a reference genome. Several implementations of the variation graph have been used for graph alignment, query, and variant calling. Graphs have also been utilized to represent partial order alignments of both DNA bases and amino acids and splicing within RNA transcripts [118–122].

In a genome graph, genomic sequence is represented along with relationships between those sequences. The most common representation is the variation graph, in which nodes contain genomic sequences and edges represent relationship between nodes;alternative representations placing the sequence on edges do exist [108]. Annotations in a variation graph are overlaid via paths, which describe collections of nodes and edges that share attributes. Haplotypes trace paths along nodes and edges in the graph. The variation graph toolkit `vg` provides one such implementation of this structure, and it has been shown to reduce reference bias and improve read mappings [99]. Tools for producing and operating on graphs remain nascent and there is ample opportunity for improvements in both theory and practice in the field.

1.6 Genomic mixtures, subclones, and uneven heterogeneity

It is exceedingly rare that an individual genome is sequenced on its own. Historically, it has been difficult to isolate a single cell for sequencing, though techniques for doing so are now routine and commoditized [123, 124]. Many organisms are multiploid and store

multiple homologous copies of DNA within even a single cell. These copies may be hard to distinguish computationally. As genome assemblies are typically presented in a linear fashion, only one haplotype is chosen to be representative. This process is imperfect and haplotigs are often merged in assemblies. Techniques for binning reads using trio data have demonstrated that assembly of individual haplotypes is possible and genetically informative [125, 126]. Techniques for separating haplotypes and removing haplotigs post-assembly and without the use of trios are under active development [127, 128].

1.6.1 Metagenomics

There has been significant interest in studying mixtures of genomes as next-generation sequencing costs have fallen and DNA barcoding techniques have improved. The field of metagenomics deals with such studies. Metagenomics is a broad field, ranging from sampling of the human microbiome [129] to collection and analysis of samples from the oceans [130] and the International Space Station [131]. Common goals of metagenomics studies include sampling of cryptic diversity from the environment [130] or detection of specific traits such as antibiotic resistance [132].

Deconvoluting mixtures of genomes is computationally expensive, and recent research has focused on data structures and algorithms for accelerating the process. Reference [110] describes Kraken, a high-performance tool which uses a large database of kmers to lowest common ancestor of all genomes containing each kmer to accurately identify species within a sample at the genus level. This tool improved significantly on the performance of BLAST, one of the most widely-used algorithms for querying a sequence against many known sequences [133]. Reference [95] used the MinHash locality-sensitive hashing scheme to compare genomes in time sublinear to the total sequence length by taking a representative "sketch" of each genome. This process has been used to approximately map long reads as well to perform overlapping for genome assembly [134]. Such approaches have made metagenomic analyses feasible for very large databases of known sequences. Reference [112] used Sequence Bloom Trees to make comparison of read sets to all known viral and bacterial sequences feasible using contemporary computing hardware.

1.6.2 HPV16 lineages and sublineages

Metagenomic analysis of HPV16 infections is of major interest to cancer researchers. Human papillomavirus (HPV) is a DNA virus responsible for over half a million cervical cancer cases each year and an estimated 239,000 deaths worldwide [135]. Persistent

infection with one of the carcinogenic HPV types is necessary for invasive cervical cancer development, and accounts for a large proportion of other anogenital and oropharyngeal cancers [136]. There are more than 200 papillomavirus types known to infect humans, with each type defined on the basis of at least 10% sequence difference in the L1 gene (major capsid protein) sequence. Not all HPV types contribute equally to infection or disease risk. Approximately a dozen of the more than 200 HPV types are considered carcinogenic, with just two types, HPV16 and HPV18, accounting for approximately 75% of cervical cancer cases worldwide [137].

HPV infection is not mutually exclusive to a specific type [138]. Concurrent infection with multiple HPV types is common, occurring in 20-50% of HPV infections [138–141]. One study reported nine distinct HPV types simultaneously in a single patient [142]. Co-infections appear to be random assortments of types with no evidence to support clustering of types or viral interactions between types [139].

Within each HPV type there are variant lineages which differ by 2-10%, and as little as 1% for sublineages, in their L1 gene sequence from other variants of the same type, and these also vary in risk for cervical precancer and cancer [143, 144]. For HPV16, the most common and carcinogenic type, there are four main variant lineages (A, B, C, and D) and ten sublineages (A1, A2, A3, A4, B1, B2, C, D1, D2, and D3) that are roughly correlated with their geographic distribution. HPV16 sublineages show strong differences in histology-specific cervical precancer and cancer risks, with relative risks exceeding 100 for specific sublineages (D2, D3 and A4) associated with adenocarcinoma [4].

Mirabello *et al.* [4] used phylogenetic methods and lineage-specific SNP genotyping to detect HPV16 lineages. While able to accurately determine the dominant lineage, Mirabello *et al.* were not able to assess whether samples were infected with multiple lineages. There is little known about the epidemiology of co-infections with multiple HPV16 variant lineages, though this is clinically relevant given the significant differences in risk associated with each lineage.

1.6.3 Tumor heterogeneity and clonal hematopoesis

Tumors often contain significant amounts of heterogeneity [46]. In many cancer types the primary tumor may have multiple subclones with independent driver mutations. Subclones may also harbor resistance mutations and it has been shown that targeted deep-sequencing of tumors has the potential to improve patient care by providing a better overall picture of tumor genomic architecture [145, 146].

Subclonal expansions are common in cancer and normal cells. Reference [147] found evidence for subclonal expansion in all metastases and 94.7% of tumors they examined. Even without the presence of high-fraction subclones, the tumor itself is a population of cells. It is well established that normal cells acquire clonal mutations over time [148, 149]. These mutations appear to be under neutral or slightly positive selection [48]. These same patterns hold for normal tissue, where it has been shown that clones with mutations conferring growth advantages colonize epithelial tissues such as the skin and esophagus [149].

Somatic clones can also colonize the blood. Mutations in several genes are known to lead to clonal hematopoesis, in which a single hematopoetic stem cell lineage produces a substantial portion of mature blood cells [150]. This phenomenon is associated with later development of hematologic cancers as well as heart disease and ischemic stroke [151]. $DNMT3A$, $TET2$, $ASXL2$, $JAK2$, $SF3B1$ $SRSF2$, and $TP53$ are commonly mutated in clonal hematopoesis [152]. These genes had previously been identified as drivers in acute myelogenous leukemia and myelodysplastic syndrome.

Clonal expansion plays an essential role before and during early tumorigenesis [153]. Mutations can be induced by exposure to mutagens [154, 62] and by predisposing germline mutations [45, 38]. Driver mutations are acquired by chance and subsequently rise in clonal fraction because they confer a fitness advantage over wild-type cells [155]. Clones with driver mutations in cancer genes colonize surrounding tissues and are under positive selection [149, 48]. It can take many decades for an initial clonal lineage to become malignant, and many will never do so [151, 155]. Clonal expansion in normal tissue therefore provides an early indicator of possible precancerous state.

1.7 Radiation-associated papillary thyroid carcinoma

1.7.1 The 1986 Chernobyl Nuclear Disaster

The Chernobyl Nuclear Disaster is one of the largest releases of radiation into the environment to date. During a planned test at roughly 01:23 Moscow Time on April 26, 1986, two explosions occurred in the number 4 reactor of the Chernobyl Nuclear Power Plant. These explosions led to dispersal of radioactive fuel, fuel byproducts, and irradiated graphite control rods into the immediate area, as well as structural damage and fires in the plant complex. Hundreds of firefighting personnel were sent in and many were exposed to high levels of radiation. A subsequent graphite fire released significant

amounts of radionuclides, particularly ^{131}I, ^{90}Sr, and ^{137}Cs. This fire would burn for two weeks before being extinguished, during which time it would continue to release radionuclides into the atmosphere.

The disaster had lasting impacts across a wide geographic region. Radionuclides were spread throughout the Northern Hemisphere including large portions of Europe [156, 157]. The triggering of radiation detectors as far away as Sweden, Norway and the United Kingdom was the first warning many nations received that a nuclear release had occurred. One hundred and thirty four first responders received sufficiently high doses of radiation to display symptoms of acute radiation sickness [158]. Approximately thirty people were killed immediately due to the initial explosions or from acute radiation poisoning.

Within four years a major increase was seen in the number of adolescent thyroid cancer cases in areas of Belarus, Russia, and Ukraine, particularly in the areas directly north of the Chernobyl site [159, 160]. While the short half-life of ^{131}I prevents direct measurement of where it was deposited, estimates can be gleaned from measurements of ^{137}Cs, which has a much longer half-life. High exposure to ^{131}I is strongly associated with subsequent risk for thyroid cancer [161]. The total number of thyroid cancer cases reported in the affected areas among people who were adolescents during the incident is roughly 5,000. It is estimated that the Chernobyl incident likely contributed to an excess of up to 4,000 - 10,000 cancer-releated deaths in those exposed, though the exact numbers are difficult to precisely estimate and political controversies over these numbers continue over thirty years later.

The Chernobyl Tissue Bank, an international program, was established to preserve and catalog blood and tissue from individuals that were exposed to radionuclides from the disaster. The Chernobyl Tissue Bank holds roughly 5,000 samples from individuals aged less than 19 years of age at the time of the incident who later developed thyroid cancer. These individuals come from either of two hospitals in Ukraine or the Russian Federation. Some combination of tumor and normal thyroid tissue (either fresh-frozen and formalin-fixed paraffin-embedded) and normal blood are available for each individual. All samples are reviewed by an international team of expert pathologists.

While ^{131}I has a half-life of only 8.4 days, the half-lives of ^{137}Cs and ^{90}Sr are both approximately thirty years. Areas impacted by fallout of ^{137}Cs and ^{90}Sr will remain measurably contaminated for centuries. Agricultural countermeasures are still required to avoid introduction of radioactive products into the food supply [158]. While there is no evidence for an increase in solid cancer prevalence (with the exception of thyroid

cancers), there is evidence that rates of leukemia may be elevated in cleanup workers. While direct health effects have impacted perhaps tens of thousands of individuals, the total human costs of the disaster have been much greater. Approximately 300,000 people were evacuated permanently from the area surrounding Chernobyl and Pripyat and a 35 kilometer exclusion zone now exists around the plant. Millions experienced psychological trauma as a result of the disaster, and depression and suicide rates for those affected have remained significant [162].

The international response to Chernobyl has provided invaluable information regarding radiation dosimetry, management of contaminated areas, and the health effects of environmental radiation. The response to the 2011 Fukushima nuclear accident, including the use of caesium binders in livestock feed, was informed by the international research collaborations built after the Chernobyl disaster [158, 157]. Chernobyl has also provided insights into the long-term effects of exposure from nuclear testing, where accurate dosimetry is not always available [163]. Continued studies of survivors and the valuable resources of the Chernobyl Tissue Bank will provide essential knowledge for responding to any future disasters.

1.7.2 Thyroid cancer

The most notable public health outcome of the Chernobyl disaster was a dramatic increase in thyroid cancer rates among adolescents in irradiated areas. Thyroid cancer accounts for 1-2% of all cancer cases each year. Though this varies significantly between sexes and continents, in general the rate of occurrence in women is roughly three times that in men.

In contrast to most cancer types, the incidence of thyroid cancer has been increasing over time [164, 165]. The total number of thyroid cancer cases diagnosed per year has doubled since 1970 [164]. This increase remains significant even when accounting for improved detection methods. Thyroid neoplasms tend to be diagnosed earlier than most cancers, with a median age at diagnosis is 46 years of age compared to 67 years for all cancers [161]. Overall mortality from thyroid cancer has declined during this period, and survival rates exceed 95% for follicular and papillary subtypes [161].

Biological sex is among the only factors that appears to explain major differences in thyroid cancer rates [161]. Thyroid cancer is roughly three-times more prevalent in women than in men, though the reason for this difference is not well understood. Thyroid neoplasms represent roughly 7% of cancer cases in women in Latin America each year but only 1-2% of cancer cases in men [165].

There are multiple thyroid cancer subtypes that display histologic and epidemiologic differences. The most common subtype is papillary thyroid carcinoma (PTC), which accounts for anywhere from 60 to 90% of cases. Prognosis for the papillary subtype is good with five-year survival rates over 95%. The follicular subtype is composed of well-differentiated cells (like papillary cases) but prognosis is slightly worse [161, 166]. Follicular cases account for roughly 10-15% of diagnoses. Medullary tumors account for 5-10% of tumors. Approximately 25% of medullary cases display familial inheritance due to mutation in the $MEN2$ gene. Prognosis for the medullary subtype is again quite good, with survival rates in the range of 80%. Anaplastic subtypes account for the remainder of cases (1-2%) and are associated with significantly worse prognosis compared to papillary cases, with many patients not surviving more than six months after diagnosis. All non-papillary cases are associated with greater age at diagnosis. Overall, thyroid cancer survival rates are higher for any other cancer type except non-melanoma skin cancer, largely due to the high number of papillary and follicular type tumors compared to the rarer, more aggressive types [165].

Exposure to radiation is a well-established risk factor for thyroid cancer development. Exposure in adolescence presents a greater risk of cancer development than during adulthood [167]. Iodine deficiency also predisposes one to a greater risk of thyroid cancer [168]. This is primarily for the follicular and anaplastic subtypes in sporadic cases, however, and not for the more common papillary subtype. It has been hypothesized that sufficient dietary iodine can reduce uptake of radioactive iodine isotopes. Ron *et al.* found that iodine supplementation with potassium iodide was associated with a three-fold reduction in risk of thyroid cancer at one gray of exposure to ^{131}I and that iodine deficiency was associated with a three-fold increse in risk for the same exposure. Iodine supplementation is often used as prophylaxis in populations at risk due to environmental exposure.

1.7.3 Ionizing radiation and its effect on the genome

Ionizing radiation is a powerful carcinogen [169]. High-energy particles can induce both single- and double-stranded breaks in the DNA, and radiation is known to act as both a mutagen and a clastogen [170, 81].

The impact of exposure to radiation is dependent on its source [169]. As an ionized particle passes through tissue it transfers energy. The amount of energy transferred per unit distance is known as the *Linear Energy Transfer* (LET). X rays and gamma rays have low LET while alpha particles are high LET; beta emissions fall somewhere between

these. The *quality* of a radioactive source is the factor by which an exposure must be multiplied to produce a normalized measure of biological damage, effectively converting an absorbed dose in gray to one of biological risk in Sieverts. *Robustness* is a measure of cells' sensitivity to exposure from a particular radiation source.

Though a full discussion of radiation quality and effective dose is beyond the scope of this thesis, it is important to consider that the samples in the Chernobyl Tissue Bank received very different exposures than tumors receiving medical radiation. These doses were received primarily from deposited radionuclides through consumption of contaminated vegetables or milk. In the case of ^{131}I, most of the energy transferred is through beta emissions. Medical doses may come from X-ray, gamma ray, alpha or beta sources. Total doses for medical regimens may be as high as 30 Gy, although this dose is usually fractionated to minimize damage to cells [171]. Previous studies have most often focused on exposure of cell lines or medical samples to gamma, X-ray, or alpha particles. Direct comparison of studies of medical radiation and those of individuals exposed from environmental radiation should be done with care.

Cornforth *et al.* provide a lengthy review of their own published research on fibroblasts and lymphocytes exposed to different radiation sources in [81]. Cytogenetic profiling by mFISH was used to look for chromosomal aberrations and the number and types of aberrations were used as a measure of dose response. It is well-established that such cytogenetic abnormalities are markers of radiation exposure [172]. Overall, all types of radiation examined (heavy ion, alpha particles, X rays, and gamma rays) produced a linear or curvilinear response to exposure.

Large inversions and complex translocations are markers of radiation exposure in cytogenetics [81, 172, 66]. These changes are stable across cellular generations and can be used for biodosimetry. In general, the ratio of complex interchanges (i.e. those involving three or more breakpoints) to simple ones is highly dose-dependent for low-LET radiation [81]. No such relationship exists for high-LET radiation. Workers from nuclear weapons plants harbor long-term signatures of exposure in their cytogenetic profiles [80]. Hande *et al.* found a strong association between dose received and the number of *intra*chromosomal rearrangements in former radiation workers from the Soviet Union. These balanced events are signs of induced double strand breaks [170, 54]. Cytogenetics-based methods have been shown to be sensitive to exposures as low as 0.3Gy [172] and stable cytogenetic markers of exposure persist for decades [80].

Cytogenetics has long been the gold standard of biodosimetry metrics and chromosomal events visible by mFISH are clearly associated with radiation exposure. The

resolution of such techniques is much lower than that of next-generation sequencing, however. [54] examined twelve radiation-associated tumors from individuals exposed to medical radiation. These tumors were whole-genome sequenced and compared to 319 radiation-naive tumors processed through the same pipeline.

Radiation-exposed samples harbored a greater number of small deletions compared to radiation-naive tumors, consistent with non-homologous end joining creating erroneous DNA repair products. These deletions were small ($\leq$100 bp), well below the detection threshold of cytogenetic techniques. Deletions were distributed randomly across the genome, whereas deletions caused by other known carcinogens are often in distinct regions or motifs. In addition, radiation-exposed tumors were significantly enriched for balanced inversions. These mutations are relatively rare in cancer samples not exposed to radiation. Large inversions like those found in [54] are consistent with reports of intrachromosomal translocations and inversions in radiation-exposed cells using mFISH [66, 172, 80].

Behjati *et al.* demonstrated that next-generation sequencing of radiation exposed tumors could detect two genomic markers of radiation exposure. The study was limited to only twelve radiation-exposed tumors. In this thesis I analyze approximately 300 radiation-exposed samples, though at much lower doses than those of Behjati *et al.* While still insufficient to detect novel radiation sensitivity mutations in the germline this number is sufficient to obtain valid calculations of significance. Results and discussion of this analysis follow in Chapter 2.

1.8 Structure of the remainder of this thesis

In Chapter 2, I describe a cloud-based analysis of over 1,000 whole-genome samples from the Chernobyl Tissue Bank. These were analyzed as somatic pairs, either tumor - normal blood or tumor - normal thyroid tissue; a total of 393 tumor - normal pairs were initially included. Normal blood was chosen over normal thyroid tissue as a germline background when available. These pairs were analyzed for somatic SNVs, indels, structural variants, and telomere length, as well as small variants in the germline. These results were examined in the context of phenotype data including radiation dosimetry estimates, as well as in comparison to cases of sporadic thyroid cancer.

In Chapter 3, I discuss variation graphs and structural variation.

Chapter 4 is a description of algorithms for working with long reads from genomic mixtures.

Chapter 2

Genomic characterization of radiation-associated papillary thyroid carcinoma

2.1 Introduction

After the 1986 Chernobyl nuclear disaster a pronounced rise in pediatric thyroid cancer cases was observed in regions north of the site. A concerted international effort was made to catalogue blood and tissues from individuals presenting with thyroid cancers in these regions. This resource, called the Chernobyl Tissue Bank, now holds approximately 4,500 samples from Ukraine and Russia. A portion of these samples come from individuals that developed pediatric thyroid cancer but who were not exposed to radiation, allowing comparison between cases and controls which are age- and population-matched.

Over the course of three years we received 1,307 samples from the Chernobyl Tissue Bank (Table 2.1). These samples came from indivduals which had a thyroid tumor that had been labeled by pathology as papillary thyroid carcinoma (PTC). In addition to pathology data we received phenotypic data on sex, age at exposure, age at surgery, dosimetry estimates, and geographical information for each individual.

To be included in our study we required that individuals have a thyroid tumor sample and at least one normal tissue sample (either blood or non-tumor thyroid) which was whole genome sequenced and passed stringent quality controls. In total, 998 samples met these criteria from a total of 381 individual patients. However, 21 of these samples had tumors with low purity ($\leq$20%) or tumor-normal pairs that appeared flipped (i.e. where

Tissue Type	Received	WGS Sequenced	Passing QC
Thyroid tumor	491	400	381
Normal Blood	331	252	252
Normal Thyroid	483	382	365
Metastasis	2	2	0
Total	**1307**	**1036**	**998**

Table 2.1 Counts of samples delivered from the Chernobyl Tissue Bank. 1,036 samples in total were sent for whole genome sequencing. Of these, 998 passed quality control metrics at the sample level and were used in the analytic set.

normal thyroid tissue showed evidence of malignancy) based on somatic analysis. These samples are excluded from some analyses where the inability to detect their full mutation spectrum might skew results. The distribution of samples available per individual is shown in Figure 2.1 (subsection 2.1.2).

In addition to the somatic pairs from the Chernobyl Tissue Bank we received fifty sporadic whole genome sequenced PTC cases from The Cancer Genome Atlas THCA study [53]. These were sequenced on a platform with slightly shorter reads and lower depth. All fifty of these samples were included in our analysis.

I examined the mutational landscape of these samples, including somatic single-nucleotide variants, indels, and structural variants; germline small variants; and mutational signatures. I compared the results between exposed cases and unexposed cases from both this study and [53]. Results from whole genome sequencing analysis were corroborated with SNP array and RNA data where possible. While data from a variety of assays was available this work relies primarily on analysis of the whole genome sequencing data.

2.1.1 Collaboration and publication notes

The work in this chapter was performed in collaboration with others. The Chernobyl Tissue Bank provided samples and phenotype data. Nationwide Children's Hospital provided whole genome sequencing of our samples. The Cancer Genome Research Laboratory at the National Cancer Institute provided some RNA sequencing as well as nucleotide array data, methylation array data, and a relative telomere length assay. I designed the germline and somatic calling pipelines with Dr. Chip Stewart and Dr. Gad Getz of The Broad Institute of Harvard and MIT, largely based on their work in [53]. Jaegil Kim from the Broad Institute helped devise early versions of the mutational

signature analyses. I worked with Dr. Danielle Karyadi and Dr. Stephen Hartley at NCI on manual quality control and variant annotation. Dr. Karyadi performed extensive manual review of the small variants as well. Dr. Mia Steinberg, Dr. Jieqiong Dai, and Dr. Joe Boland performed the mRNA- and miRNA-Seq analyses, which I used as validation for the results in WGS. Dr. Lindsay Morton was essential to study design, quality control, and statistical analyses. The work is currently being prepared for submission. I have noted where individual analyses or tools have been made available publicly.

2.1.2 Quality control and study design

The tissue samples in the Chernobyl Tissue Bank are a unique and precious resource and radiation-related incidents attract significant amounts of public attention. As an international collaboration across multiple centers there were also major logistical challenges in managing data and sample transfer. Due to these factors and possible regulatory implications of our results, I implemented stringent quality control pipelines to ensure the integrity of our data. This ensured that genomic data matched the phenotype data across all platforms analyzed. This was corroborated with extensive manual review.

We received some combination of whole genome sequencing, mRNA and miRNA sequencing, an Illumina OmniX SNP array, an Illumina EPIC methylation array, and a relative telomere length assay data for each sample. For each assay, I coordinated the individual laboratory information management system identifiers with the sample code from the sequencing center and the identifier from the Chernobyl tissue bank. This was used to generate a study-unique identifier for each sample on each assay. This global identifier facilitated cross-assay comparison later on.

Each sample was quality checked individually on each assay. QC thresholds were established by exploratory data analysis and then applied programatically. When a sample failed a scripted QC threshold we only removed it from the study after manually reviewing and verifying any inconsistencies.

Whole-genome data was checked for contamination, purity, ploidy, sex concordance (both between samples and with phenotype data), depth of sequencing, and estimated driver detection sensitivity. We required a median depth of 60X in the tumor sample and 20X in the normal. Most of our samples were sequenced to >90X for tumors and >40X for normal blood and normal thyroid tissues.

The majority of samples in our dataset had a ploidy of exactly two. I examined the WGS copy number aberration (CNA) plots of any samples that received a ploidy of less

than 1.99 or greater than 2.01 according to `allelicCAPSEG` [173]. All of these samples passed review.

We examined WGS contamination using `verifybamid` [174], `GATK` version 4.0 [175], and ContEst [176]. Any samples with contamination above 2% were flagged and manually reviewed. All samples flagged as $\geq 2\%$ contaminated were subsequently removed from analysis.

We examined WGS tumor purity using `ABSOLUTE` [173] and an allele-fraction (AF) based metric (developed by Dr. Stewart and Dr. Karyadi). `ABSOLUTE` relies on somatic copy number alterations to determine purity. In many of our thyroid tumors there was little or no SCNA activity. While purity estimates from `ABSOLUTE` and the AF purity metric were highly correlated, in the cases where the two methods disagreed the AF purity estimate provided a more accurate estimate of purity. These metrics were also compared against tumor purity estimates from the CTB pathology board.

Based on the AF purity, I calculated the approximate heterozygous mutation calling sensitivity for a given tumor-normal pair. I defined sensitivity to be the probability of observing more than three reads from the tumor which support a heterozygous alternate variant, assuming a binomial distribution. Where necessary, individual mutations were examined to give a manual estimate of tumor purity based on allele frequency. Any tumor with less than 80% sensitivity (corresponding to approximately 20% purity) was flagged during downstream analysis.

We also examined normal thyroid tissue samples for signs of clonal expansion or secondary tumors by running a flipped somatic pipeline where normal tissue samples were called against the somatic background of the tumor. The presence of somatic mutations in the normal thyroid tissue that are not present in the tumor would indicate clonal expansion of the tumor or a second malignant focus in the normal tissue. This could impact downstream variant calling, reducing the apparent mutational burden and our ability to detect driver mutations. Any normal tissue samples which were labeled by pathology review as containing tumor foci or which contained >50 somatic mutations compared to the tumor were removed from analysis.

Sex concordance was assayed by calculating the ratio of reads covering the Y chromosome and those covering the X chromosome. The software package for calculating this metric is available at https://github.com/edawson/check-sex. Sex-chromosome read ratios were plotted against the labeled sex from phenotype data to establish cutoffs. In addition, sex estimates from whole genome sequencing were compared to those from the OmniX array and phenotype data from the Chernobyl Tissue Bank. Samples which

were labeled with a particular sex but which did not cluster with that sex were reviewed manually, and any found to be discordant were removed from subsequent analysis.

Samples were initially screened to have come from the same individual by comparing OmniX results. Many samples were not sent for sequencing as it could not be ascertained that the samples were actually from a single individual. Samples were further fingerprinted across WGS, OmniX, and EPIC methylation arrays using shared sites across the three platforms. WGS BAM files from tumor and matched normal(s) were fingerprinted using BAM-matcher [177]. Any samples which were not reported with high confidence as coming from the same individual were reviewed. Several samples were flagged as discordant between tumor and normal due to significant copy number alterations but were deemed to be from the same individual upon manual review (see subsection 2.6.4).

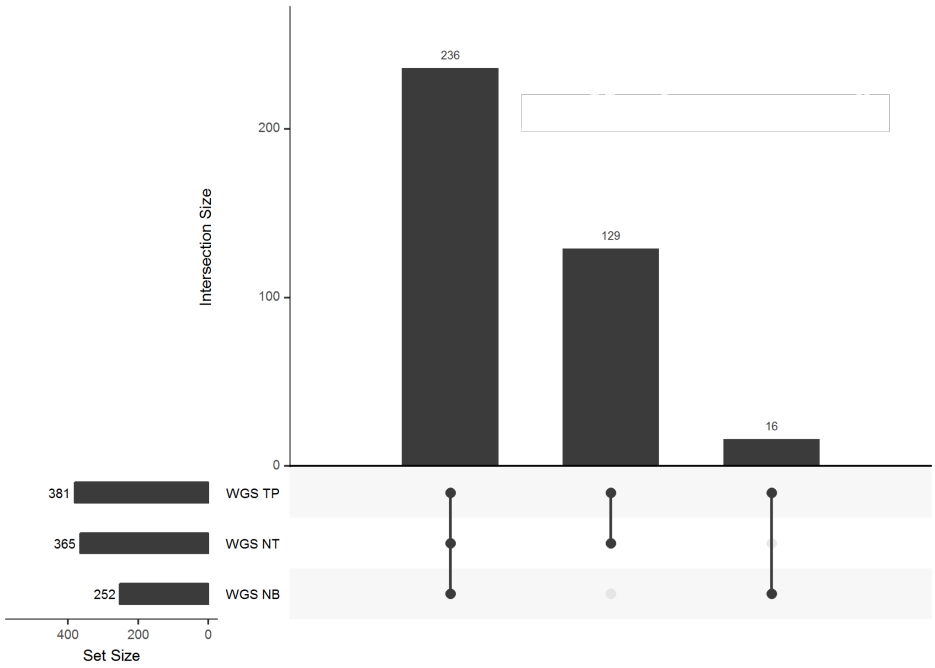


Fig. 2.1 An upset plot of the data available for 381 patients with a tumor and at least one matched normal. The majority (236 of 381) of individuals remaining after quality control have both normal blood and normal thyroid tissue. One hundred and twenty nine have only normal thyroid tissue; sixteen have only normal blood.

After sample-level QC, somatic pairs were matched to assess which individuals had samples available for study. An upset plot of available samples per somatic pair is available in Figure 2.1. For individuals with two normal samples I chose normal blood over normal thyroid tissue for somatic analysis. The secondary normal was used for validation and for analyses not presented in this work. Any individual without a tumor

or without a matched normal that was not considered high-quality was removed from the analytic set.

2.1.3 Phenotypic overview of analyzed somatic pairs

In total, 381 somatic pairs met criteria for inclusion. Among the 381 individuals available from the Chernobyl Tissue Bank dataset, 303 received a non-zero estimated raditation dose. Seventy-eight individuals were born at least one year after the Chernobyl nuclear event and are estimated to have received no radiation exposure. I refer to the full sample set of 381 patients as REBC (Radiation Epidemiology Branch - Chernobyl). Unexposed cases are labeled "REBC-Unexposed," whereas "REBC-Exposed" is used to distinguish samples which received a non-zero radiation dose. The fifty matched whole genome sequenced cases from [53] are collectively labeled "THCA" (the name of the TCGA Thyroid Carcinoma study which produced them). Counts of samples in each sample set are shown in Figure 2.2.

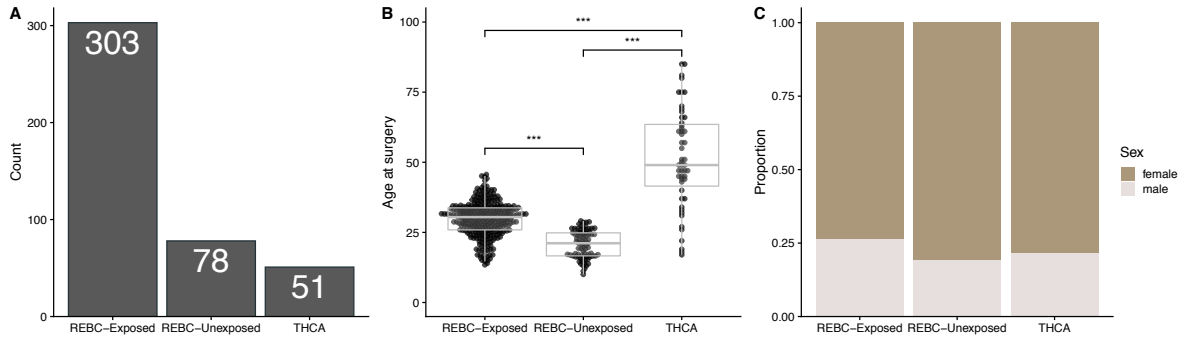


Fig. 2.2 Summary data for phenotype variables in our dataset. (A) Counts of somatic from each of the REBC-Exposed, REBC-Unexposed, and THCA sample sets. (B) Age at surgery for each of the sample sets. There is a significant difference in age between any pairwise combination of the three sets ($p < 0.01$). (C) The proportion of individuals of each sex for the sample sets.

Chernobyl-associated PTC cases are significantly younger than THCA cases

Thyroid cancer risk peaks in the fifth decade of life, with most cases occuring between age 40 and 60 [161, 178]. The THCA cases in our dataset fit within this age range (mean = 51.9, median = 49, range = [17, 85]). Both the unexposed and exposed portions of the REBC sample set differ significantly from the THCA sample set in mean age at onset ($p < 0.01$, Figure 2.2B). Exposed cases have a mean age of onset of 29 (median 30, range = [13, 45]); unexposed cases have a mean of 20 (median = 21, range = [10, 29]). These age

Dose Range (mGy)	Number of samples
$0 <$ Dose < 200 mGy	235
$200 \leq$ Dose < 500 mGy	37
$500 \leq$ Dose < 1000 mGy	14
1000 mGy $\leq$ Dose	17

Table 2.2 Summary of estimated radiation doses for exposed cases.

differences are again significant ($p < 0.01$, Figure 2.2B). Statistically, there is a ceiling to attained age since sampling for the unexposed individuals was stopped at a fixed timepoint roughly thirty years after the accident. Some of these may also be sporadic cases that were caught early due to enhanced screening after the Chernobyl disaster. These factors may explain some of the differences in age between the three sample sets.

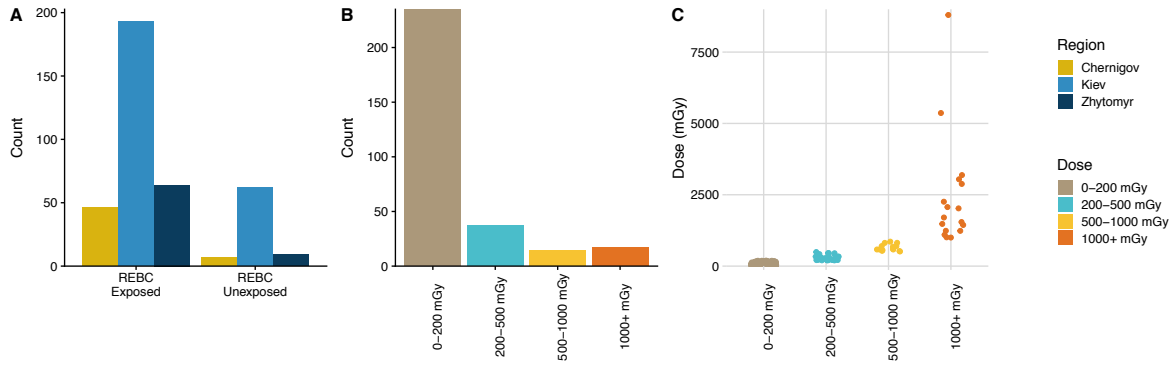


Fig. 2.3 (A) Geographic region of samples in the REBC-Exposed and REBC-Unexposed sample sets. (B) Number of exposed samples within each exposure category. (C) Summary of doses within each dose category.

For other phenotypic variables we see no significant difference among the three datasets. REBC-Exposed, REBC-Unexposed, and THCA show identical sex distributions with roughly three times as many female cases as male ones (Figure 2.2C). This difference in rates between the two sexes is well established [161]. The two REBC sample sets are also drawn from the same geographical regions at roughly the same rates (Figure 2.3A).

Seventy-five percent of our exposed samples received an estimated dose of less than 200mGy (Figure 2.3B and C, Table 2.2). This is significantly less than medical doses like those seen in [54], which can approach 30Gy of total absorbed dose.

All analysis for this study was performed in the FireCloud, a cloud-based platform hosted by the Broad Institute. DNA was extracted at Nationwide Children's Hospital and sequenced at the Broad Institute on the Illumina X10 platform. Sorted and indexed

BAM files of aligned reads were provided by the Broad GDAC team through FireCloud. FireCloud uses a workflow description language called WDL to manage analyses [179]. All WDL scripts used in this text which could be made public are available on FireCloud and at GitHub.

For both small variants and structural variants we implemented consensus-calling pipelines that used the intersection of multiple tools to improve sensitivity and specificity. This has been shown previously to improve variant calling results over the use of a single caller [180, 31].

2.2 Simple somatic variants

I analyzed somatic SNV and indel variants (collectively simple somatic variants or SSVs) using two linked consensus pipelines. Single nucleotide variants were called using `Mutect1` [181], `Mutect2` (via `GATK` version 3.8) [175], `Strelka` [182] and `Strelka2` [183]. Indel variants were called using `Mutect2`, `Strelka1`, `Strelka2`, `SvABA` [87] and `Snowman`, an older version of the `SvABA` algorithm that was used in the Pancancer Analysis of Whole Genomes [184]. `SvABA` and `Snowman` results were pulled from the structural variant calling pipeline. The remaining algorithms were part of the SSV pipeline. The outputs of these variant callers were converted from VCF into a common tab-separated format. This file was then annotated by `Oncotator` [185] to convert this to a MAF-like format which includes information such as gene name, variant context, and variant impact. These calls were then annotated with two panels of normal samples, one from this study and one from the Pancancer Analysis of Whole Genomes [31].

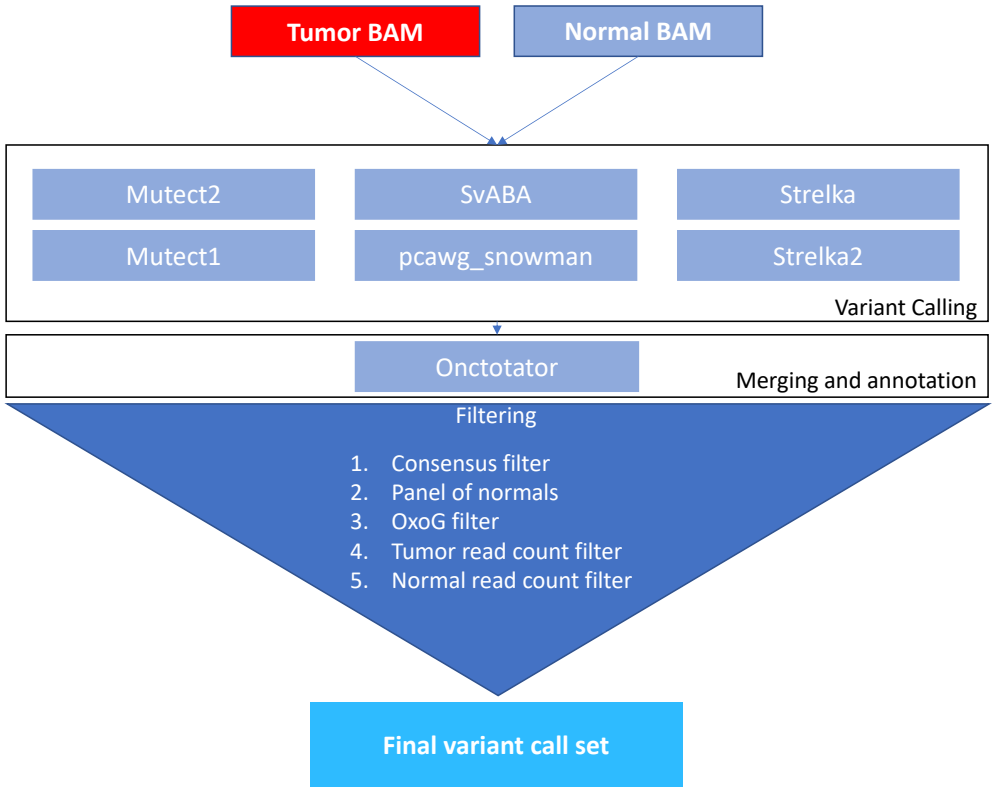


Fig. 2.4 Diagram of somatic calling pipeline.

This pipeline produced a union of roughly four million raw SSV calls among the four callers. These were first filtered using a consensus voting scheme. Each variant caller received one vote; however, callers using the same algorithm (i.e. Strelka1/2 and SvABA/Snowman) could contribute a maximum of one vote between the two of them. I required an SSV to receive at least two votes to remain in the set. Any SSV seen in either the panel of normals from this study or from [31] was removed. Any variant seen in dbSNP was labeled but not removed. Variants annotated as OxoG artifacts were removed [186]. Lastly, calls which relied on fewer than 5 tumor alt-supporting reads or where the normal sample had two or more alt-supporting reads were removed.

2.2.1 REBC PTC cases carry a comparatively low mutational burden

Across the entire study, there were approximately 350,000 SSVs after filtering (per sample mean = 920, median = 742, range = [12, 49910]). One sample exhibited a hypermutator genotype, with roughly ten times the number of calls compared to the next highest tumor

(shown in Figure 2.5A, but not within scale in Figure 2.5B or C.). Several tumors also exhibit a very small number of variant calls (<100); many of these were flagged during QC as low-purity tumor samples. The hypermutator sample was analyzed separately where necessary to prevent distortion of overall statistics. Any tumors with purity less than 20% were excluded from the calculation of summary statistics as well. With these samples removed, the median SSV count per sample was 757 (mean = 821, range = [147, 3501]).

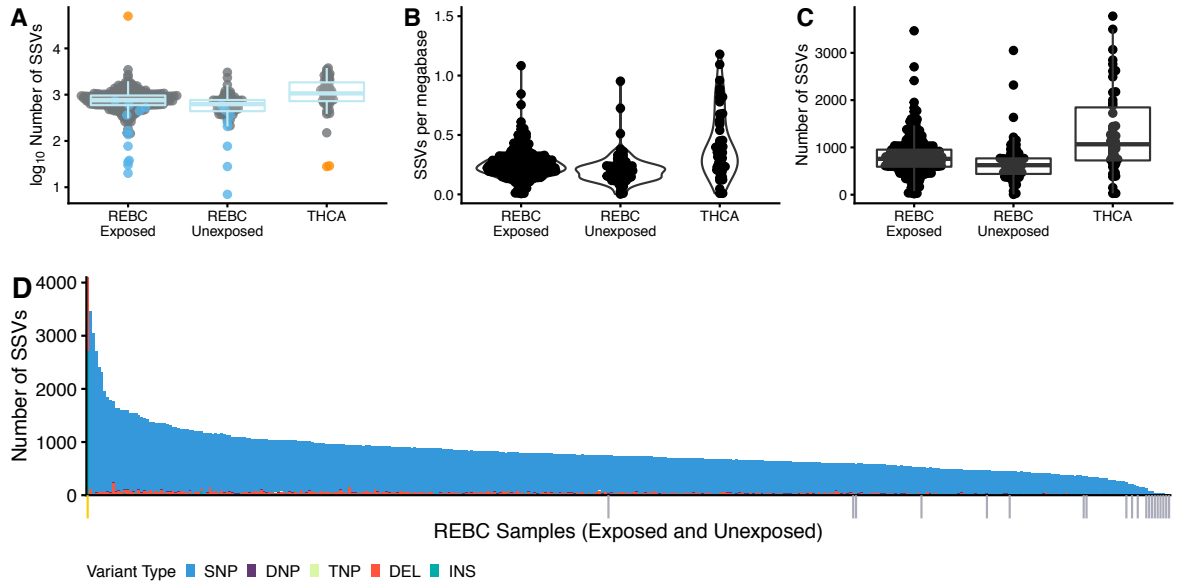


Fig. 2.5 Overview of SSV counts for REBC samples. (A) The $\log_{10}$ mutation counts for all 381 samples, with samples with more than 4,000 or fewer than 100 SSVs highlighted in orange. Samples with low tumor purity are in light blue. (B) The number of SSVs per megabase, excluding one SSV hypermutator sample. (C) The total SSV count per sample. (D) The SSV count, per sample, broken down by variant class. Samples with low purity are marked with a grey tick. The hypermutator sample exceeds the plot scale and marked with a goldenrod tick.

The median number of mutations per megabase across all 360 REBC samples with purity above 20% (and excluding the hypermutator) was 0.25. This was lower than in the THCA sample set, which had a median of 0.35 mutations per megabase (per sample mean SSV count = 1128, median = 1367, range = [54, 3832]). The REBC sample set has among the lowest coding SSV counts per sample of any cancer type in The Cancer Genome Atlas (Figure 2.6) [187]. However, the number of SSVs per sample is of comparable magnitude to other thyroid studies [44, 53].

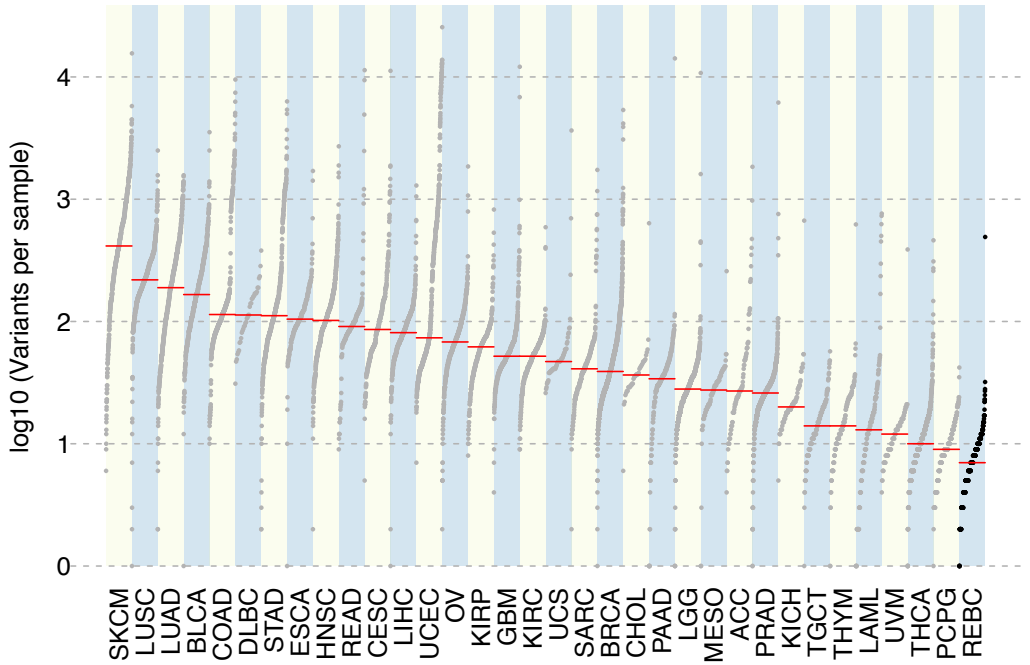


Fig. 2.6 A comparison of coding SSV counts between REBC and TCGA sample sets.

2.2.2 Somatic SSV counts increase with age

A low mutational burden is somewhat expected given that thyroid cancer has an age at onset between that of pediatric and adult cancers. Pediatric cancers tend to have mutation rates of 0.01 to 0.5 mutations per megabase [45]. The mutation counts in REBC fall within this range.

There is a significant correlation between age and SSV burden across all of our samples ($p < 0.01$, $r^2 = 0.51$, Figure 2.7). It is well established that mutations accumulate with age in both cancerous and normal cells [34]. I attempted to correct for the effects of age by fitting a linear model relating SSV counts to age at surgery across all three sample sets and then examining the residuals between the observed and expected SSV counts. The distribution of residuals is plotted against age at surgery in Figure 2.8. The randomness of the residuals and the consistency of their medians across sample sets indicates that much of the difference in SSV count between groups can be explained by differences in the age of individuals within each sample set.

No difference in SSV burden was seen between male and female cases (Kolmogorov-Smirnov test p-value = 0.5). Nor was the number of SSVs signficantly correlated with

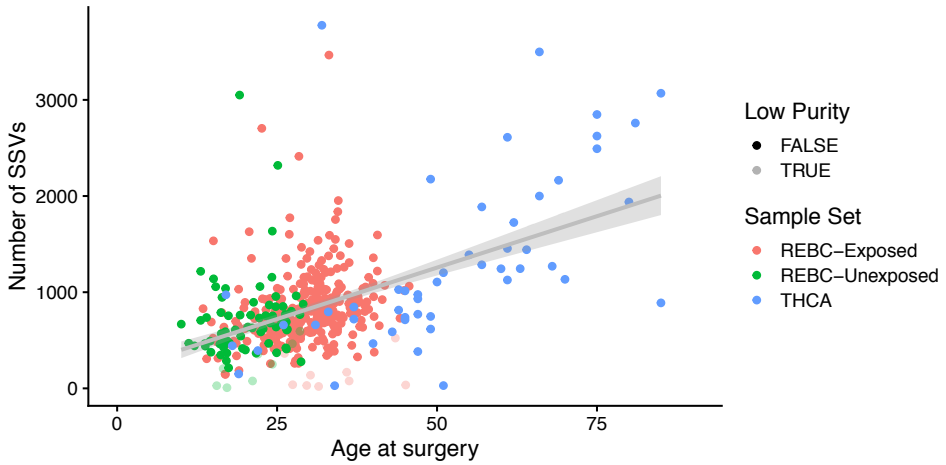


Fig. 2.7 The number of SSVs generally increases with age across all three sample sets. Samples flagged for low purity tend to have lower SSV counts regardless of which sample set they belong to.

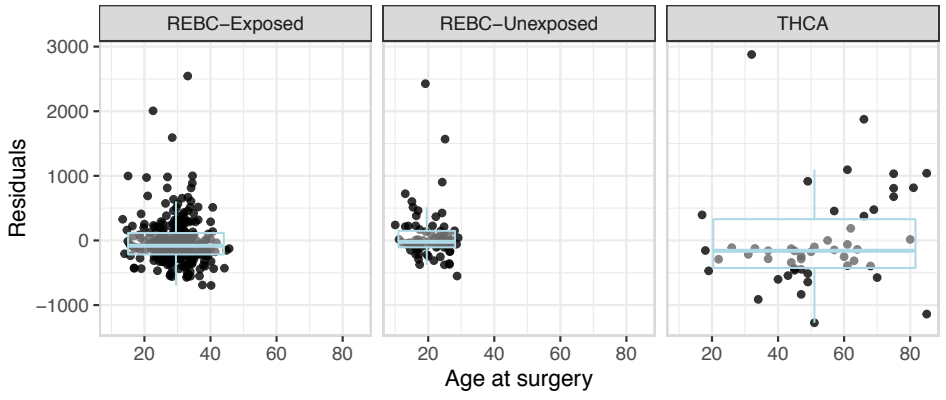


Fig. 2.8 Residuals of predicted and observed SSV counts plotted against age at surgery.

continuous radiation dose (Pearson's test, p = 0.14). When examining dose categorically, a significant difference in mean SSV count was observed between dose groups (one-way ANOVA, $p = 0.006$, Figure 2.9). However, this is likely due to the large differences in dose group size and having few samples ($\leq$20) in most of the dose groups. The medians of the groups are approximately the same (Figure 2.9). For a linear model predicting the number of mutations based on age at surgery, dose category, and sex, only age at surgery was a significant predictor ($p < 0.001$, adjusted r^2=0.247). The same was true of a model which used the same age and sex variables but continuous dose rather than dose category ($p < 0.001$, r^2=0.246).

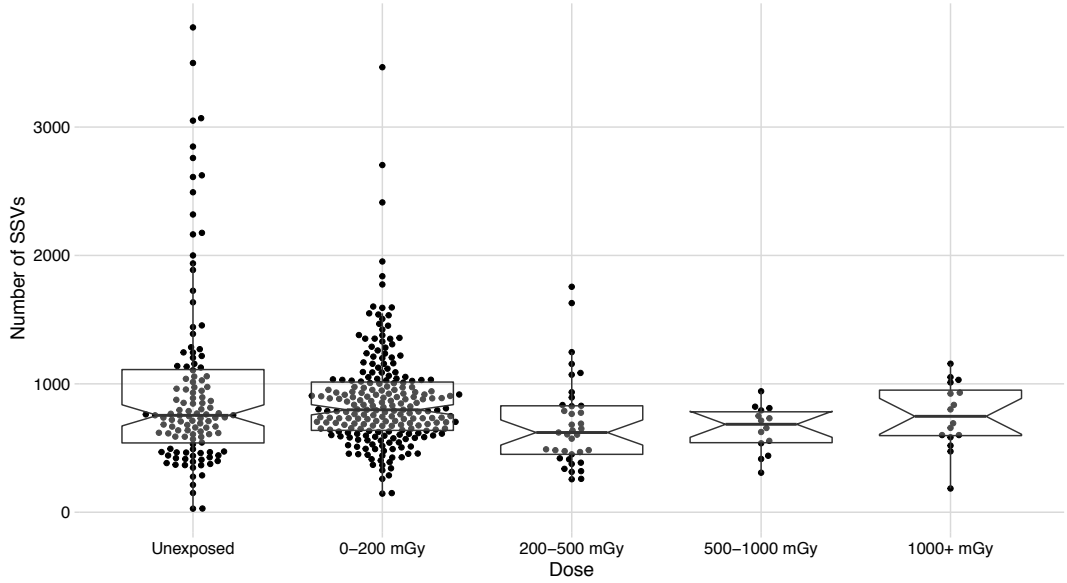


Fig. 2.9 Number of SSVs plotted per dose group.

Overall, the substitution SSV mutation spectrum of our dataset is consistent with those of previous studies on sporadic thyroid cancer. Thyroid cancer in general carries a relatively small number of somatic SSVs and more closely resembles pediatric cancers than those of adulthood in its SSV burden. There is a significant positive correlation between SSV burden and age at surgery. No other phenotypic factors we tested were associated with increased SSV burden. In addition, SSV hypermutation is rare, with only a single sample exhibiting an order of magnitude more SSVs than the median count per sample.

2.2.3 MAPK genes are significantly mutated in papillary thyroid carcinoma

Next, I applied several methods to look for significantly mutated genes across our REBC data sets. MutSig2CV [188] is designed to detect genes which have more mutations than expected by chance compared to a null model based on background mutation rates of nearby genes.

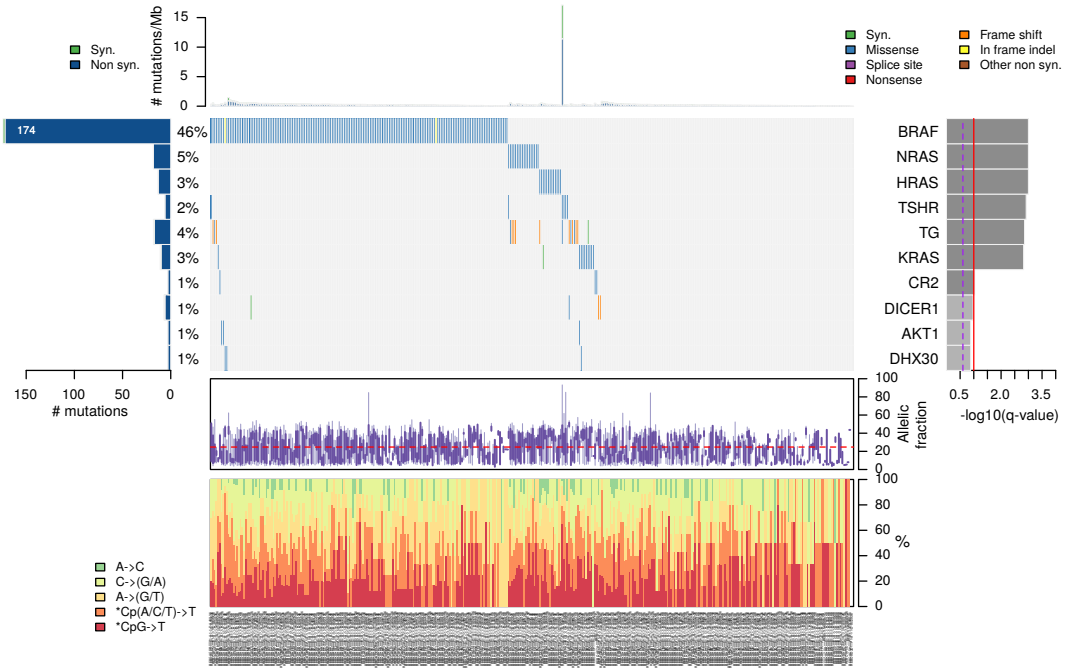


Fig. 2.10 Significantly mutated genes from MutSig2CV.

MutSig2CV reported nine significantly-mutated genes in the REBC samples. Four of the nine genes are part of the *MAPK* pathway, which is mutated in the majority of adult sporadic thyroid cancers, especially the papillary subtype [53, 189]. *BRAF* was the most commonly mutated gene. Most of the *BRAF* mutations in our dataset are the common $BRAF^{V600E}$ thyroid and pan-cancer driver mutation. 140 / 291 (47%) of exposed samples harbored a BRAF mutation, of which 134 (97%) were $BRAF^{V600E}$. Three tumors had $BRAF^{K601E}$ mutations. One tumor had an in-frame insertion at position 598 ($BRAF^{598\_599insIDFGLA}$); one other tumor had an in-frame deletion at $BRAF^{600\_601VK>E}$. All of these alternative mutations likely have similar activating effects to $BRAF^{V600E}$.

Protein Change	Sample Set	Number
$p.V600E$	REBC-Exposed	135
$p.V600E$	REBC-Unexposed	32
$p.V600E$	THCA	10
$p.K601E$	REBC-Exposed	3
$p.600\_601VK > E$	REBC-Exposed	1
$p.598\_599insIDFGLA$	REBC-Exposed	1

Table 2.3 Overview of BRAF mutations in the three sample sets.

Three *RAS* genes ($NRAS$, $HRAS$, and $KRAS$) were the next most frequent significantly mutated genes reported by MutSig2CV. These *MAPK* pathway genes are well-established thyroid cancer drivers [53]. Mutations in the *RAS* genes are almost always mutually-exclusive with *BRAF* mutations. We observed this pattern in our samples, with only one case harboring a mutation in $BRAF^{K601E}$ and $KRAS^{G12D}$. Mutations in different *RAS* genes were mutually exclusive. Across sample sets, the individual proportions of $HRAS$, $KRAS$, and $NRAS$ mutations were similar.

2.2.4 Thyroid genes are frequently mutated in papillary thyroid carcinoma

Two genes essential for thyroid function —thyroglobulin (TG) and thyroid stimulating hormone recepter ($TSHR$) —were reported as significantly mutated across sample sets by MutSig. While both genes were mutated in a sufficient number of samples to reach study-wide significance, the mutations observed in TG appear to be the result of lineage-specific mutational patterns. TG is the most highly expressed gene in follicular thyroid cells and the TG protein is essential for production of thyroid hormones [190].

We found a high number of non-coding intronic indel mutations in TG, consistent with a lineage-specific hypermutation pattern described by [191]. The ratio of noncoding to coding mutations in TG was approximately 17:1, while the same ratio in well-established PTC drivers ($BRAF$, $NRAS$, $HRAS$ and $KRAS$) ranged from 1:2 to 1:20. The majority of mutations in established drivers were primarily missense substitutions rather than indels. While [191] reported this pattern in lineage-specific genes from multiple cancers (including TG in 20-40% of thyroid tumors), they could not definitively conclude whether the high frequency of these mutations was due to positive selection or a novel lineage-specific mutational process which may be acting on highly-transcribed genes in open chromatin. We note that the majority of TG mutations occurred along with mutations

in other established PTC driver genes such as *BRAF* and *NRAS*, *HRAS*, and *KRAS*, suggesting that mutations in *TG* are not sufficient as oncogenic driver mutations on their own. None of the mutations we observed were reported previously in COSMIC, although the database does contain roughly 1000 other previously reported SSVs in *TG* and specific mutations in *TG* are associated with enlarged thyroid, goiter, and other clinical symptoms [192–195]. Given this evidence, we do not consider *TG* to be a primary driver of papillary thyroid carcinoma.

MutSig reported *TSHR* (thyroid stimulating hormone receptor) as significantly mutated in the REBC sample sets. Activating mutations in *TSHR* are found in approximately 4% of thyroid nodules [196]. MutSig reported *TSHR* mutations in 2% (6 / 381) of the REBC samples (4 Exposed, 2 Unexposed) and 1 THCA sample. The patterns observed in *TSHR* differed from those observed in *TG*. The ratio of noncoding:coding mutations in *TSHR* was 4:1 and the majority of mutations were substitutions rather than indels, suggesting that *TSHR* is not subject to the same mutational process affecting *TG*. Four of seven samples with a *TSHR* mutation had at least one primary driver mutation. Two samples had known driver mutations in *BRAF* and *NRAS* and one had a whole-genome CNA pattern; the remaining *TSHR* mutant sample was the SSV hypermutator. Considering the presence of other drivers and the relatively low number of mutations across the sample sets it is unlikely that *TSHR* is a primary driver of papillary thyroid carcinoma, though the possibility that it may be contributing to the cancer phenotype cannot be ruled out from whole-genome data alone.

DICER1 was also reported as significantly mutated by MutSig. Three samples harbored a coding *DICER1* mutation, with two samples each possessing two independent mutations. *DICER1* codes for an endoribonuclease that is essential to microRNA processing. Germline mutations in *DICER1* are associated with predisposition to endocrine cancers and pleuropulmonary blastoma, a condition known as *DICER1* Syndrome [197, 198]. Somatic mutations in *DICER1* tend to occur at known hotspots *in trans* with germline hits to produce a second hit. All three samples with reported protein-altering mutations in *DICER1* were not exposed to radiation.

2.2.5 Pancancer drivers and frequently mutated genes

Three other genes were reported by MutSig as significantly mutated. *AKT1* is a well-established pan-cancer driver with clear oncogenic properties [199, 200] and was reported as mutated (though not significantly) in sporadic thyroid cancer cases [53]. Three mutations were found in the *AKT1* tumor supressor gene, though all of these samples

also contained mutations in other genes. Two genes, $CR2$ and $DHX30$, were reported as significantly mutated but are likely not drivers. $CR2$ encodes a membrane receptor to which Epstein-Barr virus binds to lymphocytes. Two of the three observed $CR2$ mutations were predicted to be protein damaging by CADD and PolyPhen2 [201]. $DHX30$ encodes a DEAD box protein that is a putative RNA helicase. None of the mutations we observed in $CR2$ or $DHX30$ were previously observed in COSMIC.

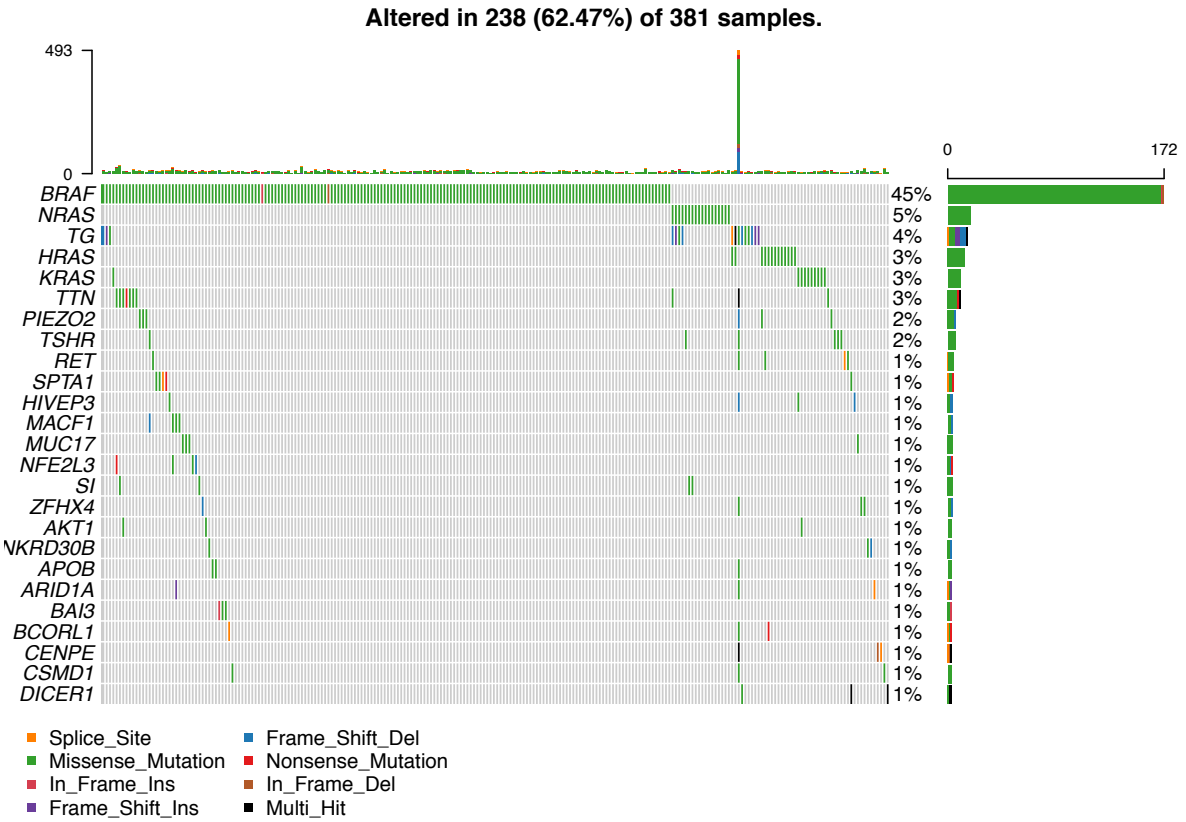


Fig. 2.11 Oncoplot of frequently mutated genes for SSVs in the REBC sample set.

In addition to significantly mutated genes I looked for genes that were mutated frequently (but not necessarily reported as significant by MutSig) using maftools [202]. *BRAF*; *NRAS*, *HRAS* and *KRAS*; *TG*; *TSHR*; *AKT1*; and *DICER1* were again reported among the most frequently mutated genes. Other known thyroid cancer related genes were reported as well. *RET* mutations were reported in four samples. *RET* is a well-established fusion partner in both sporadic and radiation-associated thyroid cancer, where activating fusions drive oncogenesis [189]. Activating mutations in *RET* are also associated with the Multiple Endocrine Neoplasia family of familial cancer syndromes

[203, 38]. RET^{K1007N} was predicted to be damaging by CADD, SIFT and POLYPHEN; RET^{R163Q}, RET^{A26V}, were predicted to be benign; and RET^{R67C} was predicted to be benign by all tools but POLYPHEN, which gave it a "possibly damaging" categorization. Frequent mutations in *TTN*, *PIEZO2* and many other genes were likely due to their size rather than oncogenic effects.

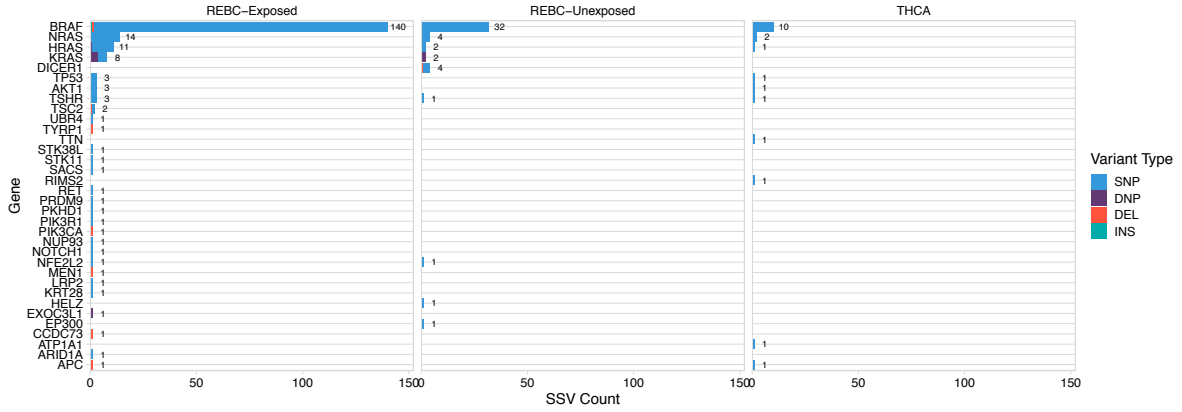


Fig. 2.12 SSV counts in genes, restricted to SSVs which overlap sites previously reported as mutated in COSMIC.

I then examined mutations in frequently-mutated genes which overlapped variants in the COSMIC database, broken down by sample set (Figure 2.12). Notably, all of the *DICER1* mutations in our dataset come from the unexposed cases. Three exposed samples and one THCA sample had $DICER1$ mutations.

Mutations in other cancer-related genes were reported. There were no *TP53* mutations in the unexposed samples. $TP53$ is the most-commonly mutated gene across all cancers [187]. $TP53$ mutations in thyroid cancer are associated with the more aggressive low differentiated subtypes [204]. Mutations in *PIK3CA*, *APC*, *NOTCH1* and *MEN1* were also reported.

2.3 Insertion and deletion spectra

I next examined the insertion and deletion (collectively indels) spectra of samples. Indels have been implicated as a possible result of radiation exposure in the genome [54]. For these analyses, I chose to exclude the SSV hypermutator sample as well as any samples deemed to have less than 20% purity, as these would not provide an accurate overview of the total number of indels. It is also important to note that the THCA samples

were sequenced using reads that are roughly fifty basepairs shorter than the sequencing platform used for the REBC sets. This likely has an effect on the ability to call longer indels.

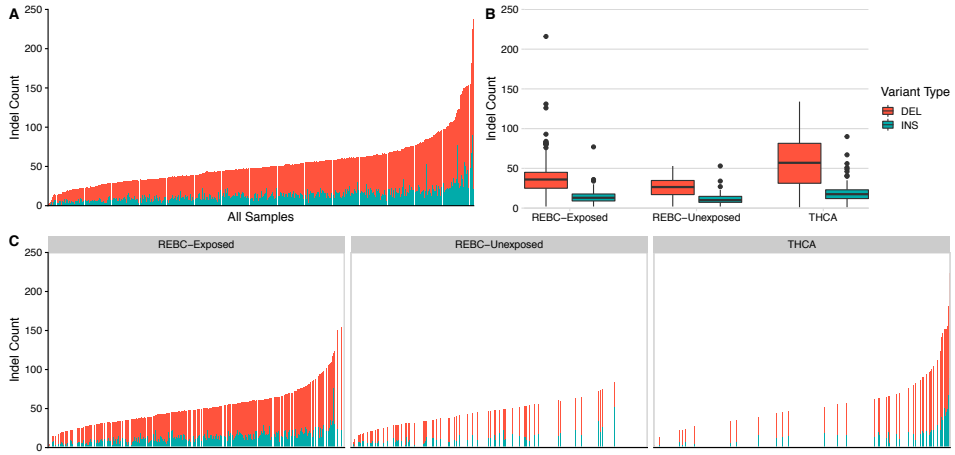


Fig. 2.13 Insertion / deletion counts per sample

The number of indels per sample varied slightly across sample sets. THCA had a median indel count per tumor of 64. REBC-Unexposed samples carried a median of 35 while REBC-Exposed samples carried a median of 49. REBC-Exposed carried a median 36 deletions (mean = 38, range = [2, 216]) and a median 13 insertions (mean = 13, range = [2, 77]). THCA carried a median 53 deletions (mean = 5, range = [9, 119]) and 16 insertions (mean = 20, range = 1, 82). The REBC-Unexposed set carried a median 26 deletions (mean = 26, range = [2, 53]) and 10 insertions (mean = 11, range = [2, 53]).

2.3.1 Indel burden increases with age

Both SNVs and indels accumulate randomly with age even in normal cells [35, 34]. The number of indels per tumor and age at surgery were significantly correlated across all samples in our study ($r^2 = 0.54$, $p \leq 0.001$, Figure 2.14).

When restricting the dataset to only REBC-Exposed and REBC-Unexposed samples, this correlation was weaker but still significant ($r^2 = 0.32$, $p \leq 0.001$). When excluding low purity samples, r^2 was slightly higher, indicating that purity may impact the ability to call indels ($r^2 = 0.35$).

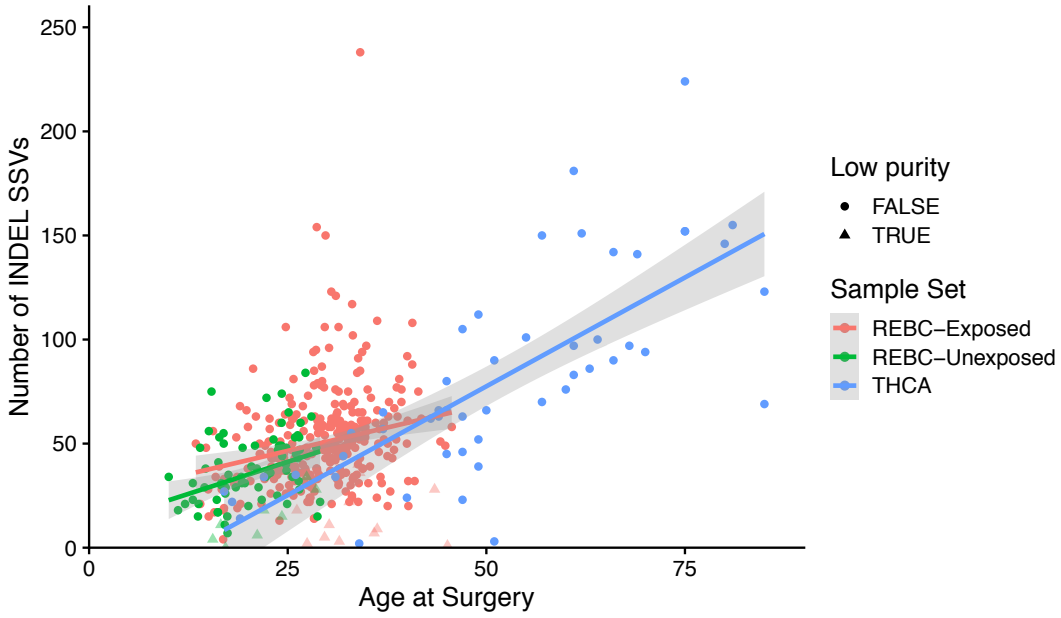


Fig. 2.14 Indel SSV burden generally increases with age. Low-purity samples also appear to have lower-than-expected indel counts.

2.3.2 The indel to substitution ratio is correlated with radiation dose

[54] reported that deletion events were the main signatures of radiation exposure, though they examined only twelve samples which were exposed to high levels of medical radiation (up to 30 Gy). The 303 REBC-Exposed samples were exposed to much lower amounts of radiation, with most being exposed to less than 100 mGy (Table 2.2).

One metric reported by [54] to be indicative of radiation exposure was the ratio of indels to substitution variants. [54] used the indel:substitution ratio as a method for correcting for the age-related accumulation of mutations, analogous to comparing the residuals of predicted indel burden based on age versus observed indel counts.

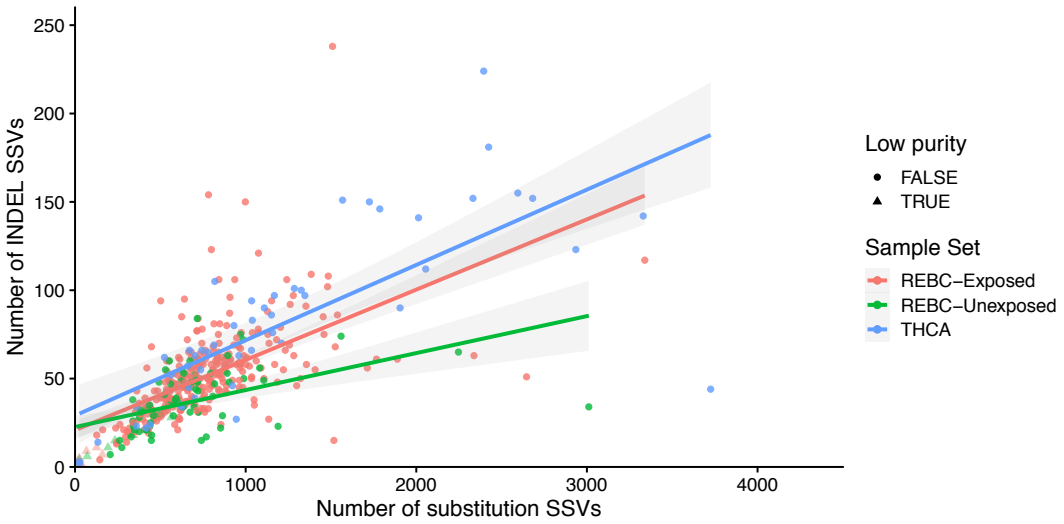


Fig. 2.15 The number of indel SSVs generally increases with the number of substitution SSVs.

After excluding the SSV hypermutator and low-purity samples, I assessed the indel:substitution ratio as a predictor of radiation dose. Figure 2.15 shows that the number of indels increases with the number of substitution SSVs. Both indel count and substitution count are correlated with age at surgery (indels: $r^2 = 0.53$; substitutions: $r^2 = 0.48$; $p < 0.001$ for both).

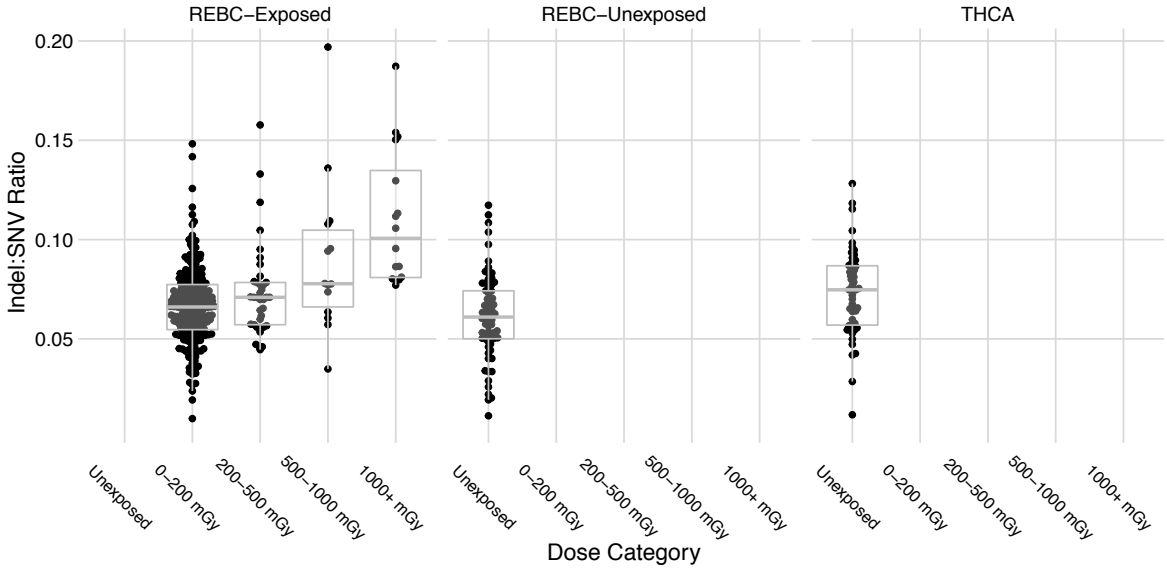


Fig. 2.16 The indel:substitution ratio, broken down by dose category and sample set.

I first examined dose broken down categorically into five groups. Across dose groups, the indel:substitution ratio increases with increasing dose Figure 2.16. A one-way ANOVA test indicates that the means of these groups are not the same ($p < 0.001$). Unexposed samples have roughly the same indel:substitution ratio across all sample sets.

I next examined dose as a continuous predictor for the indel:substitution ratio (Figure 2.17A). Excluding samples that received an estimated dose of zero, the indel:substitution ratio was moderately correlated with dose ($r^2 = 0.46$, $p < 0.001$).

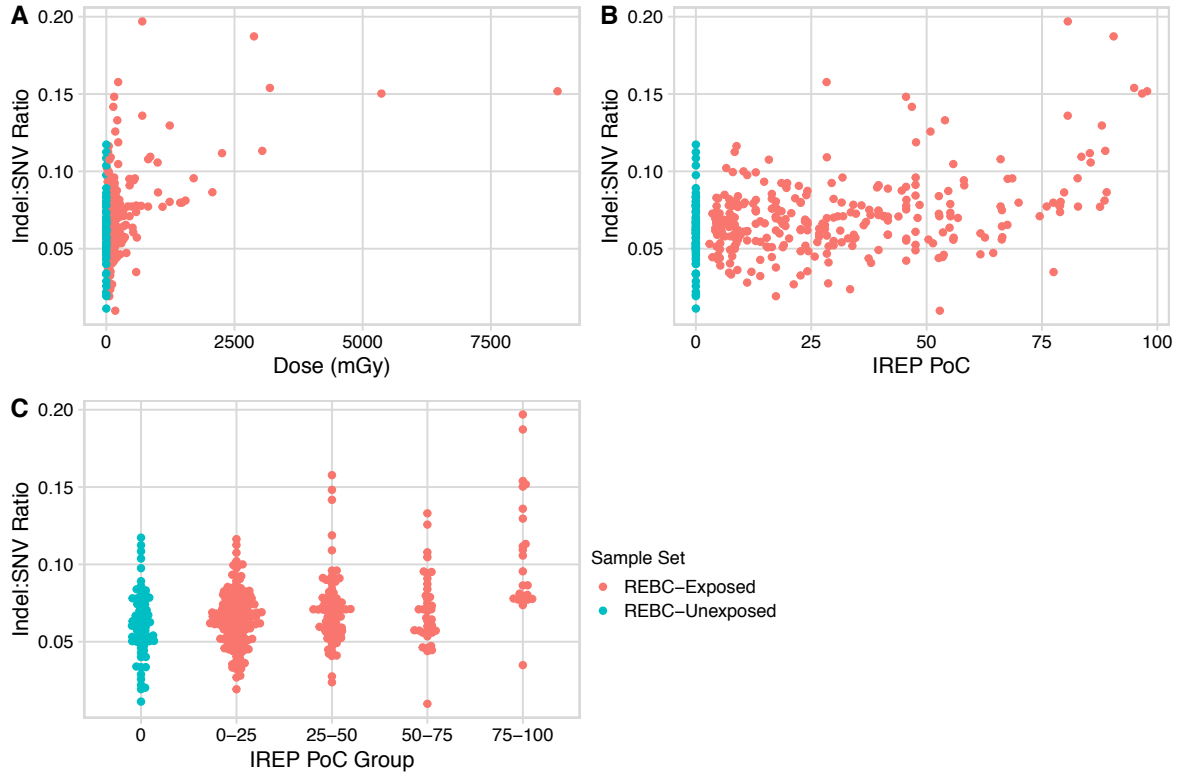


Fig. 2.17 (A) The indel:SNV ratio plotted against estimated radiation dose and (B) IREP probability of causation.

Last, I looked at how the indel:substitution ratio responded to the IREP-estimated probability of causation (Figure 2.17B). IREP was developed to estimate the probability that a radiation exposure of known dose caused an individual's cancer. The model incorporates attained age as part of its calculation and produces both an estimate and a confidence interval for the predicted probability a given case was caused by radiation. I used the 50^{th} percentile estimate for the IREP value. There was a significant correlation between the IREP POC 50^{th} percentile estimate and the indel:substitution

ratio (Pearson's test p-value < 0.001, r^2=0.39). When examining IREP as a discrete set of groups, there is a positive relationship with the indel:susbtitution ratio.

2.3.3 The deletion to insertion ratio is correlated with radiation dose

In addition to the indel to substitution ratio, [54] reported that the ratio of genomic deletions to insertions was correlated with exposure to radiation. Reference [54] found that radiation-associated deletions occurred randomly across the genome, in sharp contrast to deletions and insertions not caused by radiation. Reference [54] also reported high rates of microhomology at deletion breakpoints, indicative of microhomology-mediated non-homologous end joining as the mechanism of repair in these tumors.

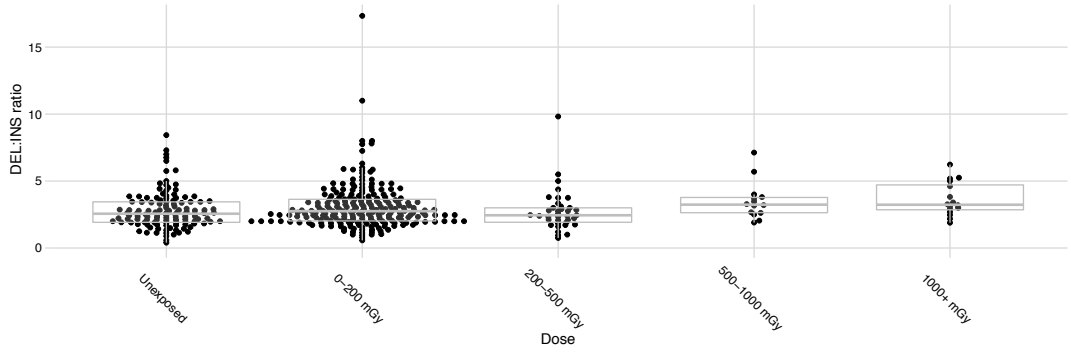


Fig. 2.18 The DEL:INS ratio.

In REBC-Exposed samples, there was a significant positive correlation between continuous dose and the deletion:insertion ratio ($r^2 = 0.12$, $p = 0.036$). However, dose was a better predictor of the indel:substitution ratio than the deletion to insertion ratio. The mean deletion:insertion ratio was not significantly different between dose groups (one-way ANOVA, $p = 0.05$).

It is difficult to explain why the insertion:deletion ratio is not a strong responder to increasing dose in the REBC sample set. The REBC study had relatively low insertion and deletion counts in general which makes statistical inference difficult. It is possible that the inflation in the number of deletions due to radiation exposure in this sample set is insufficient to see an effect.

2.3.4 The deletion to substitution ratio is a better predictor of exposure than the indel:substitution ratio

I next tested whether the deletion:substitution ratio might be a better measure of radiation exposure than the indel:substitution or deletion:insertion ratios. The deletion:substitution ratio was significantly correlated with continuous dose ($r^2 = 0.48$, $p < 0.001$). The strength of this correlation is slightly higher than that between the indel:substitution ratio and continuous dose.

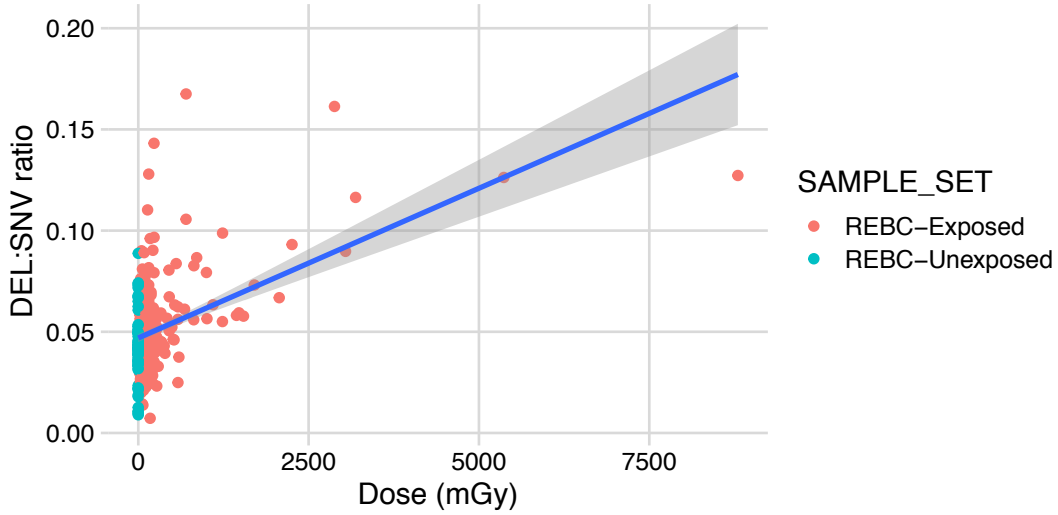


Fig. 2.19 The deletion:substitution ratio. This ratio generally increases with increase radiation dose.

A one-way ANOVA indicated that the mean deletion:SNV ratio differed between dose categories. In general, the deletion:SNV ratio increases with increasing dose.

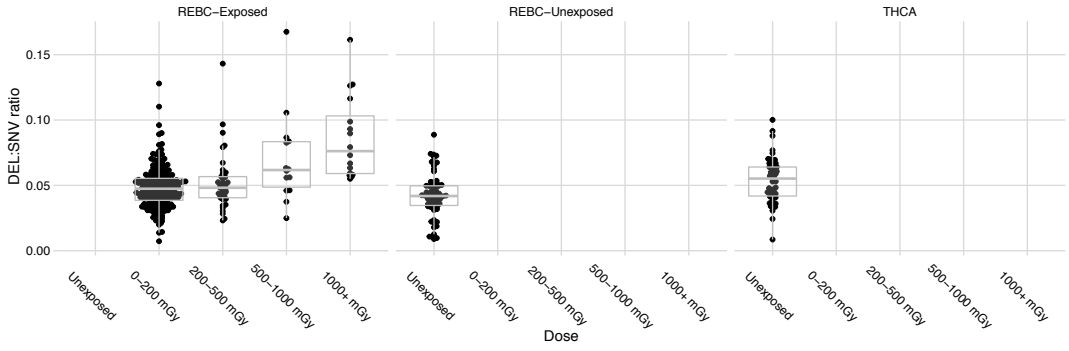


Fig. 2.20 The DEL:SNV ratio.

The deletion:SNV ratio is a slightly better response variable to radiation exposure than the indel to SNV ratio. This is in line with the findings of [54] that deletions, and not insertions, are the primary small-variant indicator for exposure to radiation.

2.4 Somatic mutational signatures

Next, I applied the SignatureAnalyzer framework to look for substitution mutational signatures in our sample sets. Signature analyzer uses non-negative matrix factorization (NMF) to estimate mutational signatures of exposures and their corresponding amounts based on counts of mutations in the 96 trinucleotide contexts. SignatureAnalyzer attempts to automatically choose the best number of signatures for decomposition using a Bayesian sampling approach.

As sample selection can significantly influence mutational signature detection, I performed mutational signature decomposition using specific combinations of the data. I first applied the method to all samples (THCA, REBC-Exposed, and REBC-Unexposed), excluding the hypermutator and low purity samples.

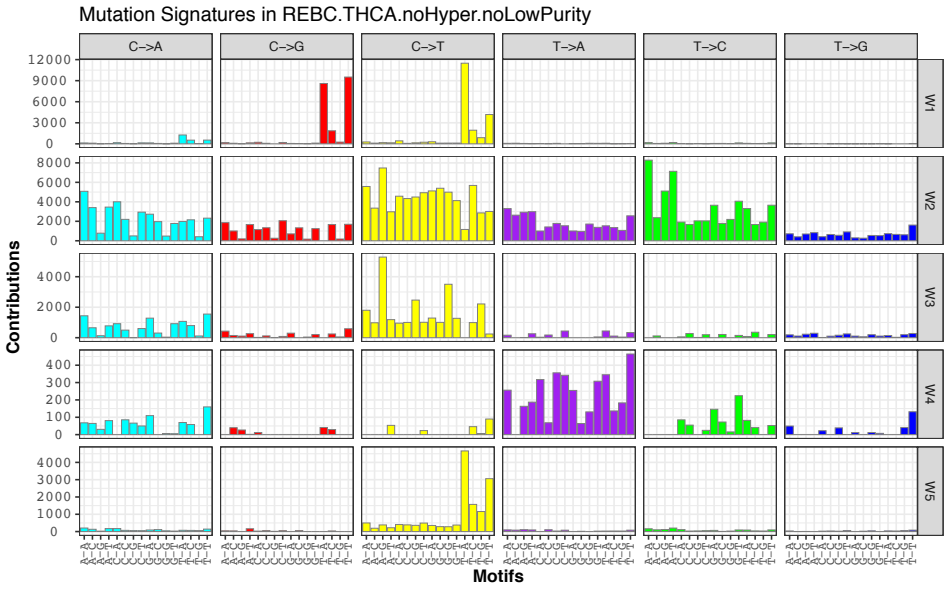


Fig. 2.21 Signatures detected in THCA and REBC (excluding SSV hypermutator and low purity samples).

The model converged to four signatures in the majority of its runs, though SignatureAnalyzer reported five signatures as the best fit in two of its ten runs. However, the five-signature model shows higher similarity to the well-defined signatures from COSMIC

(v3). I have displayed the results of the five-signature model in Figure 2.21 for this reason.

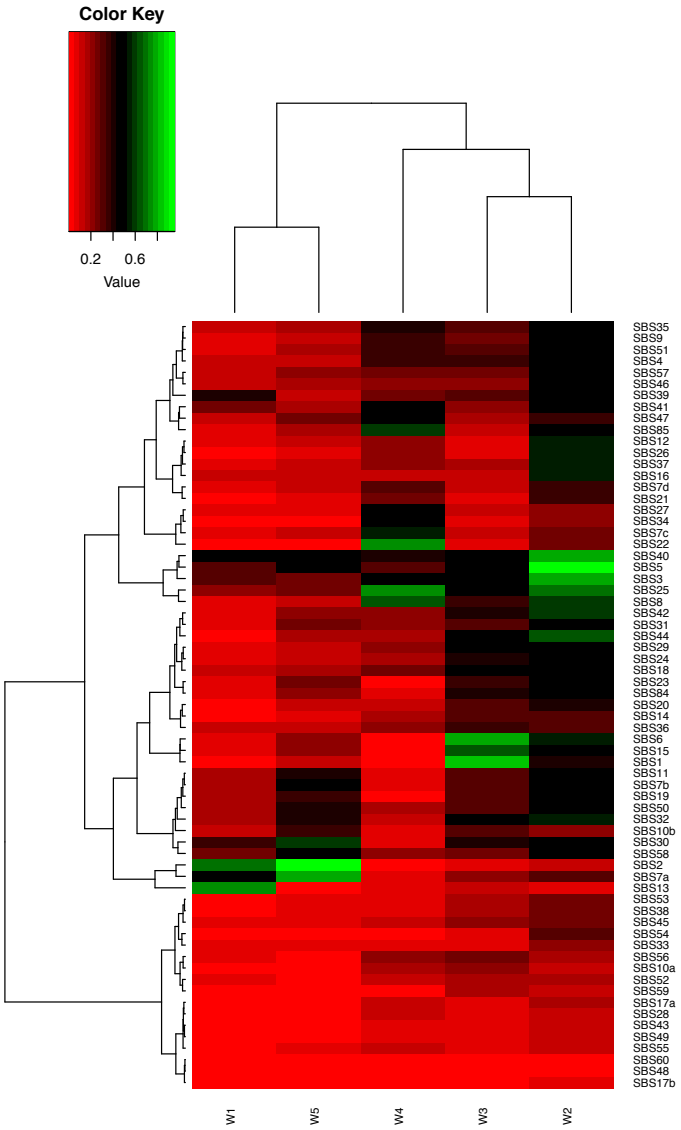


Fig. 2.22 Five extracted mutational signatures from THCA and REBC samples (excluding the SSV hypermutator and low purity samples) compared to COSMIC V3 SBS signatures.

I compared the mutational signatures found in our dataset(s) to the catalog described in COSMIC v3 which were first described in reference [44]. This catalog contains significantly more signatures than the COSMIC v2 signatures [32], and several of the older signatures have been better resolved into multiple constituent new signatures.

2.4.1 Evidence of *APOBEC* activity in PTC

Signatures W1 and W5 most resemble COSMIC signatures SBS13 and SBS2, respectively. Both of these signatures are associated with APOBEC cytidine deaminase activity [60] and have been previously reported in thyroid carcinoma as well as other cancers [44]. The activity levels of signatures W1 and W5 vary across tumors in our sample set, with some tumors harboring many more mutations attributable to the underlying mutational processes. This is consistent with unequal APOBEC activity among tumors.

2.4.2 Clock-like signatures predominate in raPTC

Signature W2 closely resembles SBS5, with relatively high similarity to SBS3 and SBS40 as well. All three of these signatures have relatively flat profiles and high similarity to one another. SBS5 is seen in all cancers and is associated with signature SBS1 in some cancer types [44]; both SBS1 and SBS5 are sometimes referred to as "clock-like" due to the accumulation of associated mutations with age in both tumors and normal cells. While aetiologies for SBS5 and SBS40 have not yet been established, both are correlated with patient age in some cancers. SBS3 is related to deficiency in homologous-recombination based DNA repair. Manual inspection and the presence of signature SBS1 (described below) suggest that signature W2 is SBS5.

Signature W3 closely resembles SBS1, which is associated with an endogenous mutational process. SBS1 has previously been reported in thyroid cancer. Overall, signatures W1, W5, W2 and W3 reflect previously established mutational patterns in thyroid cancer. These signatures were well-defined by the five-signature NMF model but not the four-signature model. The remaining signature of the five-signature model (W4) most closely resembles SBS22 and SBS25. SBS22 is associated with aristocholic acid exposure and SBS25 is associated with exposure to chemotherapy, though neither signature has been reported previously in thyroid cancer. Signature W4 is distributed relatively evenly across our samples. Given this evidence, it is likely that W4 is an artifact of overfitting.

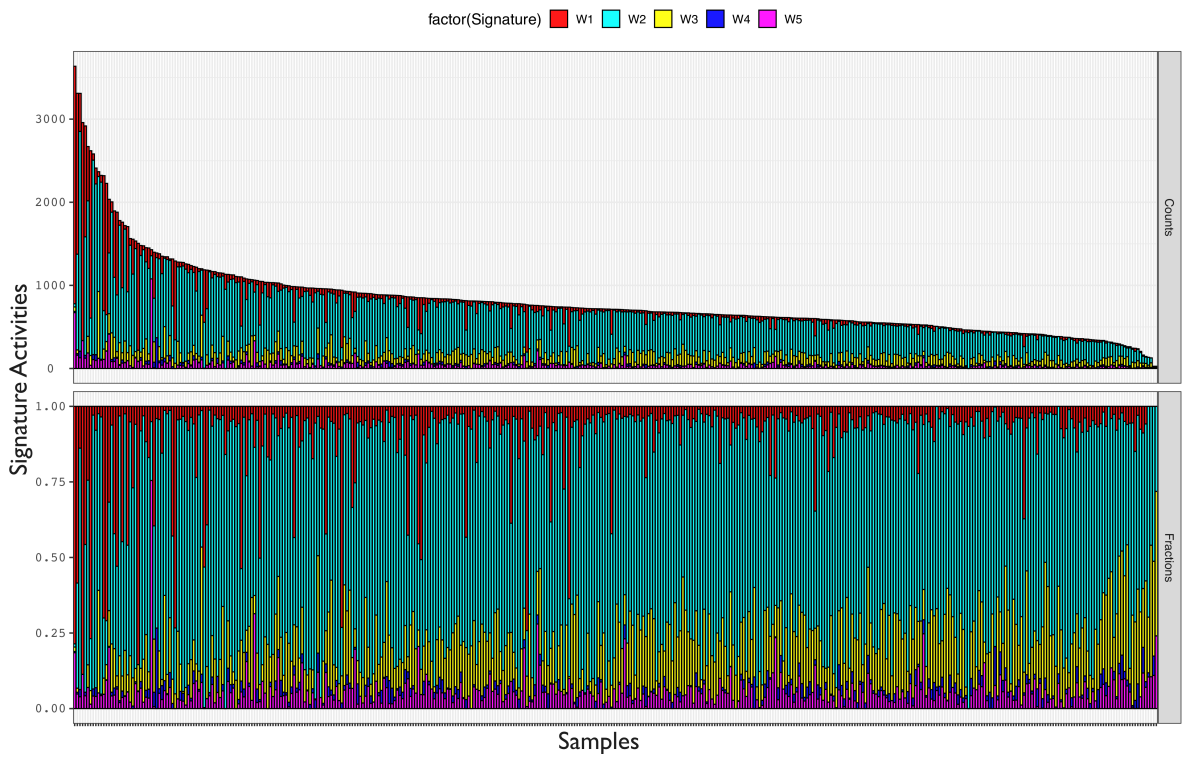


Fig. 2.23 Signature activity in THCA and REBC (excluding SSV hypermutator and low purity samples).

Signature activities vary across samples in our dataset, though some trends exist. THCA samples tend to have more mutations attributable to W3 (SBS1). These samples are significantly older, which may partially explain this difference. Most mutations in the dataset are attributable to Signature W2 (SBS5).

The SSV hypermutator sample was excluded from signature analysis because it had an order of magnitude more mutations than any other sample. Single-base substitution context proportions for this sample are shown in Figure 2.24. This profile resembles COSMIC signatures SBS1 and SBS5 as well as components of several signatures proposed to be the result of defects in DNA repair and mutations in DNA polymerase genes (including SBS6, SBS14, SBS15, SBS20, SBS21). Defective DNA repair mechanisms might help explain the high burden of mutations in this tumor. While this tumor had multiple single hits in DNA repair genes and proofreading polymerases, no mutations that could sufficiently explain its hypermutator status were found in either its somatic or germline catalogs. Further integrated analysis including expression and methylation data will be required.

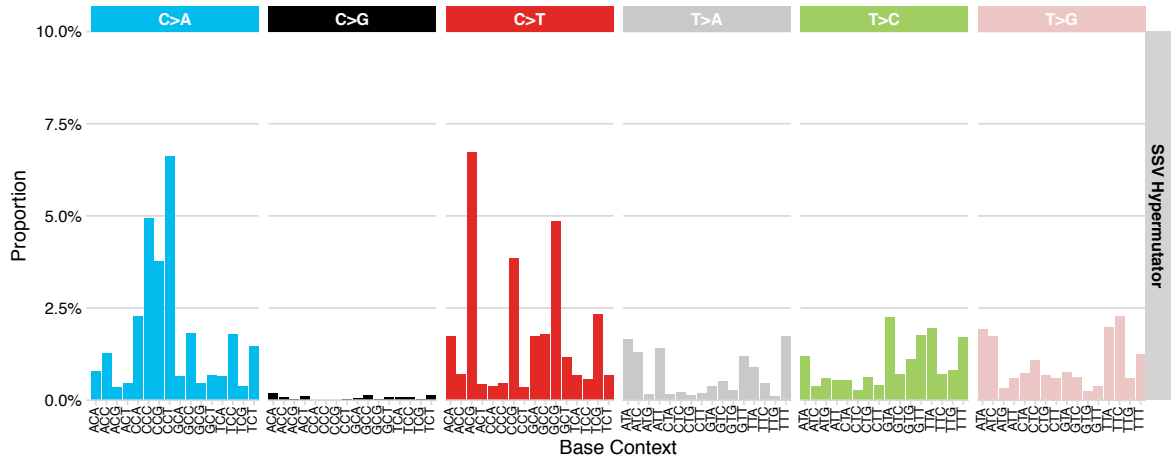


Fig. 2.24 Trinucleotide mutational context proportions for the SSV hypermutator sample. The profile of the mutational context proportions shows the characteristic peaks of SBS1 in C>T base changes. The large proportion of C>A base changes resembles a number of SBS signatures related to DNA repair deficiency combined with mutations in proofreading polymerases (including SBS14 and SBS20).

Overall, these results recapitulate established mutational signatures in thyroid carcinoma and do not suggest the existence of a radiation-related mutational pattern in the 96-context substitution model. It is possible that longer substitution contexts might yield evidence for a radiation exposure signature. However, the findings of [54] indicate that an indel mutational signature for radiation exposure would be more likely. Additionally, most of the variation in substitution SSV burden is attributable to age (Figure 2.8). If radiation is inducing substitutions in the REBC-Exposed samples, those mutations are accumulating at a rate below the sensitivity of SignatureAnalyzer to detect signatures in this dataset.

2.4.3 Indel signatures are associated with radiation exposure

Reference [54] previously reported an increase in the number of small deletions, particularly those longer than 2-3bp and with with microhomology at the breakpoint, within radiation-exposed tumors. We found a significant association between the number of small deletions and increasing radiation dose by examining the indel:substitution ratio in the REBC sample set (see subsection 2.3.2). Our analysis, however, examined deletions globally and did not take into account the length of the event or local sequence context around the indel.

Reference [44] used 83 different indel categories, or features, to generate indel mutational signatures in more than 2,000 tumors from the Pancancer Analysis of Whole Genomes study. These features incorporate information about the length of the indel, the deleted base (for single-base indels), and whether the indel lies within a repeat or has microhomology at its ends. Current versions of SigProfiler [205, 44] can take as input a count matrix of these features and extract indel mutational signatures.

I calculated counts of indel features using https://github.com/edawson/presig, which takes a MAF file as input and outputs a SigProfiler-compatible indel feature counts matrix. I then performed NMF-based signature extraction using SigProfiler (python version 0.0.5.76; fit 2 to 6 signatures; 1000 iterations; 16 cores on a 16-core Google compute-optimized VM).

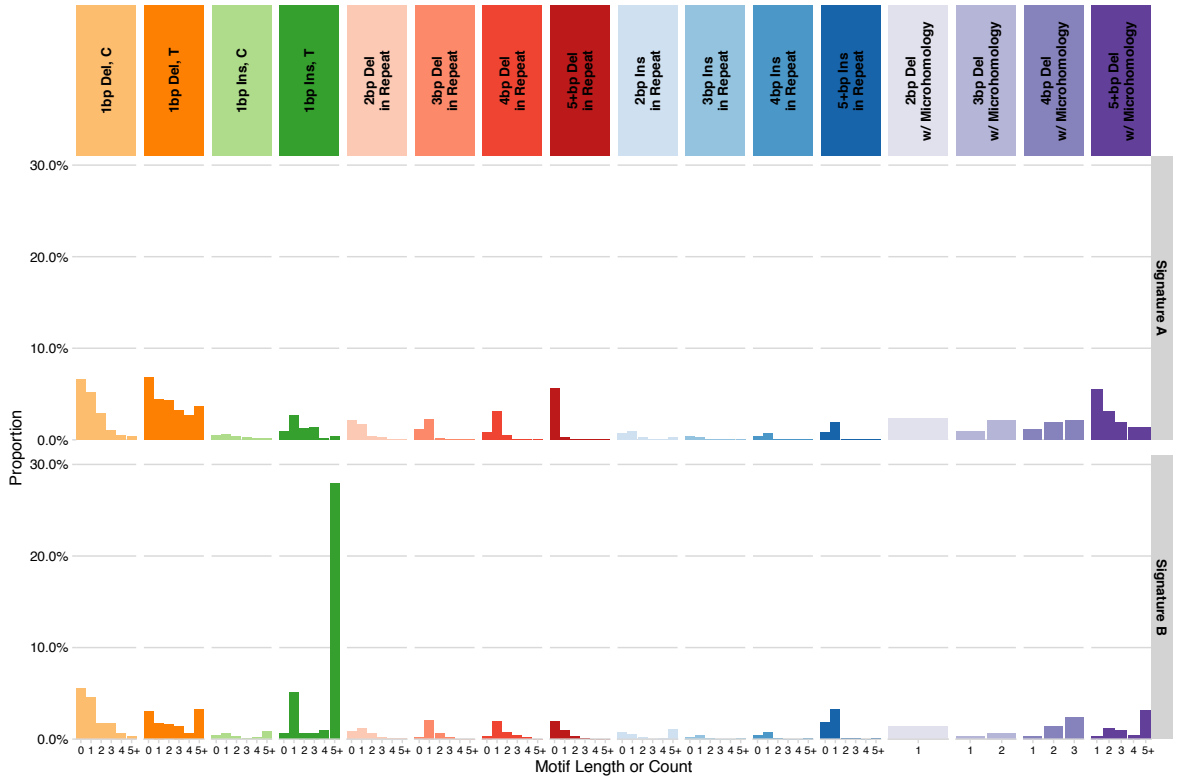


Fig. 2.25 *De novo* extracted indel (ID) signatures from SigProfiler for REBC samples. Signature A resembles COSMIC ID8, a putative radiation signature. This signature contains a major component of large deletions in repetitive elements and with microhomology. Signature B is highly similar to COSMIC ID1, a background mutational process ubiquitous in almost all samples. There is strong evidence of remaining convolution in both signatures; nearly all of the indel features that are not characteristic of ID1/ID8 are shared between Signature A and Signature B.

The suggested solution reported by SigProfiler was for two signatures (shown in Figure 2.25). Signature A resembles COSMIC ID8. ID8 was also reported in signature decomposition performed automatically by SigProfiler (shown in Figure 2.26). It has been suggested that COSMIC ID8 is associated with defective repair of double-strand breaks, possibly induced by radiation [44]. The second signature reported (Signature B) was enriched for insertions of a single base in cytosine homopolymers greater than five basepairs in length (COSMIC SBS1). This signature has been attributed to slippage during DNA replication and is found in almost all cancer samples.

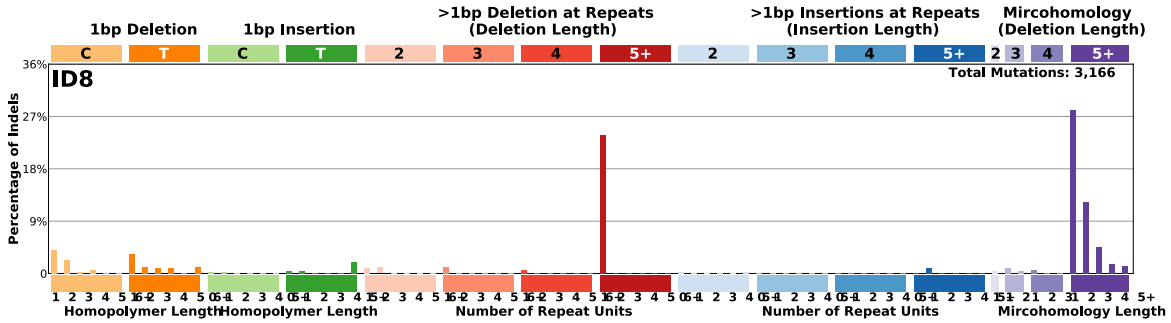


Fig. 2.26 COSMIC indel signature ID8. ID8 is composed of mosly large deletions, often within repeats or with small portions of microhomology (dark red and purple portions).

Roughly fifteen percent of all indel mutations in the REBC sample sets were attributed to Signature A (3,166 of 20,257 total indels). I next assayed whether there was a correlation between the primary features of ID8 and radiation dose.

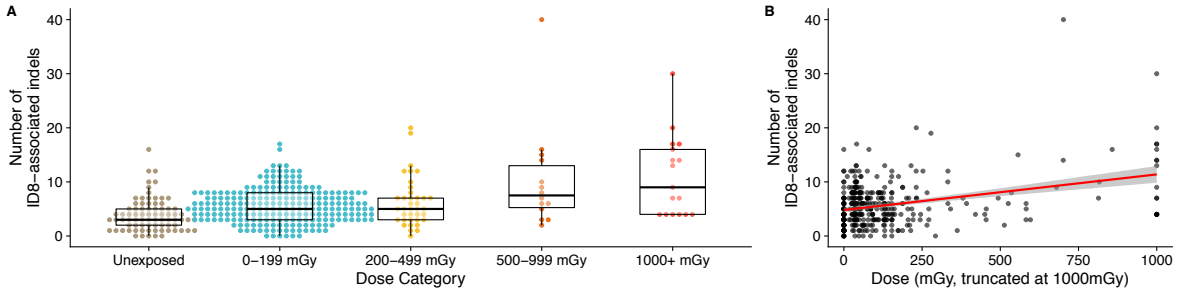


Fig. 2.27 (A) A beeswarm plot of the number of ID8-associated indel mutations per sample, broken down by dose category. The number of ID8-associated indels increases with increasing dose category. (B) The number of ID8-attributable indels plotted with continuous radiation dose. There is a signficant association between the number of ID8-attributable indels and increasing radiation dose (shown here with values over 1000 mGy truncated to minimize the effect of outliers).

The amount of ID8-attributable indels generally increases with increasing dose category (Figure 2.27A). There was a signficant association between these indel features and radiation dose (generalized linear model, $p < 0.001$) as well as age at surgery (generalized linear model, $p < 0.001$), consistent with results described in subsection 2.3.2.

These results are consistent with those of reference [54], which reported an excess of deletions with evidence of non-homologous end joining or microhomology-mediated repair at their ends in radiation-exposed tumors. The same signature was observed in [44] as well, primarily in tumors with DNA repair defects and concurrent evidence of COSMIC SBS3. The REBC tumors lack evidence for high SBS3 activity. Reference [44] also reported an association with age in some tissues. It is also important to note that the total number of such indels per REBC sample is low, with a median of five per tumor (range = [0, 40]), and that these indels also appear to be associated with age both in the REBC sets and in [44]. A method for correcting for the effect of age, like the indel to substitution ratio, may be useful as a correlate to dose. This metric would still be unable to attribute individual indels to radiation exposure, however. These results nonetheless highlight a possible signature of radiation exposure and explain why radiation, while a powerful carcinogen, does not appear to leave marked signatures in the single-base substitution spectrum.

2.5 Germline variants

Germline variants can predispose individuals to higher risk of cancer [42, 37]. The germline background also provides a genomic context that can explain cancer development in the absence of a second recessive somatic hit or clear dominant somatic driver. Germline variants in specific genes can predispose individuals to cancer after specific exposures. Established radiosensitivity variants in genes such as ATM and $RAD51C$ greatly increase the risk of cancer development after exposure to ionizing radiation [206–209]. These variants have largely been characterized from genome-wide association studies following genotyping or exome sequencing of many thousands of individuals.

The REBC study is not of sufficient size for discovery of novel predisposition variants. At least several hundred more cases and controls would have been needed to detect variants with even very large effect sizes. It is however possible to accurately detect established predisposition variants in each sample given that each normal was whole-genome sequenced to sufficient depth ($\geq$ 40X for most normal samples). When overlayed with information from somatic variant calling, a more complete picture of each tumor's

Sample Set	Dose Category	% of samples with T1 variant
REBC-Unexposed	Unexposed	61%
THCA	Unexposed	46%
REBC-Exposed	0-200 mGy	53%
REBC-Exposed	200-500 mGy	70%
REBC-Exposed	500-1000 mGy	57%
REBC-Exposed	1000+ mGy	35%

Table 2.4 Percentages of samples in each dose category that have at least one Tier 1 germline variant. Error bars are for a 95% confidence interval (binomal test).

life history emerges. In the REBC cohort, we also sought to explain the genesis of thyroid cancer in some of the unexposed individuals, who are well below the median age of onset yet still developed thyroid cancer in the absence of a known environmental exposure.

I called germline variants in these samples using a consensus calling mechanism. The Strelka pipeline I implemented output germline SNV and indel calls along with the somatic calls. I ran FreeBayes [210] in single-sample mode for every normal sample in our dataset. I also ran `GATK HaplotypeCaller` for each sample and then performed joint genotyping using `GATK GenotypeGVCFs` [92].

The results from these three callers were then normalized, combined, and annotated by Dr. Hartley. Variants were all left-aligned and trimmed to normalize their representation. Normalized variants were then merged into a single VCF with information about the caller(s) of origin for each variant. This VCF was then annotated with population allele frequency information as well as variant impact information from dbNSFP [211]. From this information, the impact of a variant was calculated according to ACMG guidelines [212] and with a custom tiering system. Only variants labeled Tier 1 (ACMG P/LP, present in ClinVar [213], or SNPEFF high-impact [214]) were used in this analysis.

One hundred and sixty-six genes had at least one Tier 1 variant. Tier 1 germline mutations included well-established cancer predisposition genes such as $BRCA1/2$ and $PTEN$; genes involved in DNA repair and microsatellite instability ($ERCC2/6/8$); Fanconi Anemia predisposition genes ($FANCA$, etc.); and genes known to contribute to radiation sensitivity (ATM, $RAD51$).

Across sample sets the number of samples with a Tier 1 germline variant was roughly the same (REBC-Exposed: 55%; REBC-Unexposed: 61%; THCA:46%; pairwise Fisher test p-values all > 0.1).

The proportion of samples with a Tier 1 variant varied across dose categories. This variation was not significant, however. We hypothesized that the number of cases

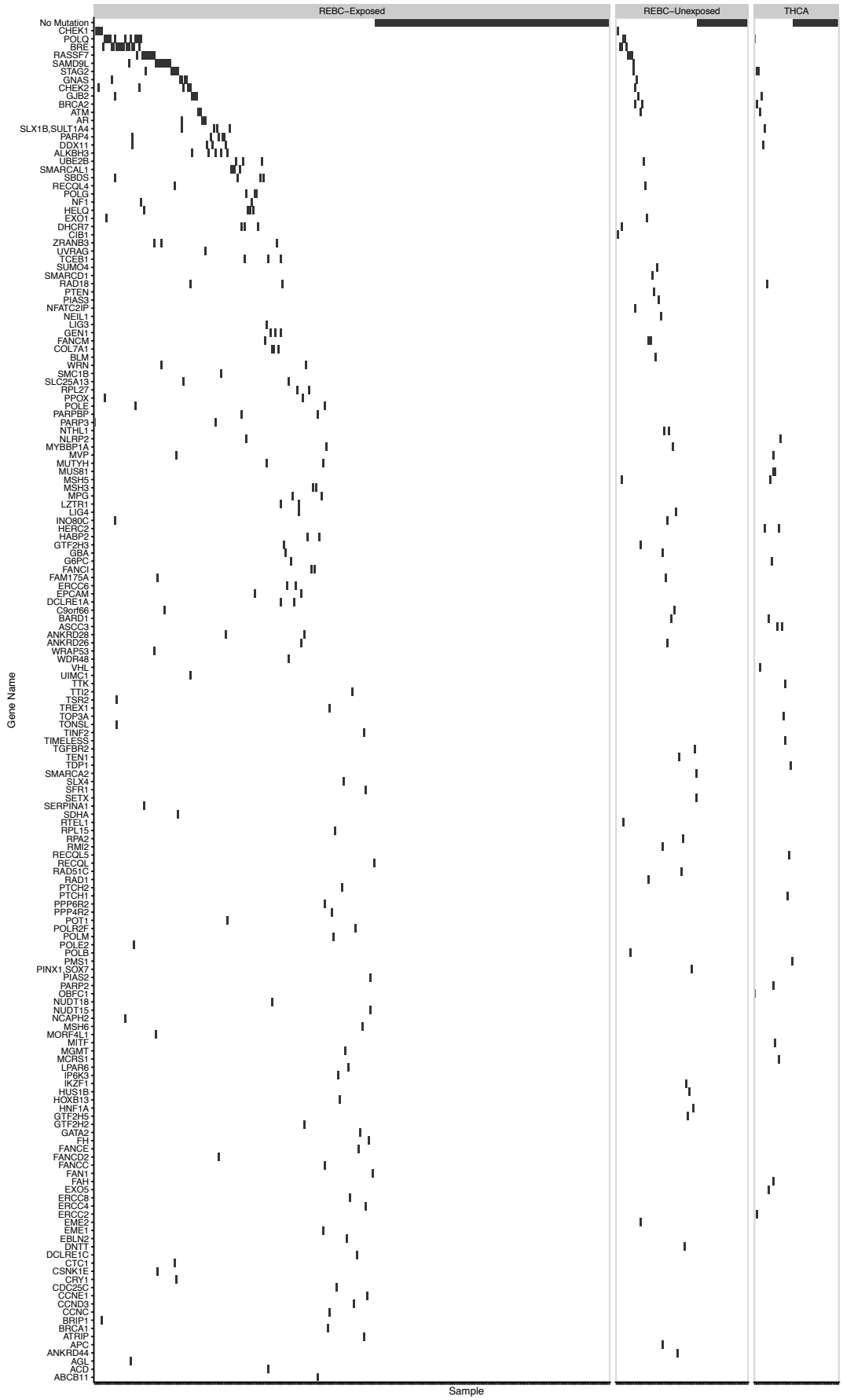


Fig. 2.28 A waterfall plot of genes impacted by germline mutations.

contributed by germline predisposition variants would decrease with increasing dose, as more cases would be caused by mutagenic exposure to ionizing radiation. However, neither dose nor the IREP POC 50^{th} percentile estimate were significant predictors of whether a sample had at least one Tier 1 variant (logistic glm, dose $p = 0.13$, IREP $p = 0.87$; proportions per dose category with 95% confidence interval error bars for the binomal test are shown in Figure 2.29). Our results suggest that there are not major differences in the rates of Tier 1 germline variants across groups. Proportions of samples with at least one Tier 1 variant are listed in Table 2.4.

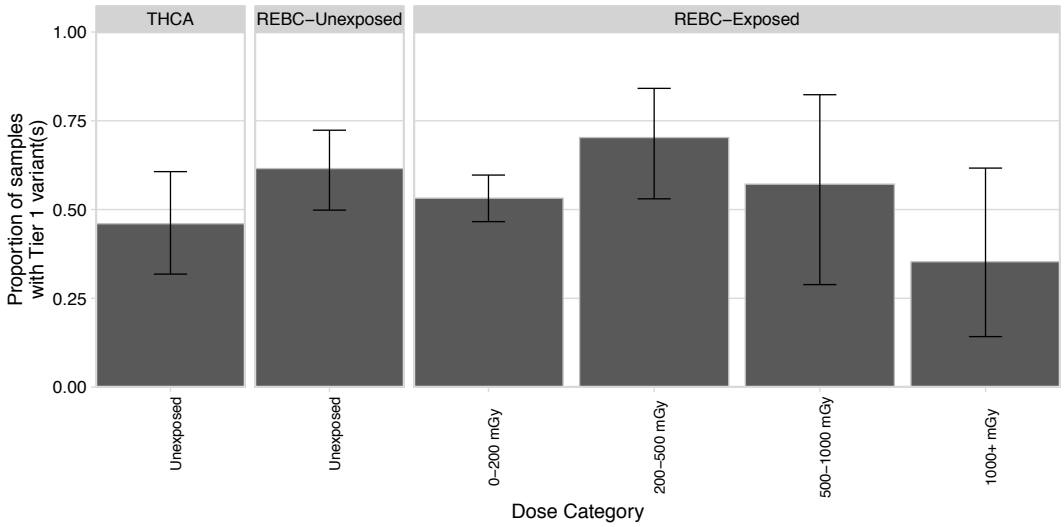


Fig. 2.29 Proportions of each sample set with at least one Tier 1 variant.

We observed some frequently mutated genes across samples sets, though we did not have sufficient power to find germline variants associated with radiation sensitivity even at exceedingly high risk ratios (Figure 2.30). I restricted this analysis further to only look at the missense, frameshift, or stop-gain mutations. *POLQ* was mutated across in 14 samples (1 THCA, 13 REBC) from all but the highest dose category. All of these mutations were the same allele at the same position. It is possible that this mutations is a germline polymorphism that is present at a higher frequency in the study population. The next most-frequently observed mutated gene was *CHEK*1. *CHEK*1 mutations were present in 7 of 381 REBC samples (1.8%). Two mutations were present in each of these samples; based on local alignment of reads, these are likely to be false positives.

*GJB*2 was the next most frequently mutated gene. Seven samples (1 THCA, 6 REBC) have mutations in *GJB*2. These variants were predicted to be pathogenic. Other genes involved in DNA repair were repeatedly reported. The *POLE* and *POLG* genes were

both reported frequently, as were genes involved in DNA repair such as *BRCA2* and *WRN*.

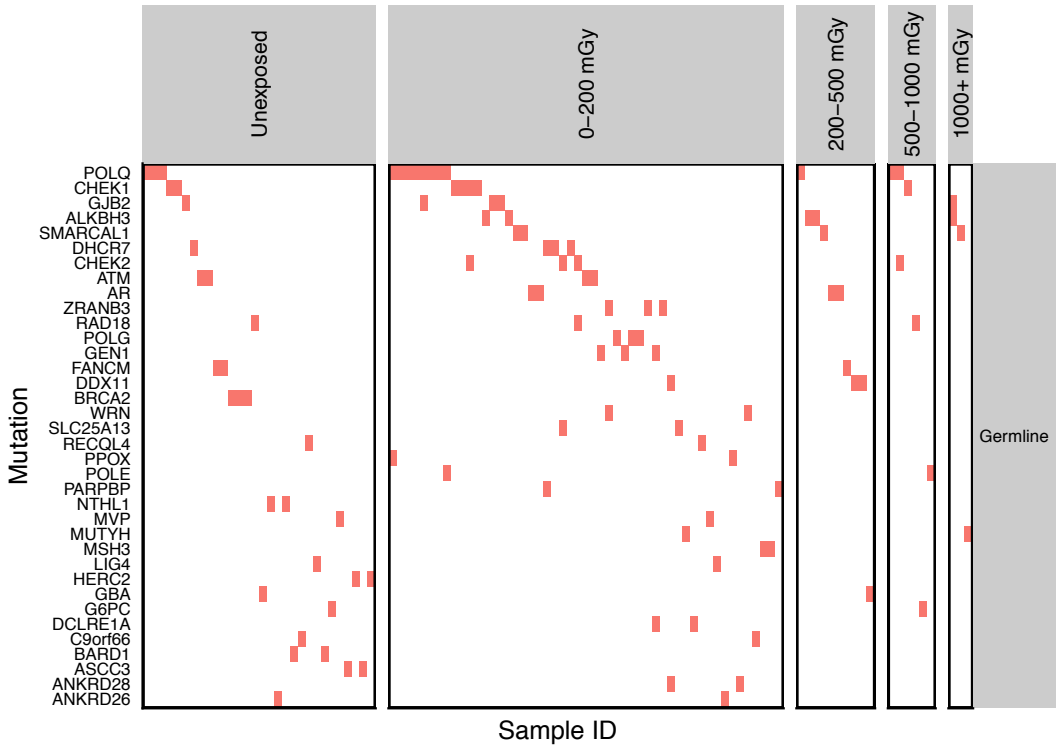


Fig. 2.30 Genes which are mutated by a frameshift, missense, or stop-gain mutation in at least two samples.

One sample possessed two pathogenic loss-of-function germline mutations in *ATM* and likely has clinical ataxia telangiectasia (AT). AT is characterized by neurological symptoms, radiation sensitivity, and predisposition to various cancers [207]. This individual was not exposed to radiation and was diagnosed with PTC at age 16. This was the only salient example of a clear predisposition disorder in the REBC-Unexposed sample set. Further analysis beyond the scope of this work — including integration with epigenetic data — is needed to better understand the germline-somatic interactions within these sample sets.

2.6 Structural variation

In collaboration with Dr. Stewart, I analyzed the structural variant spectrum of the REBC dataset using a consensus calling pipeline based on SvABA, dRanger, Manta, and

snowman. Outputs from these programs were harmonized and merged to a common format, realigned around breakpoints by the Broad's BreakPointer tool, and then annotated with Oncotator to produce a MAF-like file.

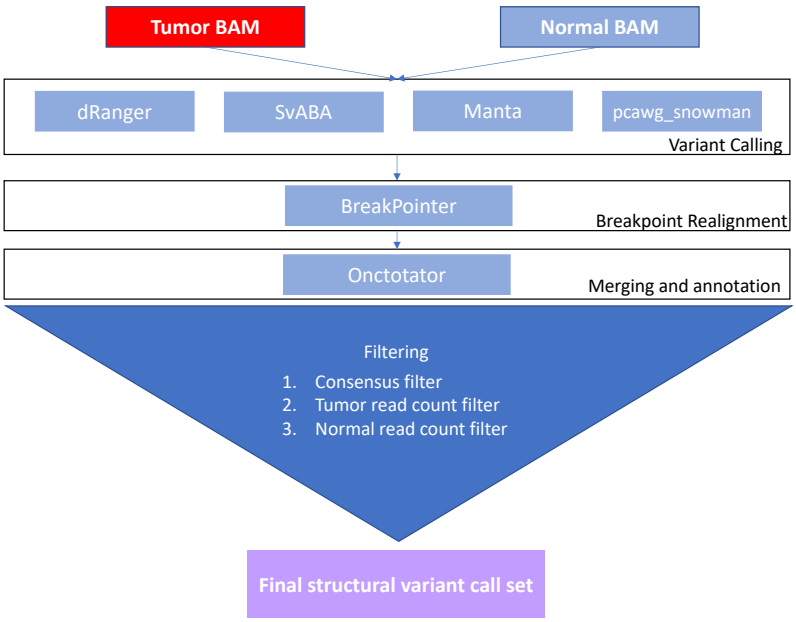


Fig. 2.31 Structural variant calling pipeline.

The raw union SV callset contained roughly 86,000 variant calls. These were filtered by a similar consensus scheme to small variants. Each caller recieved one vote, but pcawg-snowman and SvABA could contribute a maximum of one vote together. Two votes were required to pass a variant. Variants were then filtered to remove any with fewer than 10 tumor supporting reads or 1 or more alt-supporting read in the normal. All samples excluded from SSV analysis (i.e., those with low purity or failing quality control) were excluded from SV analysis, except where explicitly noted.

2.6.1 Structural variants are common in papillary thyroid cancers

After filtering, 256 REBC-Exposed samples (84%), 57 REBC-Unexposed samples (73%) and 31 THCA samples (62%) had at least one SV call. Most tumors harbored fewer than three structural variants regardless of sample set (THCA median = 1; REBC-Exposed median = 2; REBC-Unexposed median = 2).

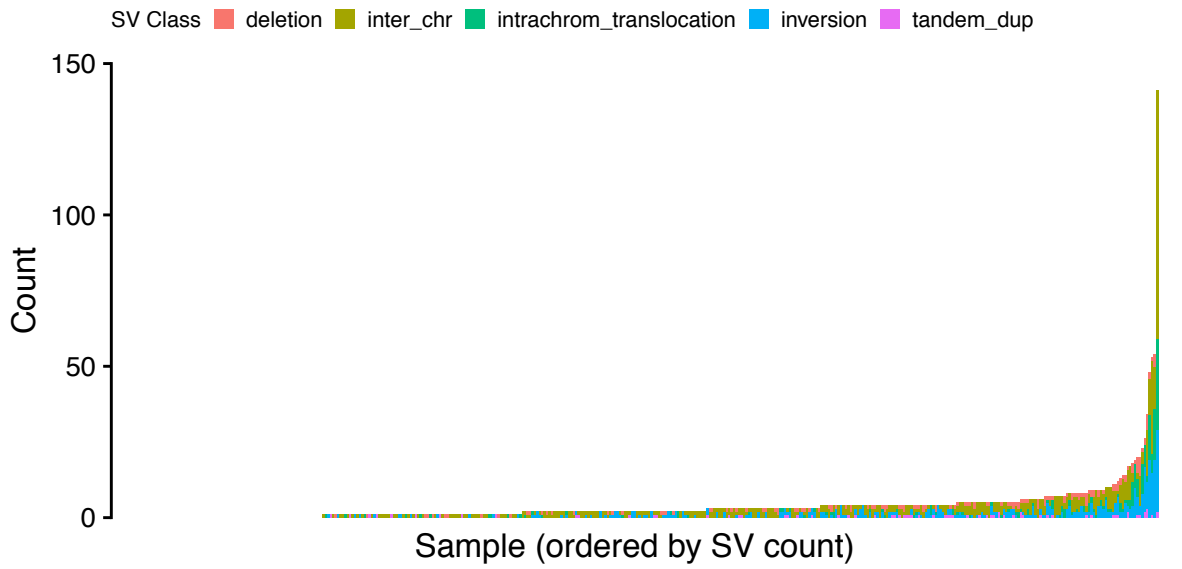


Fig. 2.32 Structural variant counts per sample, colored by SV type. Most tumors have few SVs.

2.6.2 The number of rearrangements is associated with both age and radiation dose

In the combined REBC and THCA dataset, age at surgery, radiation dose, and the number of SSVs per sample were all associated with the total number of structural variants per genome (Poisson generalized linear model; p = 8×10^{-8} for dose; p = 1.6×10^{-5} for age at surgery; p = 0.006 for number of SSVs; SV and SSV hypermutator samples excluded). The same was true when using dose category rather than continuous dose (poisson generalized linear model; dose category p = 1.4×10^{-11}; age at surgery p = 9.4×10^{-5}; number of SSVs p = 0.002). The number of structural variants generally increased with increasing radiation dose category (Figure 2.33) It has been shown previously that somatic structural variants accumulate with age in blood [215, 216].

Only the relationship between the number of structural events and dose was consistently robust to stratification by age and sample set. For samples with a reported age at surgery less than 25 only dose was still significant (n = 123; $p = 8 \times 10^{-8}$). This subset of the data also had good class balance between exposed and unexposed samples and the various dose groups. Between 25 and 45 years of age, all three predictors are significantly associated with the number of SVs (Table 2.5). As only two REBC samples were older than 45 years of age, we did not have sufficient data to make inferences about associations

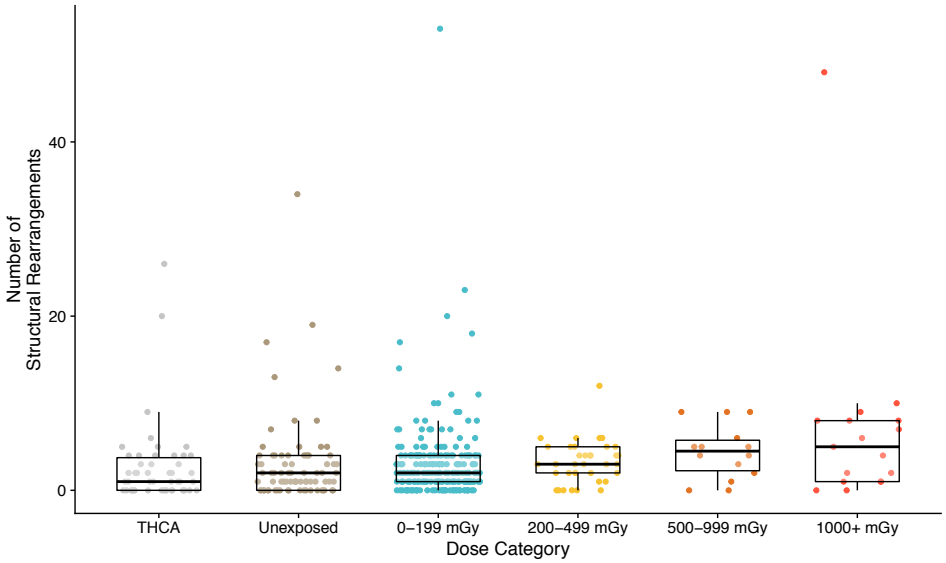


Fig. 2.33 Structural variant counts per sample, separated by dose category. The number of structural variants generally increases with increasing dose.

Sample Set(s)	Number of Samples	Age at Surgery	Number of mutations	Truncated Dose
REBC and THCA	429	$p = 1.6 \times 10^{-5}$	$p = 0.006$	$p = 8 \times 10^{-8}$
REBC (Age $<$ 25)	123	$p = 0.07$	$p = 0.22$	$p = 8 \times 10^{-5}$
REBC (25 $<$ Age $<$ 45)	269	$p = 5 \times 10^{-5}$	$p = 9 \times 10^{-9}$	$p = 2.5 \times 10^{-16}$
REBC (Dose $>$ 0)	302	$p = 9.5 \times 10^{-7}$	$p = 0.28$	$p = 1.54 \times 10^{-14}$

Table 2.5 Table of p-values and sample sizes for GLM models, stratified by sample set and age at surgery.

in this range. Among the 302 exposed samples (excluding any hypermutators), both age at surgery and dose were associated with the number of structural events (Table 2.5).

These results suggest that radiation is inducing structural variant formation in exposed samples. Structural variants also accumulate with age, as described above; however, age-related accumulation is not significant across all age ranges. It is possible that the same mechanism generating small deletions is responsible for generating large structural variants. In subsection 2.6.7 I will explore whether gene fusions — the functional products of certain structural variants — are also associated with increasing dose.

2.6.3 Chromoplexy is frequent in papillary thyroid carcinomas

Three THCA samples and 20 REBC samples (6% of both sample sets) showed evidence of chromoplexy. Chromoplexy was initially described in prostate cancer and has recently been described in thyroid cancer as well as other neoplasms [69, 217]. It is thought that chromoplexy is a result of DNA breakage and erroneous repair around spatially localized transcription forks [1]. Reference [217] observed chromoplexy in approximately 16% of 2,778 pan-cancer samples.

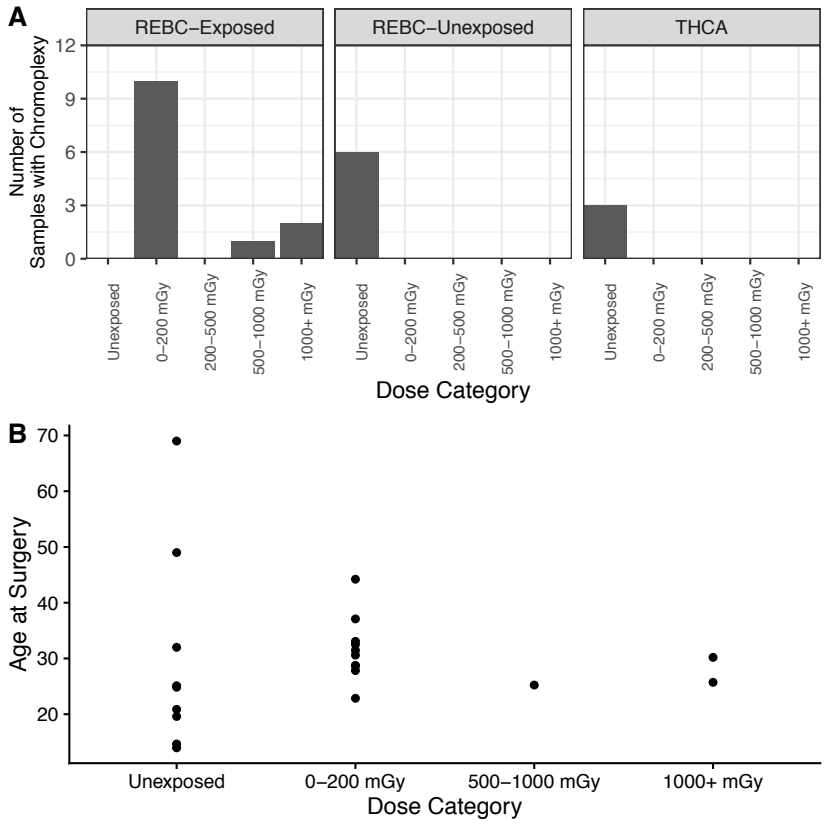


Fig. 2.34 Overview of samples with chromoplexy. (A) Number of samples with apparent chromoplexy, broken down by dose category. (B) Samples with apparent chromoplexy, with age at surgery plotted on y-axis and dose category on the x-axis.

Similar rates of chromoplexy occur across dose categories (Figure 2.34A). Rates of chromoplexy did not appear to be associated with either age at surgery or dose (Figure 2.34B).

One REBC-Unexposed sample exhibited an SV hypermutator phenotype, with three times more calls than any other tumor, caused by an apparent chromoplexy event involving chromosomes 1, 2, 10, 13, and 18. The individual with this tumor was not

exposed to radiation and had no clear germline predisposition variant that would explain its high number of SV calls.

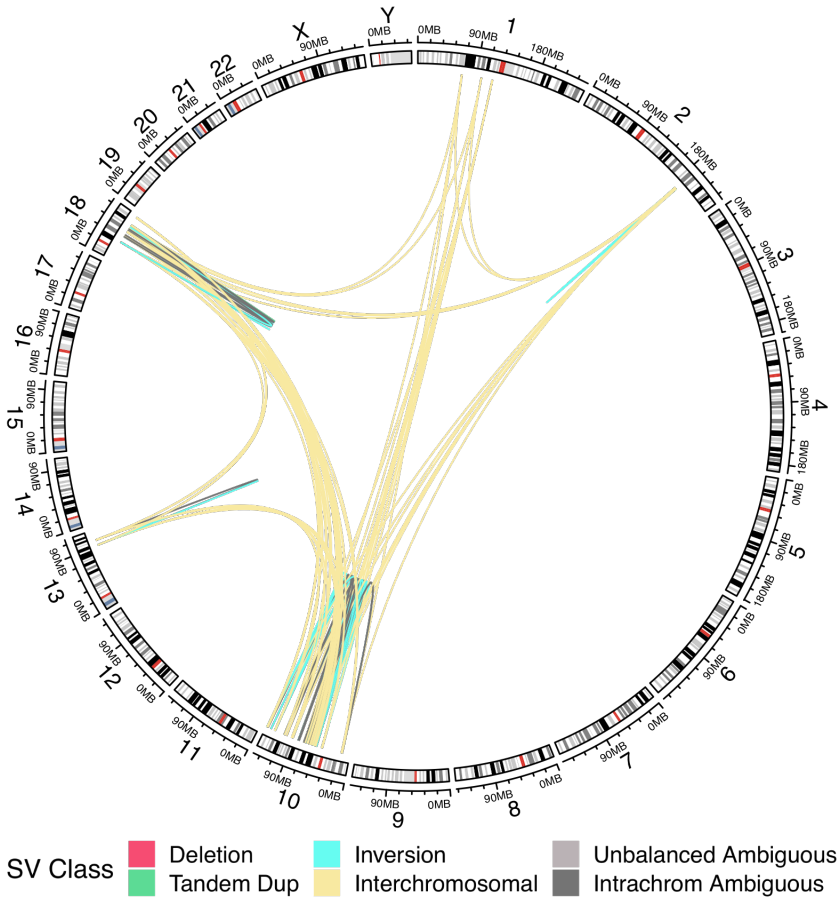


Fig. 2.35 Circos-style ideogram of chromosomes involved in structural variants in the SV hypermutator sample. This sample had 141 reported structural events (82 interchromosomal, 30 intrachromosomal deletions or translocations, 27 inversions and 2 duplications). The large number of interchromosomal and inversion events on chromosome 10 (particular the long arm) suggests a major chromoplexy event between chromosomes 1, 2, 10, 13, and 18.

The SSV (small variant) hypermutator sample has the second highest number of structural variants among all samples. This sample has lost portions of chromosomes 2, 3, 9, 11, 13, and 17. Chromosome 17 in this sample has 27 called structural variants including tiled inversions and intrachromosomal translocations with occasional changes in copy number. BreakPointer frequently reported microhomology at these breakpoints. This sample also carries a somatic $TP53^{R273C}$ mutation, which may partially explain its

genomic instability. The SSV hypermutator was exposed to radiation, though it had a relatively low exposure of 160 mGy.

2.6.4 Whole genome CNA signatures

Two samples displayed a whole genome copy number aberration (CNA) signature similar to one previously described in Hürthle cell carcinomas [218] (Figure 2.36). Hürthle cell carcinoma (HCC) is generally considered a variant of the follicular subtype of thyroid cancer and not a papillary carcinoma [219]. Reference [218] reported that more than half of HCC tumors sequenced displayed a near-haploid state with consistent preservation of both copies of chromosomes 5, 7, 12 and 20. These chromosomes are conserved at higher copy number in both whole-genome copy number aberration cases we observed. Both samples were female, received an estimated dose of 150 and 165 mGy, and had IREP 50^{th} percentile values of approximately 50%.

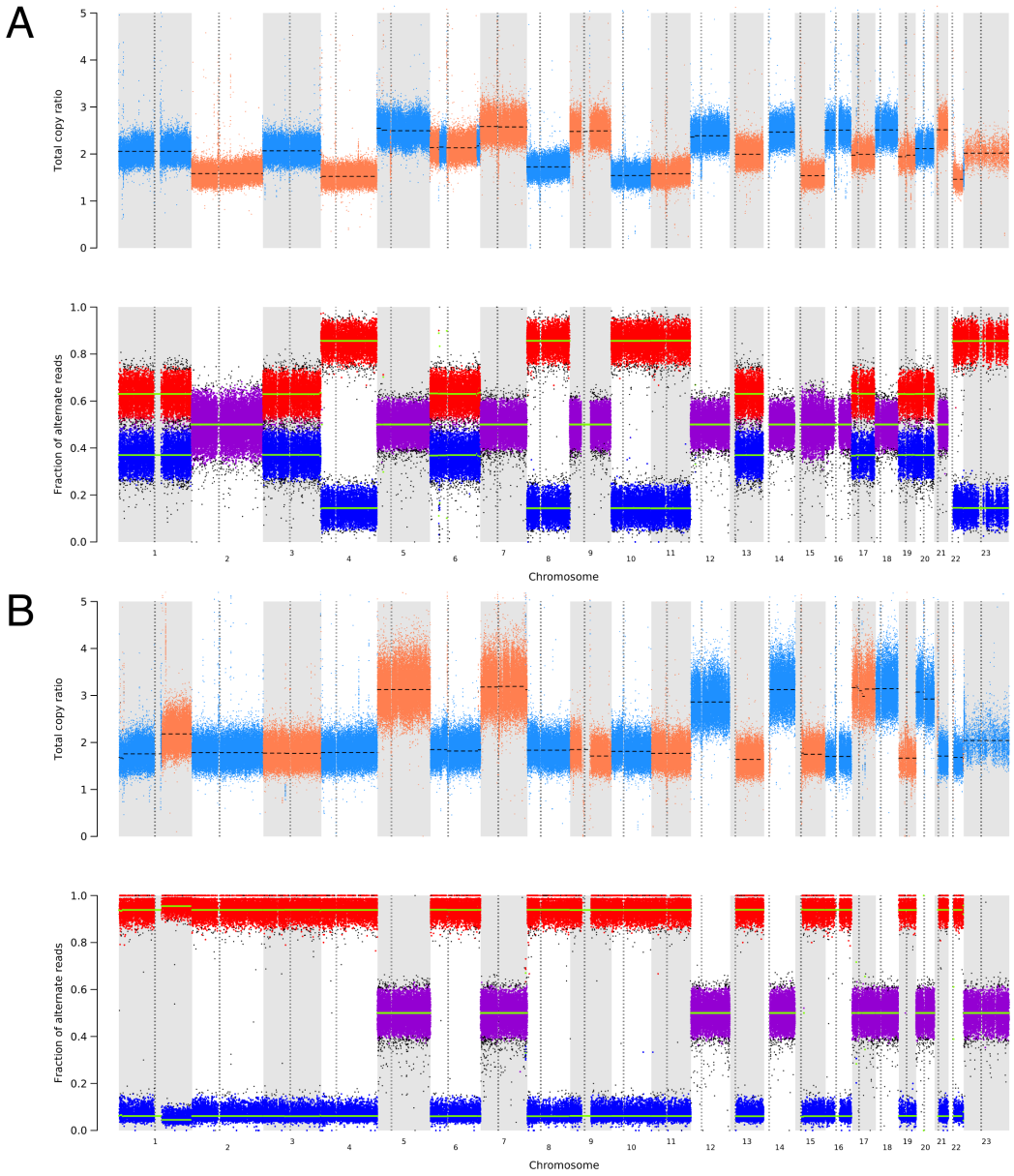


Fig. 2.36 Two samples showed whole genome copy number aberration signatures consistent with previously reported signatures in Hürthle Cell Carcinoma. This signature has been reported as a genome-halving event, where one copy of most chromosomes are entirely lost with the exception of chromosomes 5, 7, 12 and 20, which are preserved in a diploid state. (A) and (B) show total plots of copy number estimates (top) and allele fraction (bottom). (A) A sample with significant allele ratio shifts for most chromosomes. Total copy number estimates are shifted as well but appear to be confounded by sample purity, which for this sample was less than 50%. (B) A second sample with clear loss of most chromosomes except those canonically preserved in some Hürthle Cell Carcinoma samples. This sample had an estimated tumor purity of greater than 85%.

There is some debate about whether the copy number alterations observed in Hürthle cell tumors is a genome-halving or genome-doubling event (private communication, Dr. Chip Stewart and Dr. Gad Getz). The tumor copy number profile shown in Figure 2.36A is for a tumor with less than 50% purity, making proper copy number determination difficult. While the tumor profile shown in Figure 2.36B comes from a tumor with estimated purity $\geq 90\%$ and is more stable, it appears to show an increase in the copy number of the characteristic HCC chromosomes rather than a loss. Pathology data from these tumors did not support a histologic classification as Hürthle cell carcinoma, however the similarity of the genetic CNA profiles to those described in reference [218] warrants further examination of Hürthle cell carcinoma CNA activity.

2.6.5 Inversions and translocations are common in thyroid carcinomas

Certain large variant classes were more common in specific sample sets. One hundred and twenty-eight tumors (42%) in the REBC-Exposed sample set had at least one inversion compared to 32% (25) of REBC-Unexposed samples and 25% (13) of THCA samples.

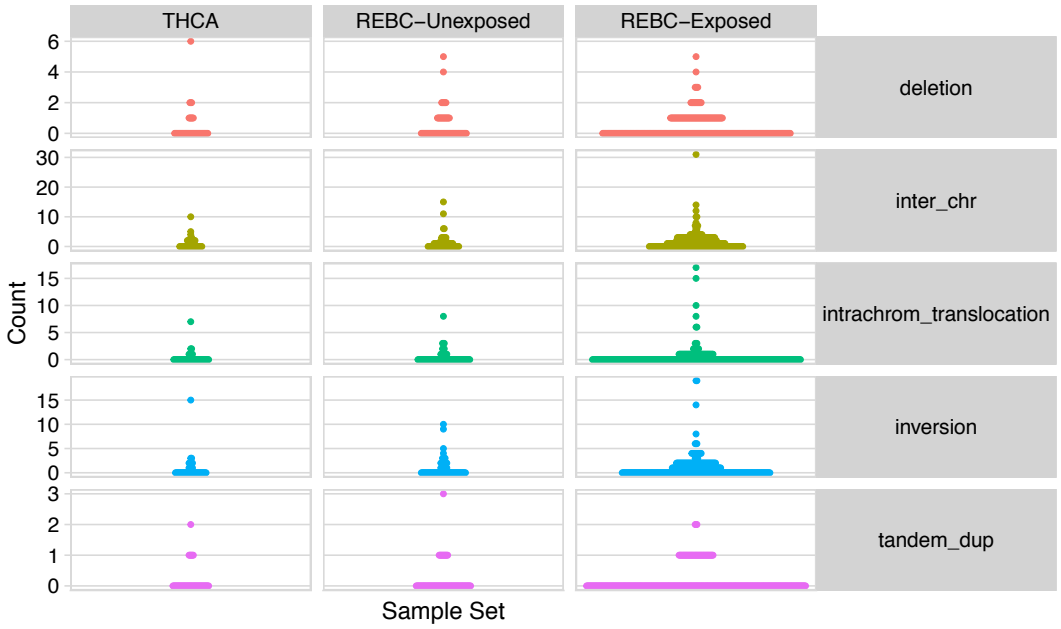


Fig. 2.37 Structural variant counts per sample broken down by variant class and sample set. The SV hypermutator sample has been excluded. In general, tumors in the REBC-Exposed set contain more interchromosomal and inversion variants. Deletion and duplication counts per tumor are similar across sample sets. Intrachromosomal translocations, which may actually represent inversions, non-tandem duplications, or copy-neutral intrachromosomal exchanges but are not reported unambiguously by the SV detection programs, are again elevated in the exposed group.

A higher proportion of REBC-Exposed tumors harbored an interchromosomal translocation compared to unexposed tumors. Sixty-three percent (192 of 303) tumors harbored at least one interchromosomal variant. Fifty-two percent of REBC-Unexposed samples (41 of 78) had at least one interchromosomal variant compared to forty-six percent (23 of 50) of THCA samples.

Balanced inversions and translocations have previously been implicated as possible signatures of radiation exposure [54]. Radiation induces single- and double-strand breaks in the DNA [170]; this damage and subsequent repair has been hypothesized to cause structural variant formation as well as formation of small deletions.

I tested the association between dose and the number of each SV class when correcting for age at surgery, the number of SNVs and the number of small indels. Dose was not associated with an increase in the number of deletions (poisson glm, REBC-Unexposed and REBC-Exposed sample sets excluding SSV and SV hypermutators, number of SNVs $p = 0.28$; number of indels $p = 4.3 \times 10^{-7}$; age at surgery $p = 0.0005$; dose truncated to 1000 mGy $p = 0.45$). Dose was significantly associated with the number of inversions

(poisson glm, number of SNVs $p = 7.8 \times 10^{-8}$; number of indels $p = 0.026$; age at surgery $p < 2 \times 10^{-16}$; dose truncated to 1000 mGy $p < 2 \times 10^{-16}$); it was also associated with the number of interchromosomal variants (poisson glm, number of SNVs $p < 2 \times 10^{-16}$; number of indels $p < 2 \times 10^{-16}$; age at surgery $p < 2 \times 10^{-16}$; dose truncated to 1000 mGy $p = 4.6 \times 10^{-5}$). While dose was associated with the number of tandem duplications the low number of duplications both per sample and across the sample sets necessitates skepticism about this result (poisson glm, number of SNVs $p = 0.028$; number of indels $p = 0.046$; age at surgery $p = 1.44 \times 10^{-8}$; dose truncated to 1000 mGy $p = 3.16 \times 10^{-7}$). As intrachromosomal translocations are ambiguously classified, we chose not to test their association with dose. These results are in line with those described in subsection 2.6.2; increased radiation dose appears to be associated with an increase in the number of structural variants. It is difficult, however, to attribute any single genomic event to a radiation strike. Reference [54] reported high levels of microhomology and balanced copy number around structural variants in tumors that were exposed to radiation, as well the increase in the indel to SNV ratio we observed in subsection 2.3.2. Work to characterize the individual breakpoints and genomic contexts of the variants in this cohort is planned as a followup to publications describing the overall mutational landscape.

2.6.6 Chromosome 10 is commonly involved in structural variants in PTC

Structural variants are not distributed evenly across the genome in PTC cases (χ^2 goodness-of-fit test; $p < 0.001$). Chromosome 10 harbors a disproportionately high number of events compared to all other chromosomes.

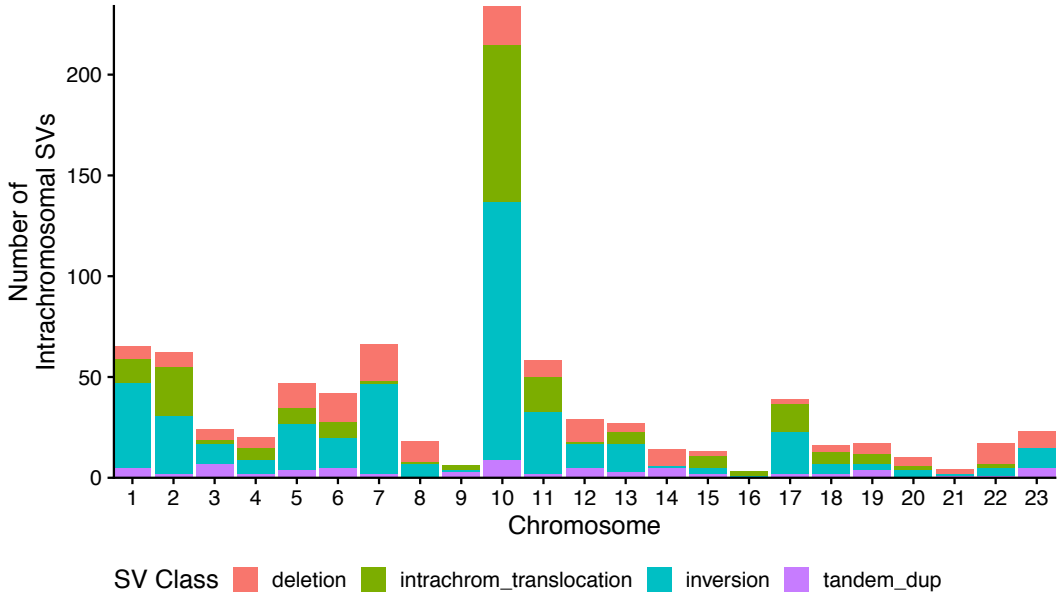


Fig. 2.38 Counts of intrachromosomal structural variants per chromosome.

Chromosome 10 had the highest number of intrachromosomal events (Figure 2.38). It also had many more inversions and intrachromosomal translocations. Figure 2.39 shows the chord diagram of interchromosomal structural variants across all three sample sets. Again, chromosome 10 is recurrently impacted by structural variants involving multiple chromosomes, although chromosomes 1, 2 and 12 have similar numbers of interchromosomal variants. Chromosome 12 and 15 were the most common interchromosomal partners, most often resulting in a gene fusion between $ETV6$ and $NTRK3$.

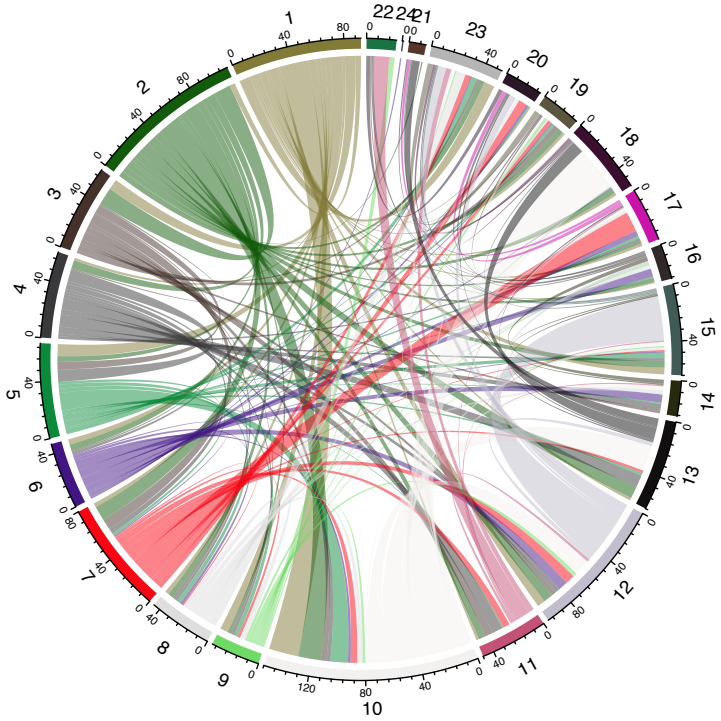


Fig. 2.39 Chord diagram of chromosomes involved in interchromosomal structural variants. Arc lengths and arc labels correspond to the number of structural events per chromosome. There is great diversity in chromosome partners across the sample sets; variants pair every autosome with at least three others. Chromosomes 1, 2, 10 and 12 are the most impacted, though this is partially driven by chromoplexy events (including in the SV hypermutator).

To correct for any samples with multiple events on chromosome 10 that might inflate this number I also plotted the number of samples which reported an event on a given chromosome (Figure 2.40).

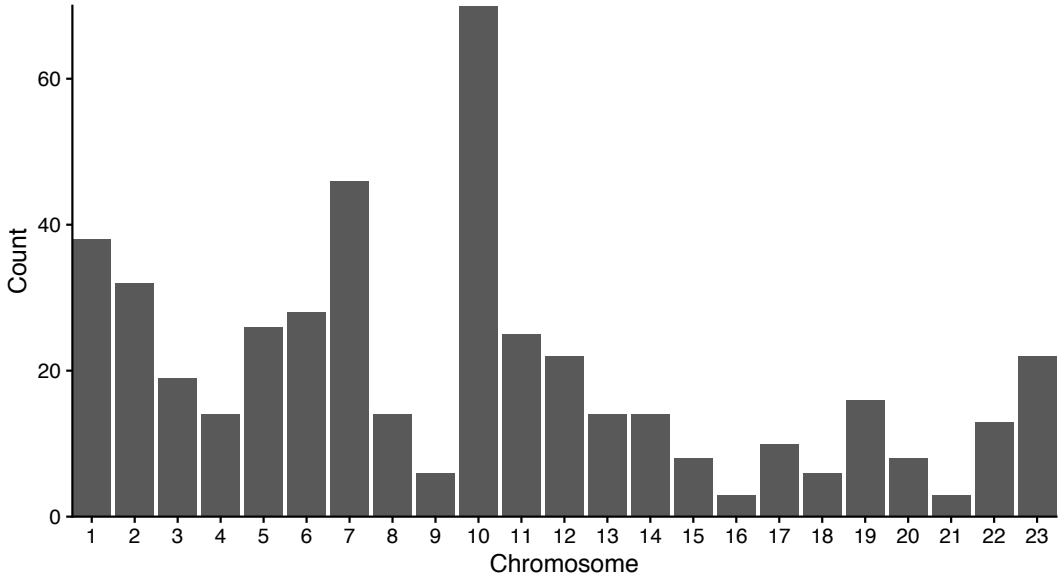


Fig. 2.40 Counts of samples with at least one intrachromosomal event per chromosome.

Seventy samples (48 REBC-Exposed, 13 REBC-Unexposed, 9 THCA) had at least one lesion on chromosome 10. There are several reasons for the prevalence of structural variants on chromosome 10. In samples with at least one structural variant on chromosome 10, 32 REBC-Exposed (66%), 8 REBC-Unexposed (61%), and 7 THCA (77%) samples had a structural variant that impacted *RET*. The *RET* proto-oncogene resides at position 10q11.2. *RET* is commonly rearranged in papillary thyroid carcinoma, especially among cases that are exposed to radiation [220, 167]. *RET* fusions are potent drivers of papillary thyroid carcinoma. There is some evidence that the 10q11.2 region of chromosome 10 may be more sensitive to radiation compared to the rest of the genome [221].

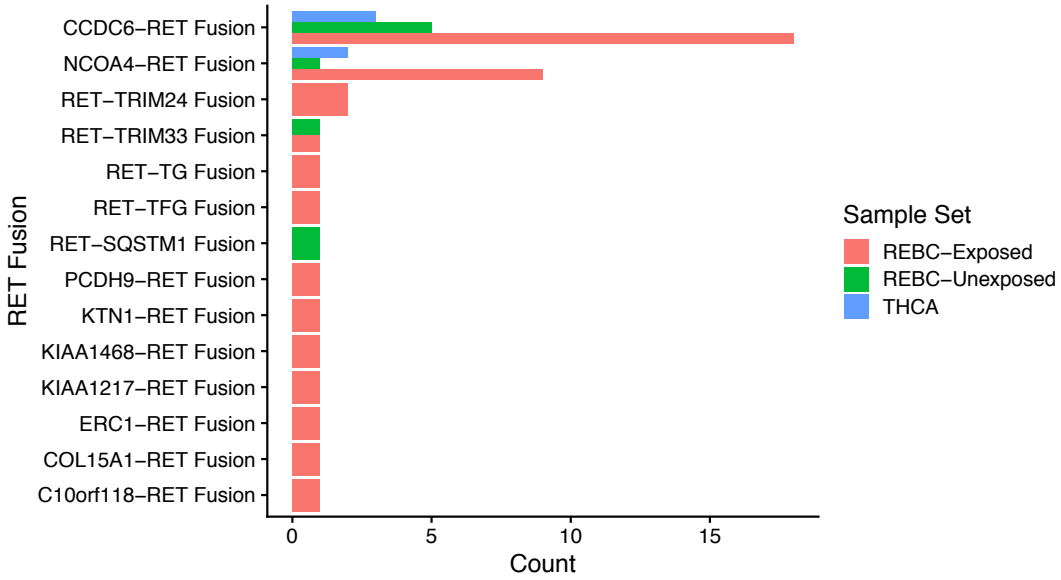


Fig. 2.41 Counts of *RET* fusions in our sample sets.

*RET-CCDC*6 was the most common *RET-PTC* fusion across all sample sets. Also called *RET-PTC*1, *RET-CCDC*6 is the most common *RET* fusion reported in studies of papillary thyroid carcinoma [220]. *RET-CCDC*6 comprises up to 70% of the observed *RET* rearrangements in some studies.

We observed other previously reported *RET* gene fusions as well as novel fusions. *RET-NCOA*4 (*RET-PTC*3) was the second most common fusion in our study. *RET-TRIM*24 and $RET - TRIM33$ have been previously reported and we observe these in our dataset. We observe a novel fusion between *RET* and *TG*. This fusion conserves the kinase domain beginning at exon 11 like all other observed *RET* fusions, and it is likely functional.

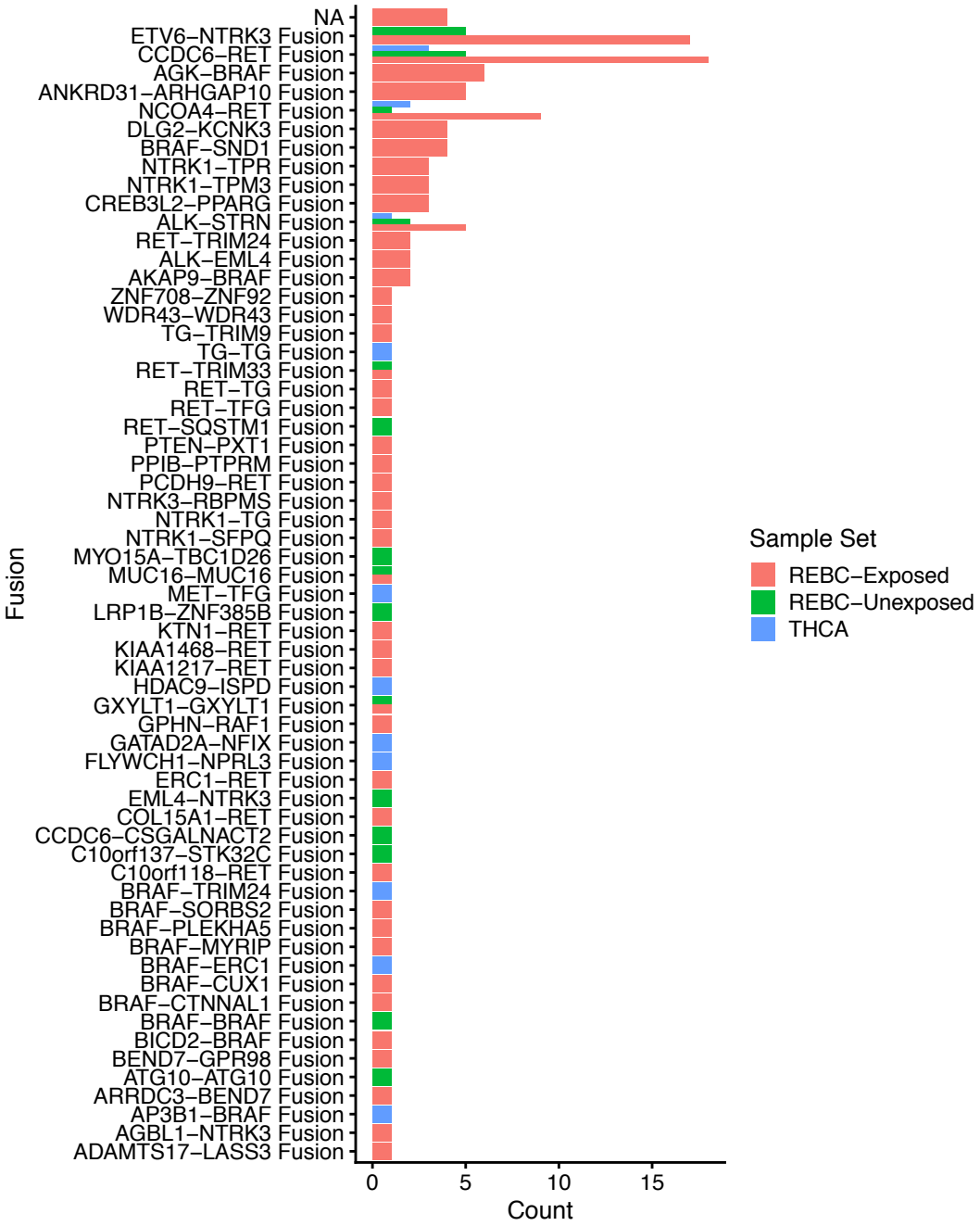


Fig. 2.42 Counts of all gene fusions in our sample sets.

2.6.7 Gene fusions are common in radiation-associated PTC

We corroborated DNA fusion calls with RNA calls where data was available. Our pipeline showed very high concordance rates for gene fusion calls between RNA and DNA. Of a total 168 RNA fusion calls, 137 (81.5%) were called as in-frame fusions by the DNA

pipeline and an additional 10 (5.9%, total: 87.5%) were called but labeled out-frame-fusion (not shown). The high concordance rates observed between the two orthogonal sequencing strategies and pipelines suggest that our DNA data is of sufficient sensitivity and specificity for further analysis.

Although *RET* was the most common fusion partner many other gene fusions were observed (Figure 2.42). [53] reported gene fusion drivers in 15.3% of their 484 whole-exome sequenced samples. Gene fusions have been reported at much higher rates in studies of children exposed to ionizing radiation, particularly the *RET-PTC* family of fusions, which are present in 50-85% of samples in some studies [222, 223].

The REBC-Exposed sample set displayed a high proportion of gene fusions, many of which are known drivers. Thirty-three percent (102 of 303) radiation-exposed samples had at least one in-frame gene fusion. Twenty seven percent of REBC-Unexposed and twenty-eight percent of THCA samples had at least one gene fusion. However, examining exposure in a binary fashion obscures the fact that different dose groups showed much larger differences in the proportion of samples with at least one gene fusion.

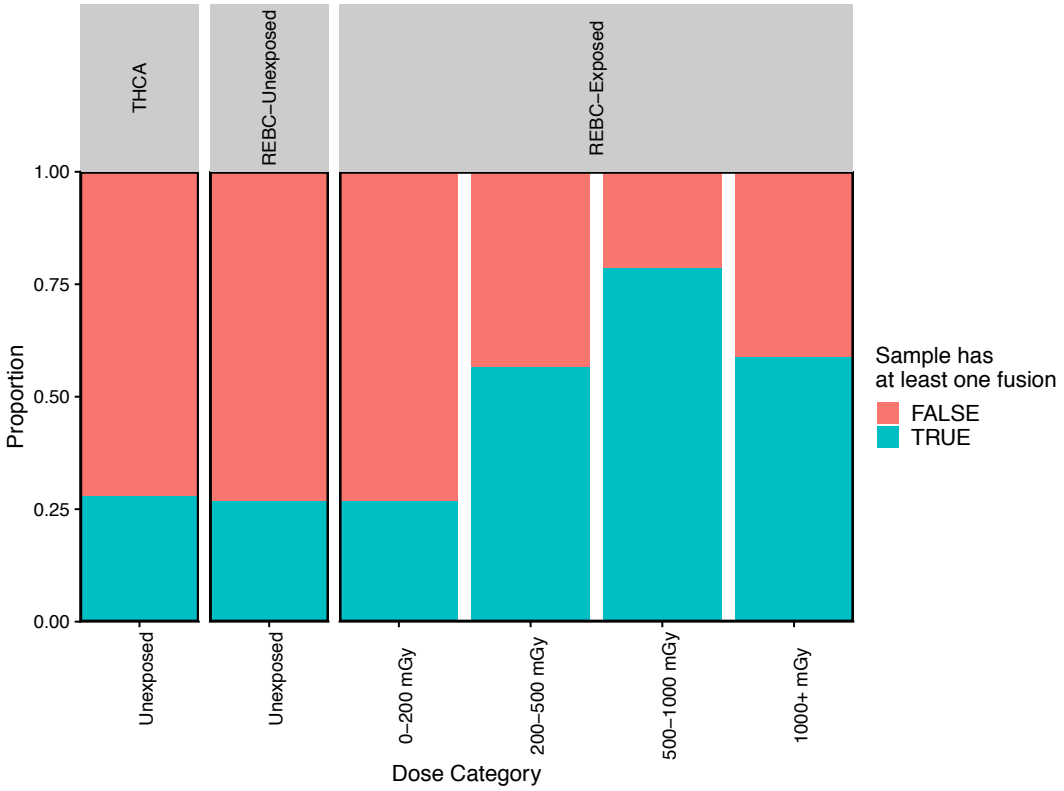


Fig. 2.43 Proportion of gene fusions by dose category. The proportion of observed gene fusions tends to increase with increasing dose category. The proportion of gene fusions ranges from roughly 25% in unexposed groups to roughly 75% in samples exposed to between 500 and 1000 mGy.

There was a positive association between radiation dose and whether a sample had at least one gene fusion when correcting for age at surgery, the number of SNVs, and the number of indels (binomal logistic glm; restricted to REBC sample set; age at surgery $p = 0.0001$; number of SNVs $p = 0.27$; number of indels $p = 0.12$; dose $p = 0.0096$). As most samples had only one fusion, we were not able to test whether the number of fusions per sample was associated with dose via Poisson regression. In line with the association between fusion status and dose, the proportion of samples with at least one gene fusion increases with categorical dose. At 0-200 mGy, roughly the same proportion (26.8%) of samples in the REBC-Exposed set have a gene fusion compared to the REBC-Unexposed and THCA sets. However, this proportion increases to 50% of samples in the higher dose groups, with more than 75% of tumors in the 500-1000 mGy dose group possessing a gene fusion. Reference [223] did not report the doses their samples received. Reference [222] found *RET* rearrangments in 84% of observed irradiated thyroid carcinomas, compared to 15% of sporadic tumors. These medically-irradiated tumors received significantly

higher doses than tumors in our study. Overall, the proportions of tumors in each of the REBC sample sets corroborates the results of past studies of radiation-associated thyroid carcinomas as well as the association between dose and the number of rearrangements described in subsection 2.6.2.

2.6.8 Breakpoints in fused genes are recurrent

We would expect structural variant breakpoints around a radiosensitive site to be highly recurrent. Recurrent breakpoints will also occur where the gene fusion produced is functional and confers selective advantage.

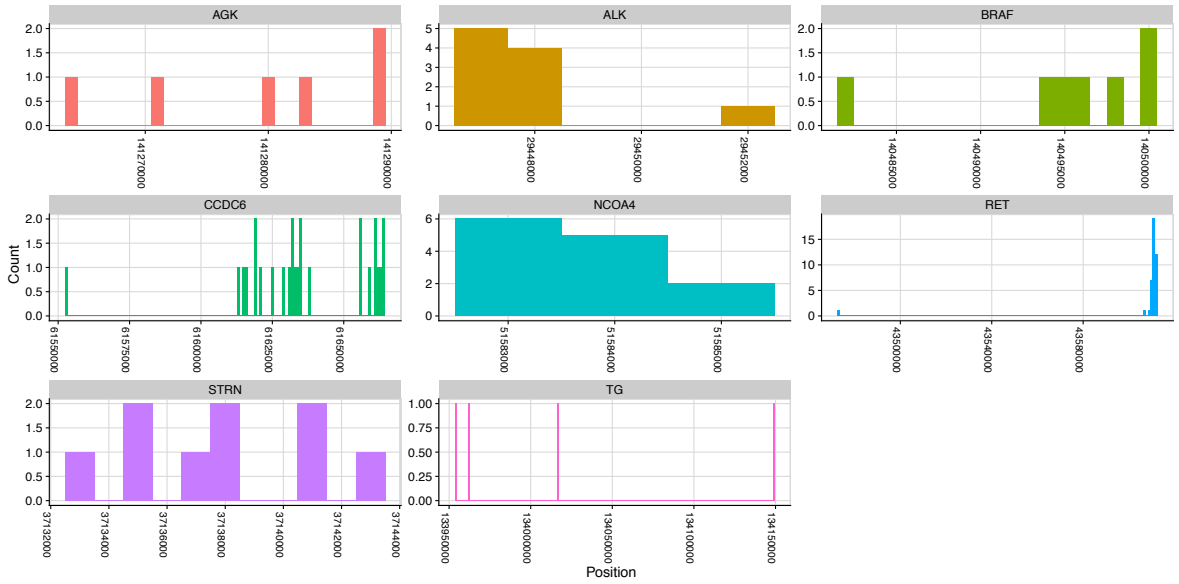


Fig. 2.44 Genomic positions of gene fusion partners for the ten most fused genes. Each bin is 1kb wide.

Breakpoints in *RET* were highly recurrent, with 32 breakpoints occuring within a single 3kb span. This region falls immediately before the eleventh exon of *RET*. Active *RET* fusions always occur in this region [224]. Transcribed fusions always include the kinase domain but lack the transmembrane and extracellular domains [220].

2.7 Papillary thyroid tumors tend to be driven by one or few mutations

Figure 2.45 shows the overall driver landscape of REBC papillary thyroid carcinomas. Most tumors have only one driver mutation. This is in line with patterns established in pediatric tumors [45] and previous studies of papillary thyroid carcinoma [53]. Gene fusions and SSV drivers are mutually exclusive, as are individual mutations in the *MAPK* pathway.

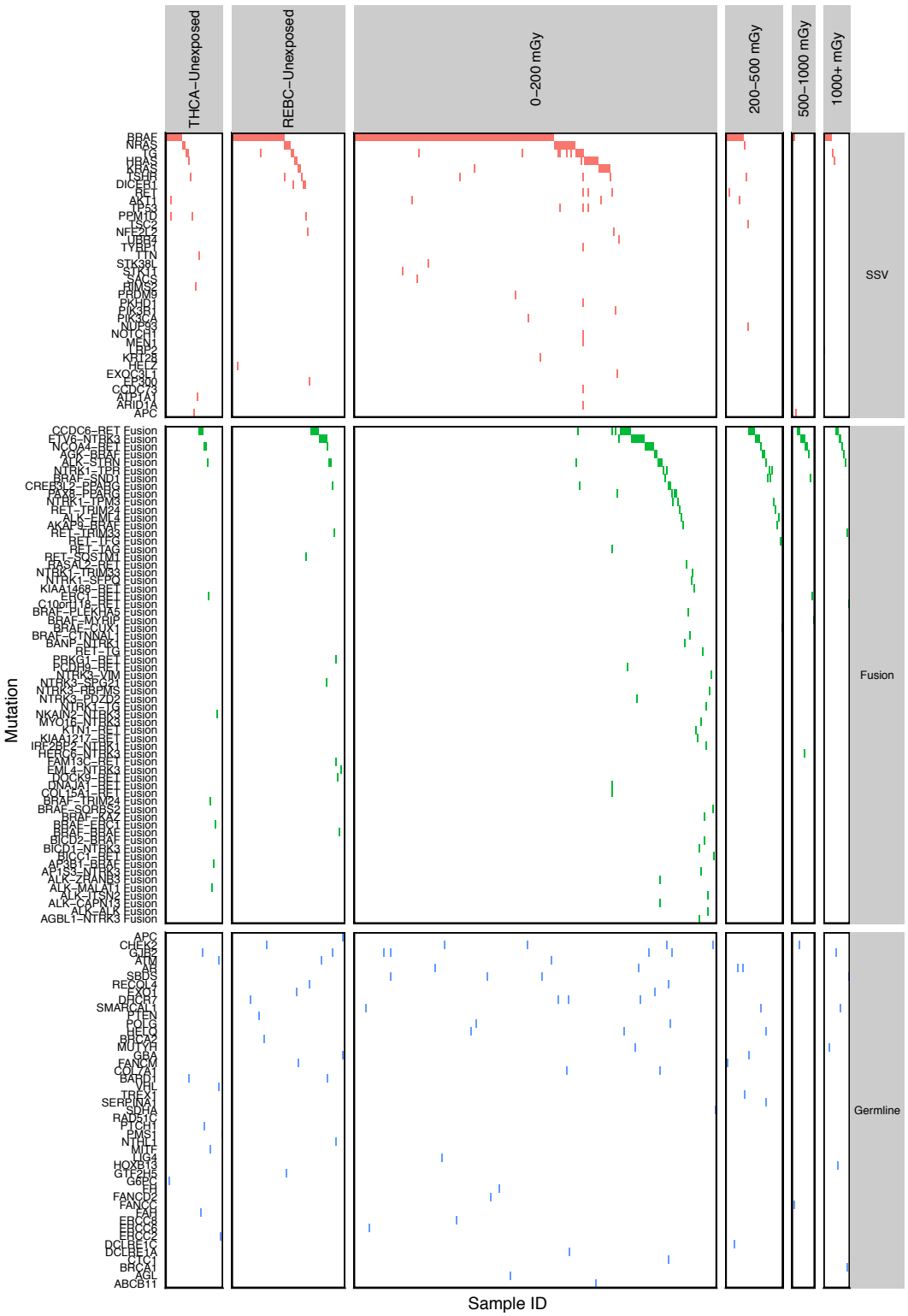


Fig. 2.45 Drivers of raPTC broken down by sample set and radiation exposure group.

Sample Set	% *BRAF* driver	% *NRAS*	% *HRAS*	% *KRAS*	Total *MAPK* %
THCA	20	4	2	0	26
REBC-Unexposed	41	5.1	2.6	2.6	51.3
REBC, 0-200mGy	51.9	5.5	4.3	3.4	65.1
REBC, 200-500mGy	29.7	2.7	0	0	32.4
REBC, 500-1000mGy	14.3	0	0	0	14.3
REBC, 1000+mGy	29.4	0	5.9	0	35.3

Table 2.6 *MAPK* mutation rates across dose groups. Higher dose groups tend to have a lower proportion of samples with *MAPK* SSV drivers.

Chromosome 22 loss, which has previously been implicated in thyroid carcinoma, does not appear to be sufficient to drive tumor progression. All samples with a reported chromosome 22 CNA also had another driver mutation.

As described in subsection 2.2.3, the majority of our samples carry a *BRAF*, *NRAS*, *KRAS* or *HRAS* mutations, all of which are well-established potent drivers of thyroid carcinoma. 172 (56.7%) of exposed samples and 40 (51.2%) of unexposed REBC samples had SSV drivers in the *MAPK* pathway. While only 26% (13 / 50) THCA samples had an SSV *MAPK* driver, the rate of *MAPK* SSV drivers in the overall dataset described in [53] was 61.7%.

The proportion of samples with a *MAP* SSV driver varied by radiation dose group (Table 2.6). As dose increased the proportion of samples with a *MAPK* SSV driver generally decreased, with a corresponding rise in the number of gene fusion drivers (Figure 2.46).

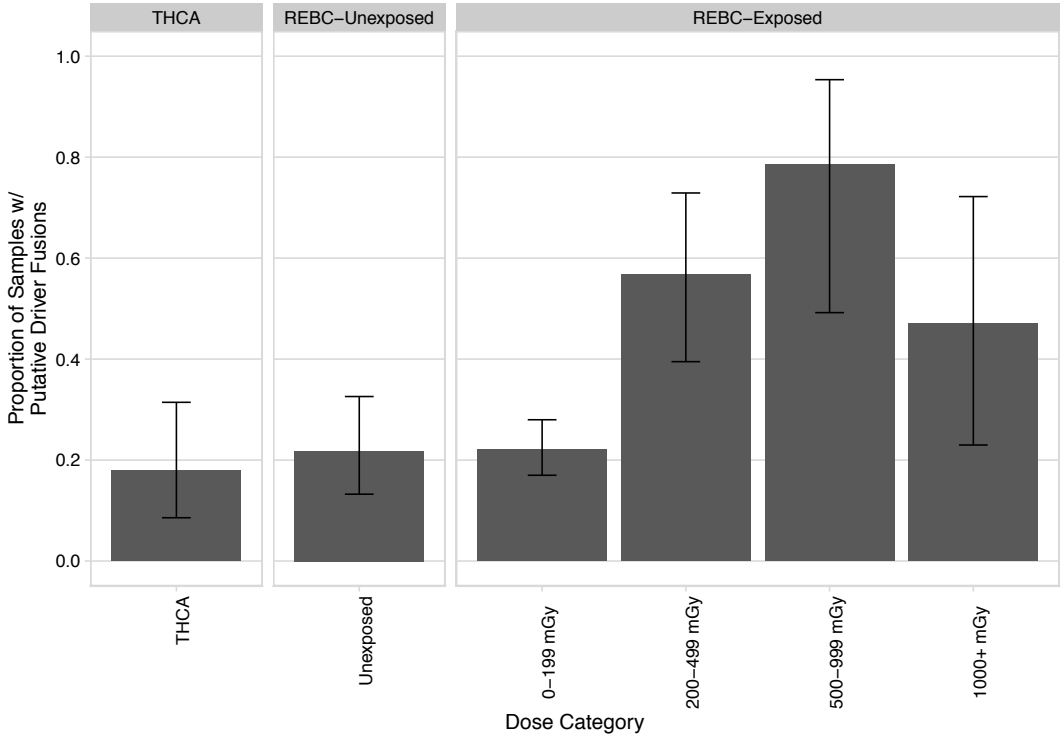


Fig. 2.46 Proportion of samples with a putative SV-induced gene fusion driver. This figure is essentially a reproduction of Figure 2.43, with the additional restriction that fusions in this plot must be putative drivers and with the inclusion of error bars (binomial test). Again, an increase in the proportion of gene fusions is observed with increasing dose group. The unexposed groups (both REBC and THCA) as well as the lowest exposure group (<200 mGy) have approximately the same proportion of gene fusions.

2.8 TERT promoter mutations are rare in pediatric PTC

Mutations in the promoter of the *TERT* gene are associated with worse prognosis and less-differentiated tumors in thyroid cancer [225, 226]. *TERT* mutations are also typically associated with greater age at diagnosis. I scanned our datasets for any mutations in the TERT gene or promoter regions. We verified that we had sufficient coverage to call mutations in all REBC samples, and Dr. Stewart performed forced calling of the *TERT* promoter region for all 381 REBC samples using `Mutect`; similar analyses were performed for the THCA data in [53]. In our fifty THCA samples, 10 tumors (20%) had canonical *TERT* promoter mutations. This rate is comparable to that found in the entire THCA [53] dataset where *TERT* promoter mutations were found in 36% of

samples. The youngest individual with a $TERT$ mutation in THCA was 49 years of age. Across all 381 REBC samples, only one sample had a canonical $TERT$ promoter mutation. This individual was 40 years old at surgery, among the oldest in our study. All samples in which we located $TERT$ promoter mutations had $BRAF^{V600E}$ mutations.

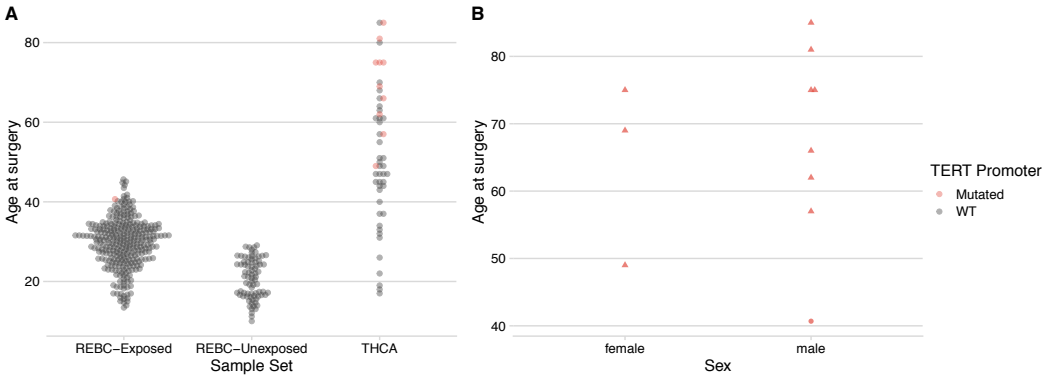


Fig. 2.47 Mutations in the $TERT$ promoter region are associated with age.

These results support previous results indicating that $TERT$ promoter mutations are associated with age. While we could not test for an association with dose, we note that the single REBC sample with a $TERT$ promoter mutation received an estimated dose of 31 mGy, among the lowest in the dataset. The IREP 95% confidence interval of probability of causation due to radiation exposure was [2.7, 31.8], indicating that this individual's cancer was unlikely to be caused by radiation.

2.9 Conclusions

Our study comprises the largest collection of whole-genome sequenced irradiated tumors to date. Exposure to ionizing radiation is a well-established risk factor for development of thyroid cancer. In 381 tumors from the Chernobyl Tissue bank and an additional 50 from The Cancer Genome Atlas, we find striking similarities between sporadic and radiation-associated cases of papillary thyroid carcinoma. We also validate markers of radiation exposure in the genome which were previously suggested but which could not be verified without a study of this size.

2.9.1 The genetic landscape of radiation-associated papillary thyroid carcinoma

Papillary thyroid carcinoma has a low somatic mutation rate compared to all other adult cancers. The median coding mutation count per tumor in our study is lower than that of any other adult cancer in the original 12 TCGA cancer types. There is a clear association between age at surgery and the number of SSVs. This association holds when looking at substitution or indel SSVs alone. Papillary thyroid carcinomas have relatively low mutation rates and are most often driven by a single driver mutation. Our data support a model of thyroid cancer as an intermediate between adult and pediatric cases.

Constitutive activation of the $MAPK$ pathway through either SSVs or gene fusions is the primary mechanism of driver formation in papillary thyroid carcinoma. The $MAPK$ pathway contains the well-known $BRAF$, $HRAS$, $NRAS$, and $KRAS$ genes as well as the RET proto-oncogene that is often involved in gene fusions in radiation-associated cases. While we observe other pancancer drivers in our dataset, these occur in tandem with a $MAPK$ mutation. $MAPK$ mutations appear to be essential to papillary thyroid carcinoma development.

$TP53$, the most commonly mutated gene across all cancer types, is rarely mutated in papillary thyroid carcinoma. Only three tumors of 431 have a protein-altering $TP53$ mutation. $TP53$ mutations were mutually exclusive with $BRAF^{V600E}$ mutations, the most common driver in papillary thyroid carcinoma. $BRAF$ mutation may alter $TP53$ expression via oncomiR-3151, and $BRAF$ mutation alone may be sufficient to alter $TP53$ activity [227]. Unlike $TERT$, all $TP53$ mutants were thirty years old or less at surgery. We did not have sufficient information to test whether $TP53$ mutation was associated with more aggressive disease, though our single SSV hypermutator was among the three samples with a $TP53$ mutation. We did not observe any pathogenic $TP53$ germline mutations, either, though germline $TP53$ mutations in Li-Fraumeni syndrome are associated with an increase risk of thyroid cancer.

No $TERT$ promoter mutations were present in papillary thyroid carcinomas diagnosed before age 40. This suggests that acquisition of $TERT$ promoter mutations is neither necessary nor sufficient for development of PTC. Also missing from our REBC sample set are any mutations in $EIF1AX$, previously reported as significantly mutated in sporadic thyroid carcinoma.

Two genes essential to thyroid function, TG and $TSHR$, are significantly mutated in the REBC sample set. Mutations in these genes always occurred in tumors which also

possessed a driver mutation in one of the *MAPK* genes, indicating that such mutations are not likely drivers on their own. *TG* mutations, many of which are indels, are likely due to a previously-described lineage-specific hypermutational process [191].

*DICER*1, a gene essential in miRNA processing, was frequently mutated in tumors in unexposed cases and was present exclusively in tumors which lacked a *MAPK* driver. *DICER*1 is an established cancer predisposition gene. *DICER*1 syndrome requires only one mutated allele to predispose one to cancer and cause miRNA dysregulation [228]. Our data indicate that somatic inactivation of one allele of *DICER*1 may predispose one to papillary thyroid carcinoma.

The dominant mutational signatures of our dataset are the same as those of sporadic papillary thyroid carcinoma. Signatures SBS2 and SBS13 mark the activity of the *APOBEC* family of cytidine deaminases. *APOBEC* is widely active in tumors [229]; *APOBEC*-induced mutations are present across all three sample sets examined in this thesis. *APOBEC* contributes to the mutational load of individual tumors but is not the primary driver of carcinogenesis.

The majority of mutations in our dataset were attributable to signatures SBS1 and SBS5. Mutations attributed to these two signatures occur in all human cancers and normal cells, often at consistent rates with age [230]. While an underlying mutational process has not been determined it is hypothesized that both are due to endogenous cellular processes and not external exposures. The remaining signature predicted most resembled signatures SBS22 and SBS25, two signatures thought to be related to chemotherapy exposure [32]. Neither signature SBS22 nor signature SBS25 have previously been reported in thyroid carcinoma, and to the best of our knowledge no patients in this dataset received chemotherapy. The signature is ubiquitous across our samples but at very low proportions of mutations. This signature may be a result of overfitting or an exposure common to both exposed and unexposed individuals. The SSV hypermutator shows evidence of defective DNA repair, possibly combined with mutations in DNA polymerases, in its mutational context spectrum.

We did not see an association between radiation dose and the number of SSVs. Substitution SSVs are strongly associated with age, and most of the SSVs in our dataset are attributable to clock-like mutational signatures that accumulate mutations with age. These facts together support the idea that radiation does not induce substitutions in the genome at observable rates for the doses in our study.

Radiation exposure was associated with an increase in the proportion of indels, particulary deletions. Deletions also accumulate with age, and ratios which corrected

for the accumulation of somatic mutations with time were better response variables to radiation dose than raw counts. The deletion:substitution ratio was significantly correlated with radiation dose as well as IREP probability of causation. The deletion:substitution ratio is a potential dose-sensitive genomic marker for radiation exposure.

We were able to fit two indel mutational signatures in the REBC dataset. One of these signatures resembles COSMIC ID8, which has previously been reported as a putative signature of radiation exposure. The indel features of this signature are positively associated with radiation dose in the REBC dataset. The remaining signature resembles COSMIC ID1, with some remaining convolution shared with the first signature. ID1 is ubiquitous across samples and cancer types and likely represents a background mutational process much like SBS1 or SBS5.

Structural variants are common in papillary thyroid carcinoma. Most tumors in our sample sets have at least one structural variant. The burden of structural variants per tumor is low, however, with a median of only one event per tumor. This event is often the driver gene fusion. Structural variants of all classes are represented, though large inversions and translocations, usually rare, are common in these data sets. Radiation dose is associated with the number of structural variants within a sample, as is age at surgery. The number of SSVs is sometimes associated with the number of structural variants, though this association does not hold across all ages. Certain types of structural variants, particularly inversions and interchromosomal translocations, were again associated with increasing radiation dose. These associations provide more evidence for the hypothesis that radiation is inducing double-strand breaks, some of which generate structural variants when erroneously repaired.

Chromoplexy occurs in roughly 6% of papillary thyroid carcinomas. This rate is consistent with that reported in the PCAWG study. Chromplexy often results in formation of a fusion gene. There was no difference in the proportion of samples with chromoplexy between exposed and unexposed sample sets. We conclude that chromoplexy is most likely not induced by radiation but is common in both sporadic and radiation-associated papillary thyroid carcinomas.

Chromosome 10 is commonly involved in structural variation in PTC. This is likely because of the presence of the *RET* proto-oncogene on Chromosome 10. *RET* is the most common gene fusion partner in our dataset and a primary driver of constitutive *MAPK* activation. Gene fusions were more common among radiation-exposed tumors, with 25% of unexposed tumors and more than 50% of radiation exposed tumors having at least one in-frame fusion. These fusions are formed primarily by balanced inversions

and translocations. Cytogenetic data indicates the rate of inversions and translocations increases with radiation dose [66]. A high proportion of gene fusion drivers (especially $RET - PTC$ fusions) was reported previously in Chernobyl-associated PTC cases [222], but the authors could not comment on whether the effect was dose-dependent. We find that the proportion of gene fusions generally increases with increasing exposure to radionuclides and that this association reaches statistical significance.

Together, these results support previous results regarding the role of ^{131}I in thyroid carcinogenesis. Iodine radionuclides sequestered in the thyroid generate double-strand breaks in the DNA in cells. These breaks can create driver mutations, especially gene fusions, by misrepair. Errors in repair also lead to an excess of small deletions. The ratio of deletions to substitutions provides a useful correlate to radiation exposure, one that may one day act as a useful biomarker for measuring radiation exposure in next-generation sequencing reads.

2.9.2 The REBC study as a possible model of multistage carcinogenesis and attributable risk

Certain aspects of the REBC study suggest it may be useful in deciphering the number of mutations required for thyroid carcinoma formation and the effects of external carcinogens, particularly ionizing radiation. The number of drivers required for carcinogenesis has been a matter of debate for decades, with Armitage and Doll describing a multi-stage model of carcinogenesis in 1954 following previous studies conducted some years earlier [231]. Armitage and Doll fit regression lines to log-transformed mortality rates on log-transformed age for a number of cancers obtained from registries in England and Wales. Based on work by Nordling [232], they hypothesized that the slope of these lines could be attributed to the approximate number of successive steps required for cancer development, with the additional complication of variable slope due to carcinogenic exposure. They concluded that a multi-stage process consisting of six or seven stages (i.e., mutations) could explain the patterns they observed, including the rapid increase with age, the irregularity of this increase in some cancer types, the latent period between carcinogenic exposure and tumor development, and that cancer incidence is often proportional to carcinogenic exposure. Their work also established a mathematical link between age, carcinogenic exposure, and mutation that has been continued by others [155, 233, 35].

Notably, Tomasetti *et al.* provided an extension of Armitage and Doll's work in the sequencing era [233]. Rather than death statistics, Tomasetti *et al.* used a mathematical

model based on incidence, the number of drivers, and the number of mutations per year to estimate the number of drivers required for tumor formation, leveraging the fact that subtypes of certain cancers have both different risks and different mutation rates. The use of the ratio of mutation rates means that this method is advantageously independent of the strong dependence of cancer incidence on age. Their model predicted the number of required driver mutations to be three for both lung adenocarcinoma and colorectal cancers. An approach such as this based on the total number of mutations (in this case in the exome) is only enabled by the ability of sequencing to capture the mutational burden of a tumor, which was not assayable in the time of Armitage and Doll.

A determination of the number of drivers (or tumors) attributable to radiation is beyond the scope of this thesis but certainly warrants investigation. The REBC dataset is comprised of whole-genome sequenced samples with a defined, acute carcinogenic exposure, including samples which received no radiation. There are notable genomic differences between exposed and unexposed samples, especially the proportion of gene fusion drivers. These characteristics suggest it may be possible with additional data to estimate the proportion of tumors attributable to radiation exposure. Many of the required variables for analysis are present in the REBC dataset. Given that the time of the nuclear release from Chernobyl is well-known, there are highly accurate estimates for age at exposure. Dosimetry data provides a continuous measure of carcinogenic exposure. IREP values give an integrated measure of probability of causation by radiation, accounting for age, sex, latency, and specific background risk. Age at surgery is also available, though this number is slightly confounded by intensive screening in the affected areas after the Chernobyl accident which likely brought many cases to the attention of medical professionals years before they would otherwise have been detected. From these two ages it is possible to derive the approximate latency between exposure and tumor detection for this dataset.

Certain challenges prevent simply applying the model of reference [233], however. The IREP values obtained in the study provide probability of causation but not estimates of relative risk between the exposed and unexposed subgroups. While estimated differences in incidence between sporadic and radiation-exposed pediatric thyroid cancer range from 2 to 7-fold at 1 Gray [167, 234, 235], the per-year SSV mutation rates are nearly identical in exposed and unexposed samples. As demonstrated in subsection 2.2.2, the number of SSVs per tumor is predominantly determined by age. This evidence suggests that SSV acquisition is not the rate limiting step in many of the REBC tumors. The structural variant or small deletion rate may provide a more useful measure, though many samples

have zero structural variants or indels. The small number of highly-exposed samples and the difficulty of determining whether a given tumor is attributable to radiation at low dose further complicates such analyses.

We leave the calculation of the driver number mutations required for papillary thyroid carcinoma formation to future studies as well, though note that it appears to be one or two based on the somatic changes we observe. Most tumors in the dataset are driven by activating SSVs or gene fusions in onocgenes. Integrated analysis of whole-genome data with RNA sequencing and methylation will further inform this observation.

Chapter 3

Structural variation in variation graphs

3.1 Introduction

In the previous chapter, I analyzed whole-genome data using algorithms which operate on linear reference genomes. Structural variants were common in this dataset and frequently generated gene fusion driver mutations.

In this chapter, I discuss how to perform comparable cancer-related analyses using graph-based references with a particular focus on variants longer than fifty basepairs. I first describe how structural variants are best represented in variation graphs and why some representations are more advantageous than others. This requires discussing strategies for selecting variants and constructing graphs. I then demonstrate how graphs can improve long variant genotyping for previously discovered variants. I briefly introduce methods for detecting structural variant breakpoints on variant graphs. Finally, I discuss applications of variant graphs in cancer-specific contexts.

3.1.1 Publication and collaboration notes

GFAKluge, which is used here for calculating graph statistics and which makes up a core part of the data interchange machinery of the variation graph toolkit, was published in [236]. The construction logic and variant forms described in section 3.2 are implemented in the variation graph toolkit, which was published in [99]. These have been extended in the svaha variant graph constructor and [3]. [99] and [3] involved significant collaboration with the authors of both papers.

3.2 Representing structural variants in graphs

While there is a single straightforward method for representing simple substitution variants in a graph (Figure 3.1), there are multiple ways to represent longer indels and structural variants. Choosing a variant representation can have significant impacts on the computational and analytical complexity of working with graphs.

Fig. 3.1 A variation graph containing five substitution variants with two alleles each. Each allele is represented by a single side of a bubble in the graph. Haplotypes would trace paths through these in a directed fashion. `vg` represents all possible haplotypes that could be generated from input variation. Graphs containing only substitutions are naturally acyclic.

Variation graphs may be cyclic or acyclic, though in practice it is common to convert cyclic structures to implicit directed acyclic ones to facilitate graph processing. In this work I will refer only to acyclic graphs.

A common operation when in linear reference space is to left-align and trim variants [92]. This produces a standardized representation of each variant and removes any flanking sequence that is already represented in the reference, facilitating comparisons between different variant callers. When operating on a variant graph, this process can be generalized by aligning the variant sequence against the local graph sequence and only incorporating bases which are not already represented in the graph. This minimizes the amount of duplicated material in the graph and ensures that complex variants are properly decomposed into their constituent variant bases. This representation in the graph is referred to as "aligned alternates" (Figure 3.2.)

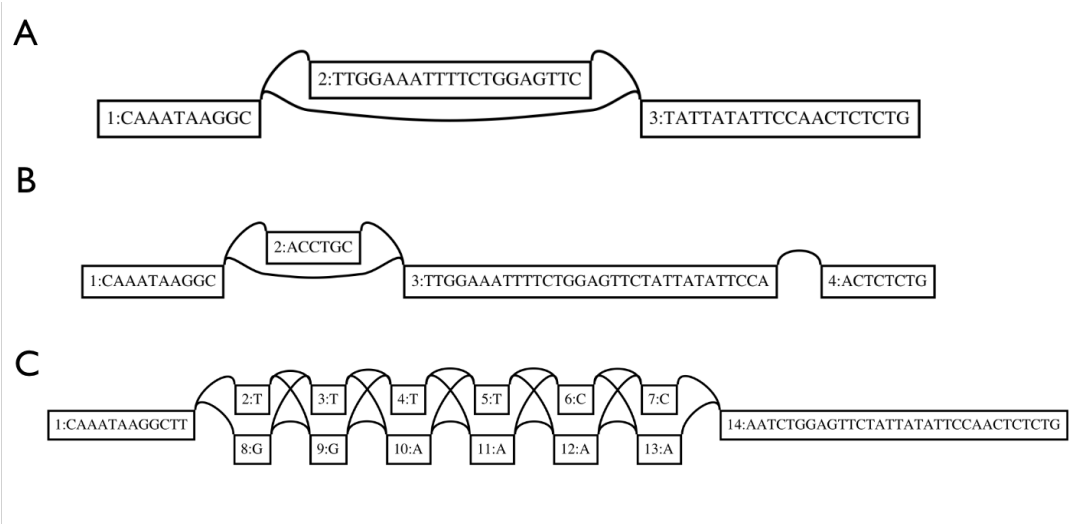


Fig. 3.2 Aligned alternate representation of (A) a small deletion (B) a small insertion and (C) a small inversion.

While the preferable representation for small variants, it presents major challenges when working with longer or more complicated variation. Aligning complex, lengthy alternative alleles is not always feasible or efficient. Significant increases in runtime are seen when performing Smith-Waterman alignment between the graph and the variant sequence. In practice, alignment of very large sequences requires significant amounts of memory and the current implementation of `vg` often crashes when aligning very large structural variant alleles.

Aligning alternate sequences produces a graph representation that does not always obviously resemble the variant in VCF form. This is especially noticable for inversions, which often decompose into complex intercalated runs of mismatches against the graph backbone (Figure 3.2C). An alternative representation is to put the literal variant sequence into the graph as-is (Figure 3.3). This representation, which I refer to as "flat alternates," has many advantages over aligned alternates.

Generating flat alternate graphs is efficient as no alignment is done. Variants also clearly reflect their representation in VCF space, although inversions resemble substitution variants and it requires further analysis to recognize the alternate sequence as the reverse complement of the reference allele (Figure 3.3C). In a flat alternate representation every allele is represented by at least one node, rather than some alleles being represented by nodes and others by solely edges as in the aligned alternates representation. This greatly simplifies working with the data structures of the graph for genotyping. It is possible to

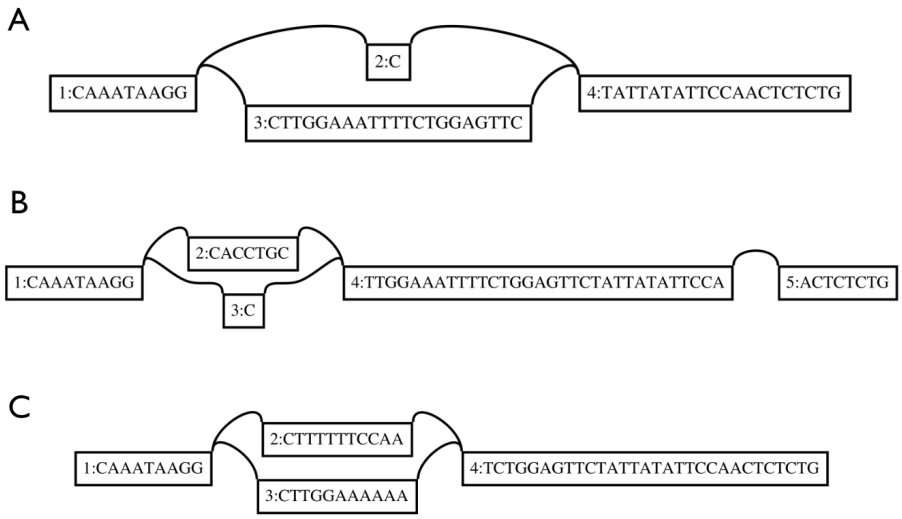


Fig. 3.3 Flat alternate representation of (A) a small deletion, (B) a small insertion, and (C) a small inversion. All three variants have all of the possible alleles represented by at least one node.

build flat alternate graphs with tens of thousands of structural variants on a personal laptop using available approaches.

Flat alternates still have major drawbacks. Flat alternates can introduce duplicate sequence in the graph. For deletions and insertions, the left flanking reference base is used as the alternate allele node and the deleted sequence (plus the flanking base) is used as the reference allele node. The single preceding base is therefore represented twice. While this should have only a minor impact on alignment, it significantly increases the number of nodes and edges in the graph without adding additional meaningful seqeunce.

For inversions and duplications, the entire variant sequence is represented at least twice in the graph. When mapping reads, this causes the mapping quality for reads aligned within the variant sequence to drop to zero, as duplicated sequence in the graph makes it impossible to map a read unambiguously. This greatly complicates downstream variant calling and genotyping. In addition, flat alternates often come directly from the VCF reference and alternate allele fields. Files containing full structural variant sequences can rapidly grow very large.

I use the term "graphical" to describe a more natural representation for structural variants in the graph (Figure 3.4). This form combines aspects of both the flat and

aligned alternates representations. Graphical variants are represented by at least one node and/or edge, but care is taken during the construction process to add edges rather than duplicating sequence in nodes. Graphical SVs only add sequence to the graph when that sequence was not present in the reference. They also provide a clear mapping between the VCF description and graph structures.

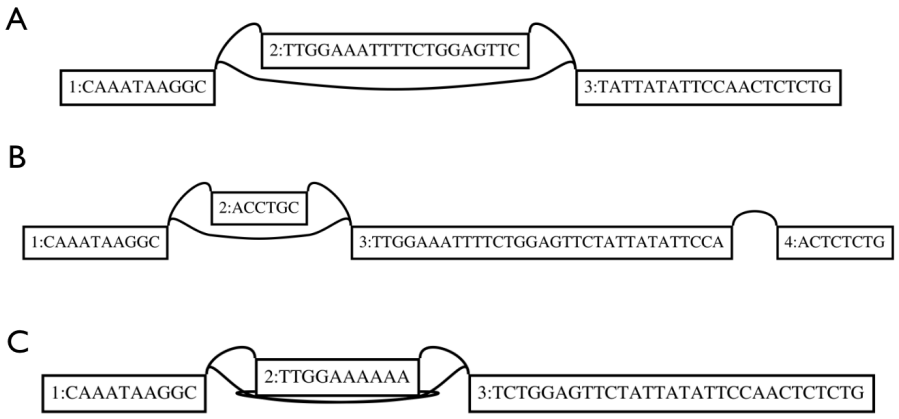


Fig. 3.4 Graphical alternate representation of (A) a small deletion, (B) a small insertion, and (C) a small inversion. Graphical alternates look similar to aligned alternates for deletions and insertions. Inversions are defined by edges rather than duplicate graph sequence.

There are still some tradeoffs that must be considered in implementations that create graphs using this representation. Some variants (e.g. deletions) are represented only by a single edge. This complicates the storage of paths in the graph, requiring them to be defined by both nodes and edges. By default, vg paths are node-based, and variant alleles that are defined only by an edge are not indexed as paths correctly. Properly indexing these variants requires redefining paths as a series of nodes and/or edges or defining the variants paths based on breakpoints and flanking reference nodes. Future designers of variation graph implementations should consider this carefully. In addition, constructing graphical alternates requires well-defined structural variant information tags (specifically the SVTYPE and END tags) to be present in the input VCF. The VCF format was designed primarily with small variants against the linear reference in mind, and these SV-specific tags are not always consistently implemented across callers or well-defined in the specification.

Building structural variant graphs in linear time

All forms of alternate allele representation can be built in linear time with regards to the number of breakpoints in the graph assuming that all incorporated variants are biallelic. In practice, however, the aligned alternate representation takes significantly longer than the other forms. I discuss why this form should be avoided for structural variation in section 3.2.1.

I implemented an algorithm for building graphs from FASTA and VCF files in low memory in the svaha package. This algorithm is able to construct graphs containing deletions, insertions, inversions, and translocations. It could easily be extended to included small variants. Graphs are output in GFA format and streamed to disk. A similar algorithm is used in [99] and [3], although these construction algorithms do not perform as well on large variants. An outline of the svaha graph construction procedure follows.

Construction starts with set of structural variants V and a FASTA reference R. Create an ordered set BP_c for each contig c in R, intitializing it to contain 0 (the start of the contig) and the last position on the contig. Initialize a map P which maps from genomic position to variant v and a map N which maps from genomic position to node.

For each variant $v \subseteq V$, calculate the breakpoints $[s, e]$ represented by the variant's position (v_s) and its END (v_e), CHR2, and SVTYPE info tags. Add s to BP_{chrom} and e to BP_{chr2}. Store v in P at e.

Many variation graph implementations require nodes to be shorter than a certain number of basepairs. When this is the case, perform the following. For each contig c, retrieve its list of breakpoints BP_c. Iterate from 0 to the length of the contig in increments of the desired maximum node length, adding each of these values to BP_c. BP_c now contains all of the positions on this contig where a node will start or end.

For each contig, perform the following. Create a pointer *prev* and initialize it to *null*. This will hold the last node seen on the reference sequence. Iterate over each break b in BP_c, starting from index 1. Emit a new node *node* with the substring $c[b, BP_c[index-1]]$. Store *node* in N at both b and $BP_c[index-1]$. If *prev* is not null, emit an edge between *prev* and *node*. This links up contiguous reference sequences. Point *prev* to *node*. When all breaks and all contigs have been processed the output file will contain a valid graph of the reference sequence.

To process variants, iterate over the map P of genomic position to variant v. If b is present in P, retrieve the variant v (which ends at b). Emit an edge from the node in N at v_s to *node*. If v is an inversion, this edge should point instead from *node* to $N[v_s]$.

Once all contigs have been processed the output file will contain a valid graph. Implementers must still handle interchromosomal variants so that both contigs of the variant are constructed before the relevant nodes are queried.

The current implementation of this algorithm can construct the COSMIC graph described in section 3.2.1 on a personal laptop with 8GB of RAM [1] in less than an hour. This is largely limited by the write performance of the system's disk. On a system with a high-performance NVMe SSD[2] this graph can be constructed in under ten minutes.

There is still room for improvement in this implementation. The two-pass algorithm for construction (once over all breaks, then over all variants) described above can be reduced to a single pass. A previous version of the program used a single-pass implementation. This was changed to facilitate prototyping as the single-pass version requires considering dependencies across contigs. It is also not necessary to store nodes which do not abut a variant site; removing them from the map N could reduce the RAM usage significantly.

Certain steps in the algorithm could be parallelized as well. Generating BP_c is embarassingly parallel across contigs. Graph construction could be parallelized but with significantly more difficulty. Node IDs in the graph are coordinated by a single counter to prevent having to properly namespace them later. However, a node ID space for each contig could be precomputed given that the number of nodes and edges is known after breakpoint sorting.

Using file-backed data structures could reduce the amount of RAM required to construct graphs with svaha. I implemented file-backed random indices for sequences in GFA files in the tinyGFA package but have not yet integrated these into svaha. Modern SSDs have excellent random read performance compared to hard drives and have become standard on even low-end personal computers. Integrating random-access indices reduce RAM usage with very little cost in terms of runtime on these systems. Paths in the graph are expensive to store in RAM and it is currently not possible to output them in a streaming fashion to GFA format. Erik Garrison designed and I have implemented a stream-compatible equivalent to GFA paths (the "W/Walk" line), but this is not a standard GFA tag. With these changes it would be feasible for users without access to high-performance or cloud computing systems to construct variation graphs in reasonable time.

[1] 13" Apple MacBook Pro Early 2015, Intel Core i5-5287U, 8GB RAM, 512GB SSD
[2] Lenovo T480s, Intel Core i5-8350U, 24GB RAM, 1024GB NVMe SSD

Dataset	Variant Representation	Time	Nodes	Edges	Total Sequence Length (bp)
GGSV Simulated Variation	reference-only	0.195 (s)	31250	31249	1,000,000
GGSV Simulated Variation	aligned alternates	70 (s)	200826	296857	1,227,043
GGSV Simulated Variation	flat alternates	0.32 (s)	43590	44588	1,337,866
GGSV Simulated Variation	graphical	0.35 (s)	42947	43946	1,336,868

Table 3.1 Construction times and node / edge counts for graphs from simulated data.

3.2.1 Constructing structural variant graphs

I constructed graphs from a number of data sources to demonstrate the effects of variant representation on construction time and graph complexity.

Simulated random reference sequence

I first simulated a reference genome and associated structural variation using the ggsv simulation package. Using truly random reference and insertion sequences ensure the Smith-Waterman alignment process used in the aligned alternate variant representation is able to align variants to the reference backbone and between alleles. I simulated a reference genome 1 Mbp in length, then simulated a random assortment of approximately 1000 variants. I restricted these simulated variants to biallelic variants that do not overlap and which are between 2 bp and 5 kbp in size. The final data set contained 327 deletions, 361 inversions, and 321 insertions of known size and sequence.

I then constructed graphs using each of the alternate allele representations for these variants as well as constructing a graph containing only the reference sequence. All graphs were limited to a maximum node size of 32. Because of this, even the reference graph contains many nodes and edges despite incorporating no variation. I used `vg` to perform graph construction then exported the graphs to GFA and calculated statistics using `GFAKluge`. The time to construct the reference-only, flat, aligned alternates, and graphical SV graphs as well as the number of nodes and edges in each is shown in Table 3.1.

Constructing the reference-only graph, which has roughly 31,000 nodes and edges, takes 0.195 seconds. The aligned alternates graph takes roughly 350 times longer to

construct (70 seconds), and contains 200,000 nodes and almost 200,000 edges. Including variation in flat alternate form imposes a much smaller penalty, taking only 0.32 seconds and producing a graph that is not much larger than the reference graph and the variation sequence. Constructing the graphical SV graph takes time comparable to the flat alternates graph and produces a graph with roughly 600 fewer nodes and 640 fewer edges.

Constructing SV graphs in aligned alternates form rapidly becomes infeasible, and it is hard to imagine a case where it would be the preferred representation for longer variants. Even in a dataset without overlapping variation and using alignable sequence, the penalty for aligning alleles against each other is apparent. In addition, the graph contains nearly five times as many nodes/edges as the flat and graphical representations. This has the effect of increasing graph complexity rather than reducing it for large variants.

The flat alternates representation contains roughly 100,000 more basepairs than the aligned alternates representation but five times fewer nodes and edges. This is an acceptable tradeoff as the penalty for increasing the number of basepairs in the graph is much less than increasing the number of nodes and edges. While the single-base reference nodes added for indel variants do not significantly impact mapping, the duplicate sequences of inversion and duplication varaints introduced in flat alternate representations would impact downstream mapping of reads. While it should be possible to adjust for this type of duplication in the mapping algorithm this introduces additional complexity.

The time required to construct the graphical SV representation is comparable to the time to construct the flat alternate graph, imposing only a small penalty. The graph produced is smaller in both sequence length and the number of nodes / edges. In addition, the graphical SV representation does not require a VCF which contains the variant sequences. The VCF used to construct the flat alternates graph was roughly four times larger than that used to construct the graphical SV representation. While the example used here is small and represents a best case scenario with low sequence divergence, there is clearly an advantage to using graphical or flat alternates over aligned representations.

The Catalog of Somatic Mutations in Cancer structural variant graph

I next performed similar construction experiments using a set of structural variants from the Catalog of Somatic Mutations in Cancer. I downloaded the full set of structural variants in the database (CosmicStructExport.tsv) for COSMIC v89 (hg38). I wrote

Variants	Allele Representation	Nodes	Edges
COSMIC	reference-only (hg38)	96508420	96508444
COSMIC	aligned alternates	Not Feasible	Not Feasible
COSMIC	flat alternates	Not Feasible	Not Feasible
COSMIC	graphical	96872672	96872477

Table 3.2 Build times and graph statistics for graphs built from structural variants in the COSMIC database.

a python script to transform the description field of the file into a VCF record with the proper SV tags; this is available in the COSMIC2VCF repository. The final VCF contained roughly 288,000 variants of the original 294,270 included in the database file, restricted to variants that are easily representable as a single event (e.g. excluding complex events like fold-back inversions). I excluded duplications as they are not handled well by current variation graph implementations and insertions because inserted sequence data was not available. This left roughly 230,000 structural variants which were used to build the graph.

Generating flat and aligned alternate representations of the COSMIC graph was not feasible. Many of the variants in the database are large; adding the full sequences of them to the VCF file created a file over 100 gigabytes in size. Reading this file and optionally aligning all of these variants would have been prohibitively difficult for most users. It is much easier to build the graphical SV representation. This graph takes roughly forty minutes to build with `vg` or svaha and has 96 million nodes and edges.

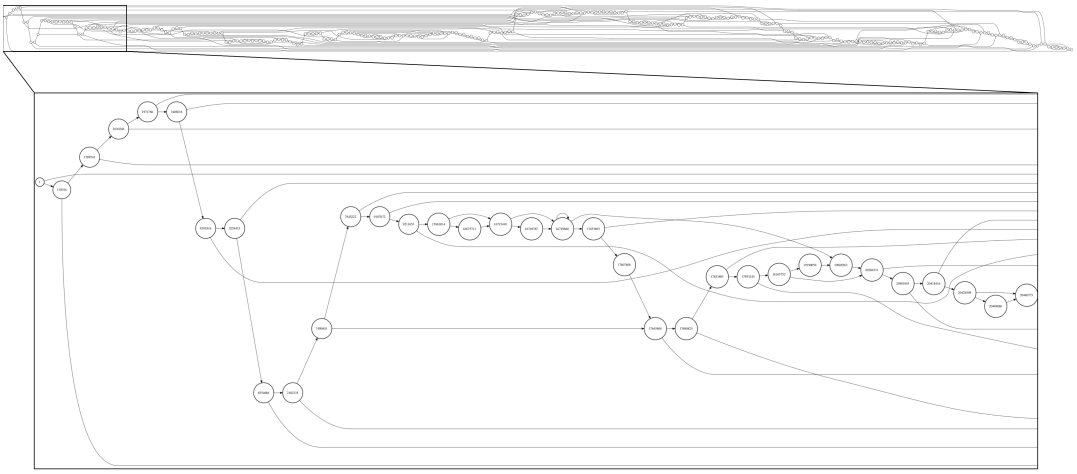


Fig. 3.5 A variation graph of well-formed COSMIC deletions on chromosome 22.

Variants	Nodes	Edges
hg19 (reference)	96739928	96739904
REBC (SV hypermutator)	96740200	96740343
REBC (all)	96872672	96872477

Table 3.3 Graph statistics for graphs built from structural variants in the REBC sample set.

A structural variant graph from Chernobyl samples

To test the utility of variation graphs on real data, I constructed a graph of all structural variant calls from our Chernobyl cohort. The calls used for construction included those produced by any caller. In total, this set contained 86550 variants, significantly more than our filtered set described in section 2.6. I also constructed a graph containing variants from one of our samples with apparent chromoplexy to show that complex events can be represented in variation graphs as long as they can be decomposed into a series of simpler events.

I also generated a graph of the variants in the structural variant hypermutator (Figure 3.6). This sample had multiple interchromosomal structural variants in a pattern that resembled chromoplexy (Figure 2.35).

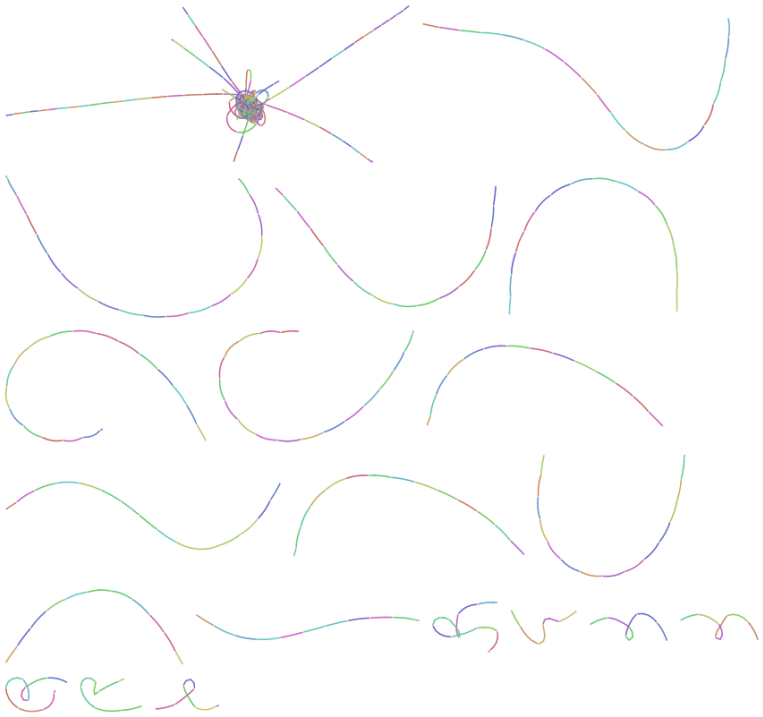


Fig. 3.6 The whole genome variation graph of the SV hypermutator sample. Most chromosomes are linear because they harbor no SVs. The chromosomes involved in chromoplexy are condensed into a single subgraph.

Constructing the graph displayed in Figure 3.6 takes less than twenty seconds to build on a 2015 MacBook Pro. A similar graph with nodes limited to 32 basepairs in length takes roughly 40 minutes, roughly the same as the reference genome graph and the pan-REBC graph with the same restriction.

Structural variant graphs

Using `vg` or svaha it is possible to construct structural variant graphs for simulated data, personal genomes, and mutational catalogs. Graphs containing hundreds of thousands of variants can be constructed on a standard personal laptop in minutes. In the following sections, I will discuss how such graphs can be used and potential areas of improvement in graph-based analysis.

3.3 Typing structural variants in a variation graph

A primary benefit of graph-based representations of variation is the ability to map reads and directly genotype variants without first going through the process of rediscovering variants. I term this process *variant typing* to distinguish it from *variant calling*, in which variants are discovered without *a priori* knowledge of their existence. While the process cannot discover variants *de novo* it is significantly faster at recovering variants which are recurrent across samples, as is common in cancer.

3.3.1 Existing approaches

Variant typing has been performed previously for both small and large variants. FreeBayes allows force-calling of variants [210]. Many tools exist for genotyping structural variant calls from a VCF and a matched set of mapped reads [88, 237, 86, 85]. These genotypers often require that SV calls come from specific *de novo* callers, limiting the variants which can be genotyped, and often perform poorly on insertions. Paragraph [238], a newer tool, uses a graph-based genotyper that relies on the same graph Smith-Waterman algorithm as [99] and performs much better on insertions.

3.3.2 Typing structural variants with vg recall

I implemented a graph-based structural variant typer, accessible via the `recall` subcommand of vg. In vg, this process relies on storing the VCF variant calls as paths in the graph. These calls may come from any variant caller which produces the standard SVTYPE, SVLEN, and END tags. Insertion calling requires the SEQ tag or an external FASTA file of insertion sequences.

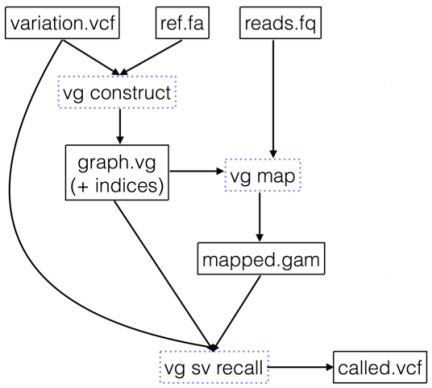


Fig. 3.7 An overview of the `vg` recall pipeline.

`vg` recall takes as input a VCF file of variants to genotype and a GAM alignment file from the `vg map` commands. Variants must be in the graph and stored as paths to be typed. In addition, the current implementation relies on the flat alternate representation of SVs, though this could be extended to graphs utilizing graphical alternates if `vg` supported paths that contain only edges.

The `vg` recall process works as follows. Iterate over the alignments in a GAM file. For each alignment, check whether the alignment contains any nodes which correspond to a variant in the graph path index and the VCF file. For any alignments that match a known variant, check whether the alignment supports the reference or alternate allele path. Each variant in the VCF is then annotated with the alignment counts for the ref/alt alleles. The genotype likelihood formula from [239] is then used to calculate an estimated genotype for the annotated variant alleles.

I demonstrated the performance of `vg` recall on a set of simulated variants from ggsv (section 3.2.1). I compared these results to the calling and genotyping pipelines implemented in delly [85] and lumpy [86], which uses the svtyper genotyper [239]. The simulated variant set included deletions, insertions, and inversions Figure 3.8. `vg` recall had the highest sensitivity of the three callers, correctly calling the genotypes of all simulated homozygous alt sites.

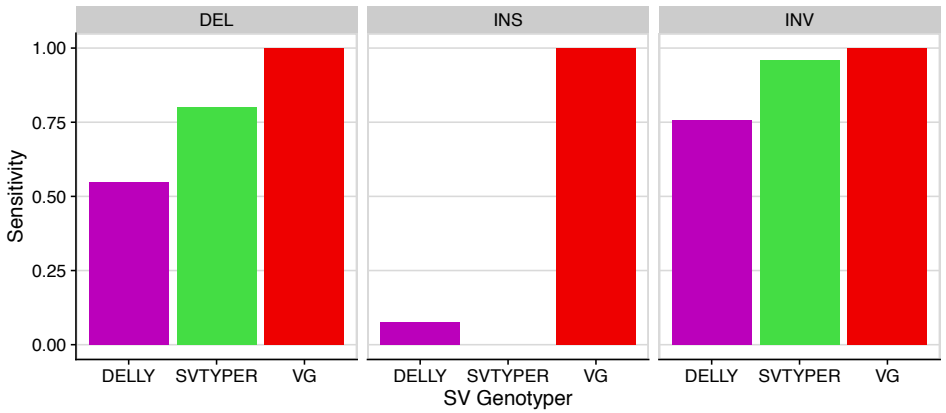


Fig. 3.8 Comparison of correctly-called structural variants from each caller across deletions, insertions, and inversions. `vg` performs slightly better for deletions and inversions but significantly outperforms both callers on insertions when the inserted sequence is known. These results are for a VCF with no error in the position of variants.

In addition, the typing pipeline was much faster than the time required to call and genotype variants using the traditional linear reference-based callers (Figure 3.9). The mapping time however was much greater. Aligning reads to a graph introduces additional complexity to the mapping process, and this has resulted in runtimes up to ten times greater than for linear mappers [107]. As graph aligners become more common, improvements in implementation should decrease this significantly.

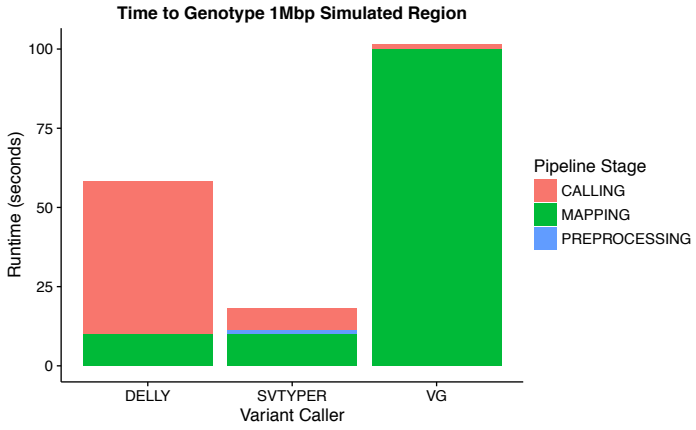


Fig. 3.9 Runtimes of the three calling pipelines. While `vg` is much slower at mapping reads, it is much fast than either `DELLY` or `SVTYPER` at genotyping variants.

In [3], we used an extended version of this approach to type variants from simulated data as well as from The Human Genome Structural Variation Consortium (HGSVC) [240]. This method allows variants to be typed when present in any form that can be

represented by the snarl decomposition described in [241] (Figure 3.10A). We compared the performance of `vg` to three existing SV genotypers (BayesTyper [242], DELLY [85], and SVTYPER [239]) across variant classes (Figure 3.10B). BayesTyper is a graph-based approach that uses exact matching of kmers across SV junctions to genotype reads while DELLY and SVTYPER rely on alignments to a linear reference. Both graph-based methods consistently outperformed the other tools when errors are not present; `vg` is more robust to errors in variant position. At a depth of twenty reads, both graph-based methods approach an F1 score of 1.

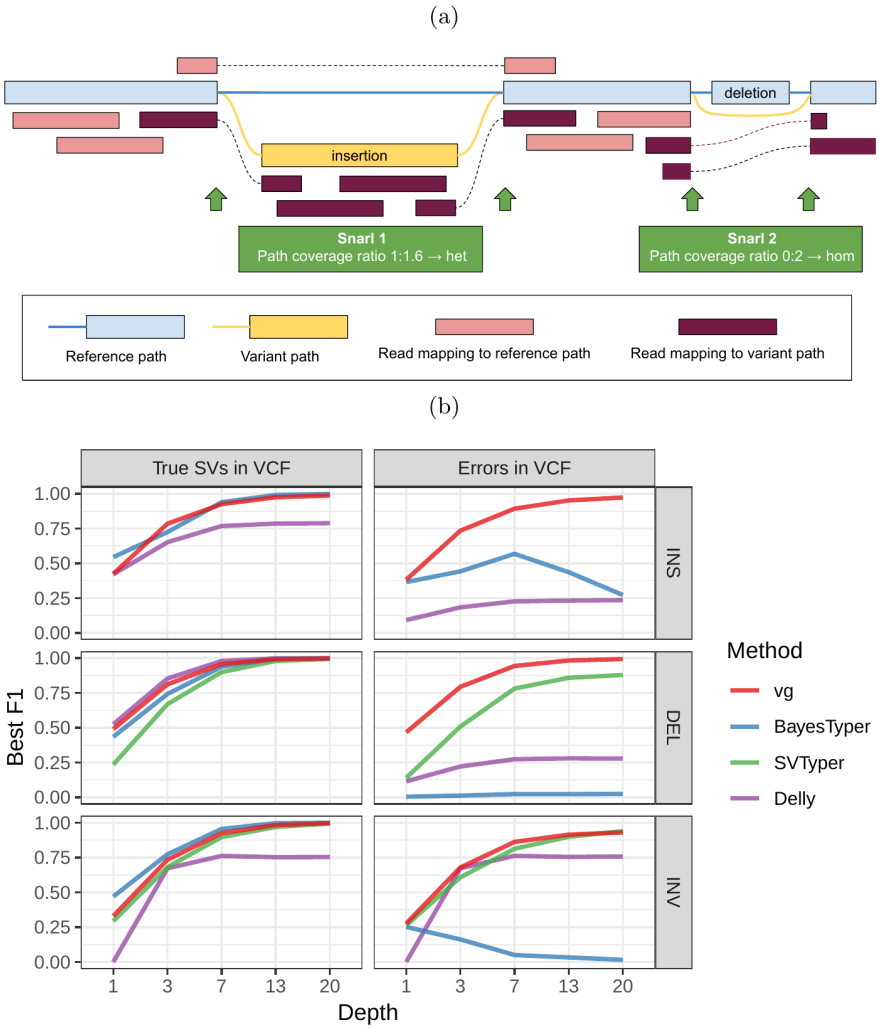


Fig. 3.10 Figure one reproduced from [3]. (a) vg uses read coverage information to calculate support for an allele across a snarl in the graph. The algorithm for doing so is described in the Methods section of the paper. (b) Performance (F1 score) at various depths of variant genotypers for true VCF variants (left panel) and a VCF file with up to ten bases of error in position (right panel). vg performs similarly to the best-performing algorithm across all depths when no errors are present. When errors are present, vg outperforms existing methods.

The HGSVC and Genome in a Bottle data used in [3] included both long reads from short and long read platforms. Structural variants were called using ensemble methods described in [91] and [243], respectively. The two graph-based methods again outperformed DELLY and SVTYPER (Figure 3.11). BayesTyper had the best overall F1 score, although `vg` had comparable performance in non-repetitive regions. The performance

advantages of using a graph were consistent across all variant sizes (Figure 3.11B). The distribution of variant sizes in the sample set is shown in Figure 3.11C.

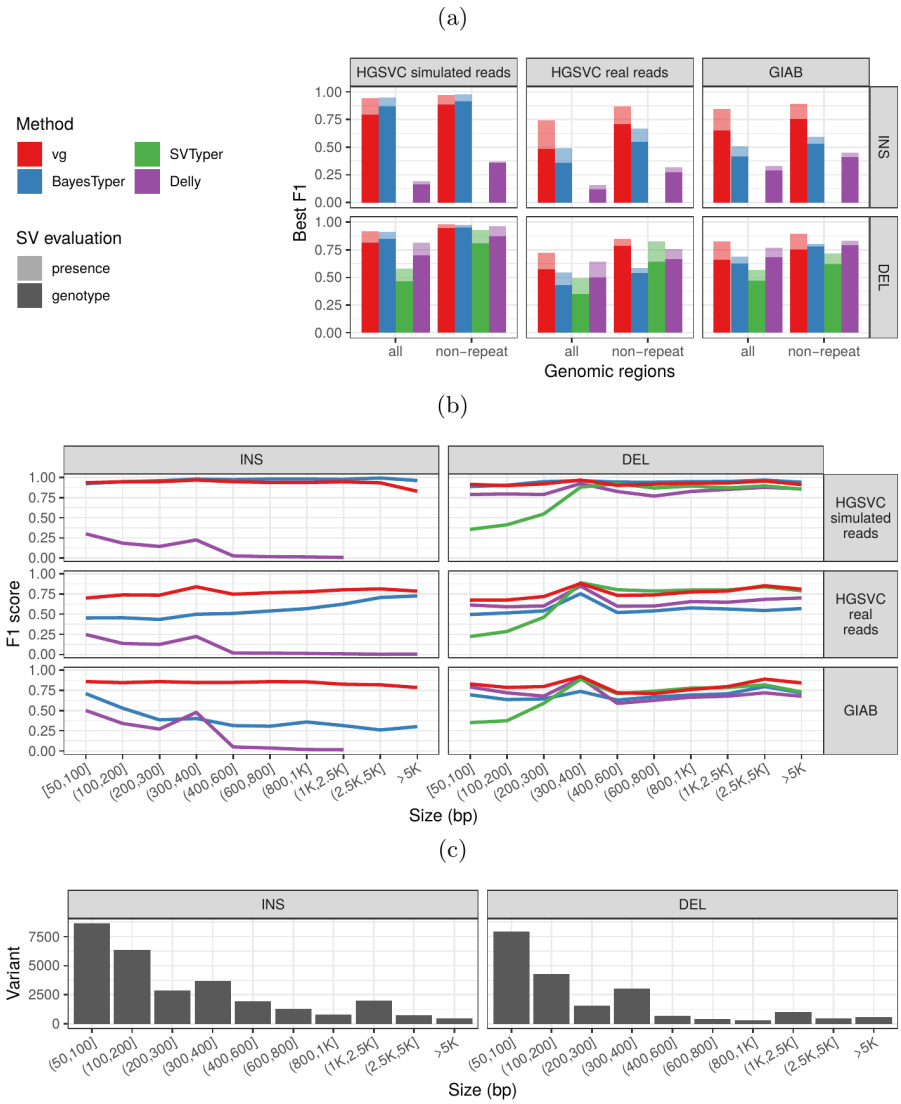


Fig. 3.11 Figure two reproduced from [3]. We assessed the ability of each algorithm to determine the presence of a variant (low alpha) and its genotype (solid) in both repetitive and non-repetitive regions across simulated reads from HGSVC, real HGSVC Illumina reads, and real Illumina reads from the Genome in a Bottle sample (a). (b) shows the maximum F1 score for each variant class and read set across variant sizes in the genome. (c) shows the variant size distributions in the HGSVC and Genome in a Bottle variant calls.

Improvements are still needed to cover the full spectrum of variation. In [3] we only extend our interrogation to variants that vary by up to ten basepairs in their position.

While many breakpoints are recurrent in cancer samples (as shown in section 2.6, even the most recurrent breakpoints can vary by hundreds of basepairs (Figure 2.44). Future work should focus on being able to accurately type variants even when there is error in the called breakpoints. In addition, the implementation in `vg` requires significant amounts of RAM and requires splitting the genome into 2.5 Mbp bins. This will exclude some very large but potentially relevant variants from analysis; it also means that interchromosomal variants cannot yet be genotyped. Nonetheless, these results highlight the promise of graph-based approaches in quickly and accurately genotyping recurrent variants.

3.4 Exploring detection of novel variants with graphs

In section 3.3 I demonstrate the utility of variation graphs in genotyping known structural variants. However, this approach is restricted to variation that has been previously discovered.

I explored several ways of calling structural variants *de novo* from graphs and graph-relative alignments. Linear reference callers most often produce a VCF file which describes variation relative to a reference genome. Section 3.4.1 describes a method for extracting variant calls in VCF from bubbles in the graph relative to an embedded reference path. Section 3.4.2 describes how structural variant mismapping signatures in graph alignments could be used for structural variant calling.

3.4.1 Graph to VCF conversion

Graphs produced by alignment of long high-fidelity sequences contain bubbles that are due mostly to variation rather than error. It is possible to call variants directly from these without relying on aberrant read mapping signatures. Long reads also do not produce the same signatures as short reads, as they are much more likely to span breakpoints of even moderately-sized structural variants.

In the case that a graph contains a reference genome, or another spanning path, a reference-relative VCF can be produced based on the bubbles within it. I implemented this method, called graph deconstruction, in `vg`.

The procedure for deconstruction is as follows. Given a graph g and a reference path P, compute a bubble decomposition of g. For each bubble, compute its bubble sides by performing a depth-first search from the bubble's entry vertex to its exit vertex. Test whether any of these bubble sides are spanned by P. If so, the path formed by this

bubble side is the reference allele. Compute the left-most coordinate in P touched by any bubble side. This is the genomic coordinate of the VCF variant. Each alternate path from the left-most coordinate to the next coordinate on the reference path constitutes one allele of the variant.

Initially, deconstruction relied on a superbubbles decomposition according to [244]. A superbubble is a directed acyclic subgraph which has one entry and one exit vertex with no edges between nodes inside the subgraph and those outside (except for the entry and exit nodes). Deconstruct was significantly extended by Dr. Glenn Hickey, Dr. Adam Novak, Dr. Benedict Paten, and Dr. Jordan Eizinga to utilize the snarl decomposition described in [241]. Dr. Hickey also solved many of the scaling issues and enabled running on whole genomes.

Deconstruct is designed for long, high-fidelity sequences such as PacBio CCS reads or finished assemblies. Short read alignments will rarely span structural variant breakpoints to produce a well-formed bubble. Because the total number of reads is much larger for short read experiments than long read ones, many more paths are stored in the graph's path store at the same depth of sequencing. These factors make deconstruction intractable on graphs of short reads data. Short read experiments are still common due to their low cost and high per base accuracy, however, and a method for calling structural variation on graphs from short reads would be an important addition to the `vg` ecosystem.

3.4.2 Detection of structural variants from discordant read signals

Known structural variants that are already present in the graph can easily be typed in short or long reads using the methods described in section 3.3. Novel variation, however, cannot be typed and must be called *de novo* from the readset. While deconstruction can call new variation, it is not practical or effective with short read sets.

Modern structural variant callers that operate on the linear reference primarily detect the presence of structural variation by their signatures in short read mappings.

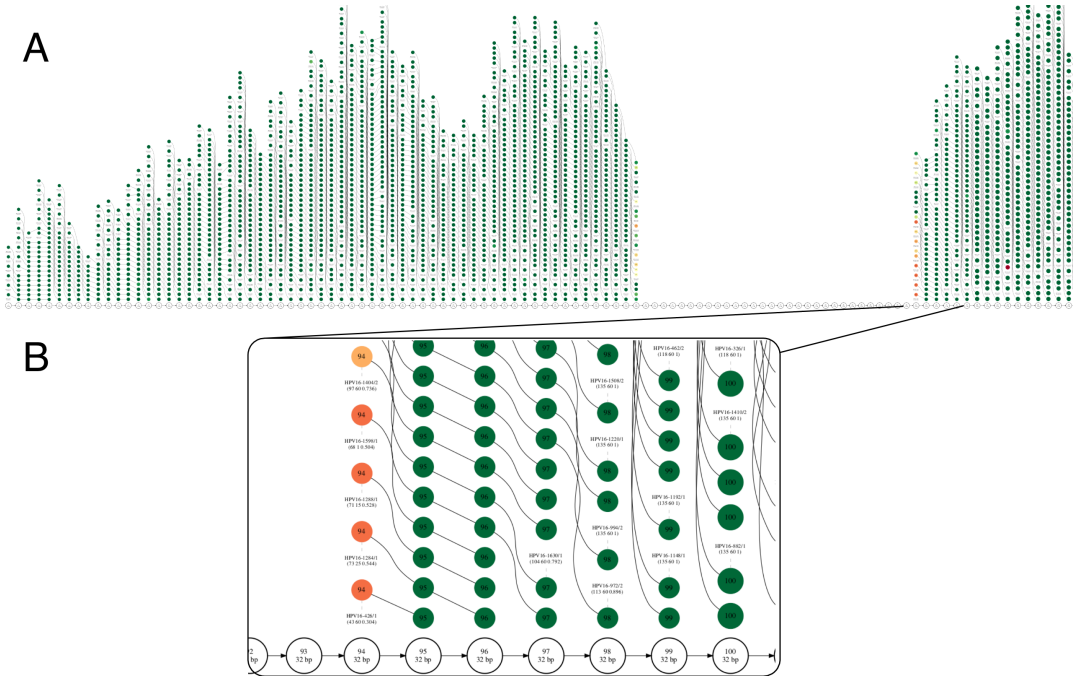


Fig. 3.12 (A) Simulated reads from a genome with a 1kbp deletion aligned to the HPV16 reference genome. Reads are colored by their mapping quality. Reads from the simulated genome do not cover the deleted portion and low mapping quality reads (many of which are soft clipped) appear at the breakpoint (B).

I implemented an analogous procedure on the graph that can extract common structural variant signatures from graph-read alignments. This functionality is available in the `vg sift` subcommand. This subcommand filters read alignments for specific hallmarks of reads that span structural variant breakpoints. This information could be used to produce structural putative variant calls for short read data aligned to the graph.

For each read pair in an alignment file, the pair is tested for SV signatures. This includes whether the mates are mapped too close or far apart, whether one or both reads are unmapped, whether either mate is soft-clipped, and whether mates are mapped in the proper orientation. Reads are output into a separate file for each signature. This essentially constitutes a multimap of SV signature to read pairs supporting that signature.

Soft-clipped reads can indicate that a read spans an SV breakpoint, especially if the soft-clipped portion aligns to a distant portion of the genome. An example of this is shown in Figure 3.12.

The soft-clipped reads extracted above could be used as the basis for *de novo* calling of putative structural variants. Soft clipped reads which map near each other in the graph can be used to stablish a putative breakpoint location. Reads that map near this position

and which were extracted as representing possible SV signatures could then be used to corroborate the breakpoint and establish a variant type. Lastly, a reference-relative position would need to be generated from the putative breakpoint. The functionality for this exists in `vg` but is not yet fully coordinated as an accessible framework.

3.5 Improving structural variant representations with graphs

In the following sections, I discuss the tradeoffs involved when selecting variants for inclusion in a graph. While incorporating variants initially increases the possible sensitivity of alignment, adding more variation can eventually lead to decreasing alignment performance and increasing complexity. I then discuss a possible measure of read mapping performance that is robust to variants that are not well-resolved in the graph. Finally, I describe a procedure for refining breakpoints that demonstrates how the mutability of variation graphs can be used to breakpoint-resolve structural variants.

3.5.1 Augmenting variation graphs

In `vg`, we rely on a process called graph augmentation to call structural variants in the graph *de novo*. In graph augmentation, read alignments are compared to the graph and any new sequence or path in the alignment is incorporated as a node or edge in the graph. The read name and its alignment is also stored as a path in the graph's path store.

Graph augmentation greatly increases the graph complexity, as variation and sequencing errors are incorporated as new graph sequence and a path is stored for each read. Paths are expensive to store for variation graphs and this process rapidly generates graphs that are hard to store and query. We removed the requirement that a graph be augmented with read alignments in [3] to make typing analysess more tractable at the cost of losing the ability to call such variants *de novo*.

A preferable approach would be to implement a method that maintains reads separate from the graph. This prevents the computational complexity of operating on a graph that contains significant amounts of incorporated read sequence. Operating out-of-graph is analogous to how structural variants have previously been called on linear reference genomes, where discordant signatures in read mappings are used to identify the presence and type of structural variant.

The error rate and total number of reads cause the computational blowup that impairs graph augmentation. If the number of sequences to be incorporated were relatively few and of high quality it would be feasible to operate on a graph composed of them. This is essentially the case for graphs generated by aligning genome assemblies. Whole-genome assembly graphs can be produced by `vg`, HAL [245], CACTUS [119], and seqwish. While graph augmentation should be avoided in the case of noisy, high-error rate reads and when many reads are present, it can be useful in cases where sequences come from polished assemblies or high-fidelity reads.

3.5.2 Selecting variants to include in the graph

When implementing graph construction algorithms for structural variants it is important to consider not only the choice of representation but also how specific variants impact the graph topology. While naively one might assume that adding all possible variation to a graph would yield the best alignment performance, this is not always the case. The computational complexity of operating on the graph rapidly increases as more variation is included. Runs of incorporated variants can also generate sequences that spuriously match other regions of the graph, which can lead to decreased alignment performance.

How variants are selected for inclusion into variation graphs has been discussed previously. We used an allele-frequency filter of 1% prevalence in the 1000 Genomes VCF to balance graph complexity and alignment performance in [99]. As more variant sites were added, read mappings improved at those sites when compared to the linear reference. However, containing reference alleles sometimes received worse mappings, and this effect increase as more variation was added to the graph.

Algorithms operating on the graph tend to have computational complexity on the order of nodes, edges, or paths in the graph. As more variation is included in a graph, the number of nodes, edges, and paths can increase significantly.

Reference [102] thoroughly examined various strategies for balancing the tradeoff between improved alignment sensitivity and increasing computational complexity when incorporating more small variants in the graph. These models incorporated both population frequency information and a measure of computational complexity. Alignment was performed using HiSat2 [104] (a graph-based variation-aware aligner) or the Enhanced Reference Genome [101], which adds alternate allele sequence to a known reference and aligns reads to the modified linear sequence. Previous studies had relied on filtering strategies based on allele frequency, database inclusion, or ethnicity. For most models the best performance occurred when 8-15% of the total variants were included, though

some models performed best when including up to 30% of variants. Significant decreases in the percentage of reads aligned (both correctly and incorrectly) and increases in computational complexity were seen above these levels. Only the simplest models were able to scale to the entire human genome.

In cancer, deciding whether a variant is included in the graph requires considering more than its allele frequency and whether it increases graph complexity. Factors such as a variant's penetrance or effect on treatment can outweigh the cost of adding it to the graph. Missing variants that significantly increase lifetime cancer risk or which can inform personalized treatment can impose a major cost on both the patient and the healthcare system.

Fig. 3.13 A Bandage plot of the a graph constructed from all simple COSMIC intrachromosomal structural variants.

Structural variants present additional challenges when constructing graphs. Interchromosomal variants are common in disease. The best practices for genome graphs encourage parallelizing construction and indexing by disjoint subgraph, often by contig or chromosome. Adding interchromosomal variants to the graph entangles these subgraphs (Figure 3.14), creating a graph that is much more complicated than one containing only intrachromosomal variants (Figure 3.13). If small variants are also included in the

subgraphs, or if many structural variants are present, such graphs can quickly become intractable.

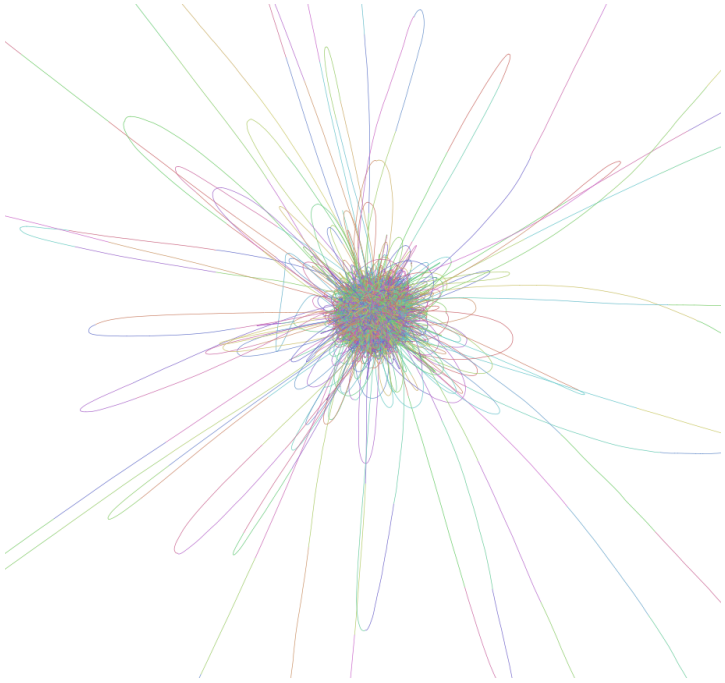


Fig. 3.14 Bandage visualization of COSMIC structural variants, including interchromosomal variants. Variants that have breakpoints on multiple chromosomes entangle the subgraphs containing these chromosomes. For COSMIC structural variants, all chromosomes become entangled into a single subgraph.

Because interchromosomal variants entangle subgraphs, I recommend not adding them to graphs already containing significant amounts of variation. Instead, analysis should be split across multiple graphs and a common, stable reference system (such as the reference genome) can be used to map all variation into a common coordinate system. As graph coordinate systems become more developed this process should become practical [246].

3.5.3 Common coordinate systems will facilitate alignment to multiple graphs

The naive approach to assaying the full spectrum of variants is to align reads once to a single graph containing all variants of interest. For the reasons described in subsection 3.5.2 this approach is suboptimal when the graph contains many complex or adjacent variants. Common graph coordinate systems will facilitate the joint analysis

of reads mapped to different graphs. This means that reads from the same experiment could be mapped to multiple graphs containing different variant sets and then collated for downstream analysis.

There are notable advantages to such an approach. The majority of reads in human resequencing experiments will exactly match the reference genome or points of common variation. These reads do not need to be realigned further as their alignments cannot be improved. Common variants can then be quickly typed from these alignments. This would save significant amounts of work compared to performing alignment on a complex graph and then variant calling the entire readset.

Reads which do not exactly match a graph of common variants may contain rare variation, structural variant signatures, or sequencing errors. This subset of reads should be a significantly smaller percentage of total reads than those that map perfectly. The reason a read or pair did not exactly map can sometimes be determined by its alignment. For example, reads which contain soft clips may fall at structural variant breakpoints and reads with mismatches at their tail end may contain sequencing errors rather than true polymorphisms.

This subset of reads can be aligned to more complex graphs containing variants of interest such as one composed of known disease predisposition variants. In the case of cancer, this may be known pathogenic germline variants (e.g. in the $BRCA1/BRCA2$ genes) or somatic variation from COSMIC. Reads which map to the reference allele at these positions can be pulled from the initial alignments to support accurate genotyping. Improved alignments combined with low-pass whole genome sequencing could one day replace custom gene panels and microarrays in genetic testing to provide a more complete genomic picture.

Finally, reads containing structural variant signatures can be aligned to graphs containing structural variants of interest. While the tools for construction of graphs and read alignment already exist, it will take the development of a stable, common graph coordinate system and translation machinery to make the analysis described above practical.

3.5.4 Measuring concordance between imperfect alignments and the graph

Mapping quality and percent identity are not well-suited for assessing the quality of alignments to graphs containing structural variants. Modern short-read sequencers

produce reads that are 100-150bp in length, though reads from specific contexts such as ancient samples may be as short as 30 bp. Many aligners do not produce full or proper read alignments for short reads which span structural variant breakpoints.

One measure that I started to explore was an alignment concordance score calculated for a pair of reads as follows:

$$C_a \propto -(S + 2^P) + E + C$$

where S is the total number soft-clipped bases, P is the z-normalized insert size of the pair, E is a fixed penalty for one-end-anchored read pairs, and C is a fixed reward for reads being on a contiguous sequence. This score is proportional to these values and not exactly defined because it may be useful to scale it to a specific range such as PHRED or the interval $[0, 1]$. Incorporating these values into a single metric allows one to more accurately score alignments which may be perfect at the per-base level but which map to repetitive regions of the genome as well as those that support a variant that is not well-represented in the graph. This measure may can be combined with percent identity / mapping quality to provide further information about how well a read or pair is mapped. Because no graph-specific information is contained in this score it is useful as a means of comparison to linear mappings.

For single-end or long reads, there is no information about insert size or one-end-anchoredness. In this case information about the mapping position of the soft-clipped portion can be used to generate a similar measure. Having a measure of graph-read concordance that does not rely on alignment identity allows one to compare read mappings even when reads map inexactly or variation is not represented perfectly in the graph.

3.5.5 Homogenizing breakpoints in an SV graph

As previously discussed, it is rare that structural variants reoccur at exactly the same breakpoints. There is significant interest in determining the breakpoint-exact location of structural variants. Regions of recurrent structural variants may constitute fragile sites in the genome, which are susceptible to illegitimate recombination or sensitive to break-inducing mutagens such as ionizing radiation.

The process of determining exact breakpoints from inexact locations is termed "breakpoint refinement." Traditionally, this has been performed by local assembly, though this approach is computationally expensive and can be confounded by microhomology, which is common at SV breakpoints. Variation graphs can be modified to include more

variation without losing information about the core reference. This property suggests that a process for breakpoint refinement could easily be defined on mutable variation graphs.

I implemented a procedure for refining breakpoints on variation graphs based on augmenting the graph with read mappings and generating candidate alleles by realigning graph tips, which represent possible structural variant breakpoints. I call this procedure *homogenization*, as the many alleles introduced by graph augmentation are homogenized to one or few best representations. While graph augmentation is expensive and in general should be avoided it is particularly useful in breakpoint refinement. This process could be improved by iteratively realigning tips, scoring alignments, and regenerating the candidate allele set. It is possible to avoid augmentation when performing breakpoint refinement; however, I have not implemented such an out-of-graph approach.

Breakpoint homogenization of a given variant proceeds as follows. The graph is first augmented with a set of alignments. Alignments which are softclipped are collected. The softclipped portions are remapped to the graph, generating a set of edges between the last anchored portion of the read and the softclipped portion's new alignment. These edges are the candidate alleles of the structural variant. It is important that the soft-clipped portion be of sufficient length and uniqueness within the genome to map accurately. Reads can then be realigned to the graph containing the newly-generated candidate alleles.

Breakpoint homogenization suffers from many of the same challenges as assembly-based refinement. Repetitive regions or microhomology near the breakpoint can impair local alignment of the clipped sequence. Alignment of longer clipped portions should be slighty more robust than deBruijn graph based assembly methods, which are highly-sensitive to repetitive sequence, but less robust than methods relying on string or overlap graphs. Homogenization is also computationally expensive, requiring mutliple rounds of read alignment and graph indexing. The procedure could likely be parallelized across allele candidates though no such implementation yet exists. Adding rounds of approximately scoring the candidate alleles and removing low scoring ones could make the process both more accurate and more efficient.

3.6 Potential applications for variation graphs in cancer genomics

The main purpose of variation graphs is to improve sequence alignment by reducing reference bias. Better alignments have been shown to improve downstream variant calling and other processes that rely on accurate sequence-sequence alignment [247, 109]. While previous work has established these effects for small variants the same had not yet been done for structural variants.

In this chapter I have discussed how to construct variation graphs containing structural variation and use them to genotype variants. I also discussed potential methods for calling novel variation and improving graph-based structural variant analysis.

There remain many areas in cancer genomics that could benefit from graph-based analyses. Graphs composed of germline variation could improve the sensititivity of detecting cancer predisposition variants, especially indels. Early and accurate detection of such variants is a major goal of consumer genetic testing. While I briefly discuss a potential method for doing so, to date there is little work in this area. Graphs may one day function as digital gene panels, providing increased sensitivity to detect consequential variation by algorithmic improvement rather than laboratory technique.

Somatic analysis may also benefit from graph-based approaches. `vg` includes many haplotype-aware methods for alignment, variant calling and phasing. These could be extended to create tumor-normal graphs for individuals or cohorts. Reads could then be mapped to germline and somatic variation simultaneously. The integration of somatic and germline variation calling into a unified framework mirrors the biological phenomena proposed in Knudson's two-hit hypothesis. Such joint mappings could also improve our ability to assess tumor purity and subclonality, both of which rely on accurate allele frequency estimates of somatic variants and the germline background.

In the special case where tumor reads from a single individual are mapped to a graph containing private variation from that patient's normal sample the process of mapping reads and calling variants is directly analogous to contemporary single-sample somatic variant calling methods. A graph coloring approach (like those used in [248] and [118]) could be applied to indicate which nodes were incorporated from tumor or normal variation. The addition of normal variation from other samples to the graph is analogous to comparison of the tumor sample against a panel of normals. Graphs generated in this manner contain both germline and somatic variation and rely on annotations to distinguish the origin of a variant. Graph-based somatic filtering can

then be performed based on the graph annotations and measures of read depth in the tumor and normal. An alternative approach is to map reads from a normal sample to a graph of patient-specific tumor variation, then perform a bubble-popping procedure to remove any nodes supported by the normal genome but not the reference (essentially removing the germline background from the graph). The remaining graph structures represent a patient-specific tumor genome graph. Such graphs present the most exciting opportunities for exploring new approaches for analyzing tumor subclonality, but creating such structures will require new algorithms for performing the bubble-popping step and careful tuning of variation discovery pipelines to improve sensitivity to tumor variation while minimizing the incorporation of errors.

Lastly, graphs can play a role in better assessing repetitive elements in the genome. Short tandem repeats (STRs) are frequently mutated in cancer, can be hallmarks of mutations in specific DNA repair genes, and may function as driver events [249]. They have also been implicated as possible markers of radiation exposure in both germline and somatic cells [250, 251]. Reference [249] used global alignment to assess STR variation in tumor genomes. A variation graph could be used to represent the reference allele, alternate alleles, as well as established variation in these regions. Reads could then be aligned exactly to these sequences.

Variation graphs provide demonstrated benefits for working with small variants and structural variation. As our catalog of known variation expands there are ample opportunities to improve upon existing analytical pipelines using variation-aware algorithms such as variation graphs.

Chapter 4

Further algorithms for examining genetic heterogeneity

4.1 Introduction

In Chapter 3 I explored how variation graphs can improve genomic inference. Graphs are variation-aware because they incorporate previously-observed or putative variation. The advantages of variation graphs come from their abilities to improve read alignment where the genome differs from the reference.

However, graphs are not the only variation-aware data structures or algorithms. In this chapter I explore alignment-free variation-aware algorithms. Alignment-free algorithms perform genome inference without performing alignment, relying instead on information theoretic methods or subsequences such as kmers or minimizers. In certain applications this can lead to significant improvements in efficiency.

Alignment-free algorithms are considered the state-of-the-art in RNA-seq transcript quantification [252, 253] and have been applied extensively in metagenomics [110]. In this chapter I present a toolkit for viral coinfection analysis using MinHash. MinHash is a kmer-based method that can calculate approximate Jaccard similarity by comparing only a representative subset of full data. This gives it performance that is sublinear to the total sequence length.

I developed a software package, `rkmh`, to calculate MinHash-based similarities and classify individual reads. I apply this toolkit to analyze highly-similar HPV16 sequences. HPV16 is the primary cause of cervical cancer. Specific genomic variants confer sigificantly different risks for cancer progression upon infection. I first show that this toolkit is robust to varying read lengths and error rates. The toolkit I developed is able to classify

individual reads at the HPV16 lineage and sublineage level across multiple sequencing technologies. Coinfections can then be detected based on the classifications of individual reads.

I then discuss future applications of similar kmer-based approaches. I show how kmers spanning heterozygous sites can be used to bin long reads from different haplotypes before genome assembly. Kmer-based methods have been used previously to remove duplicated portions after assembly [128, 127]. They have also been combined with trio data to separate maternal and paternal haplotypes in assembly [254, 126]. My approach for haplotype separation does not require sequencing the full trio but does require orthogonal read data from both short and long read technologies. Read classification using `rkmh` (including for HPV16 analysis) does not require both long and short reads; any contemporary long or short read technology may be used alone or in combination.

4.1.1 Publication and collaboration notes

The work in section 4.2 has been published in [96]. I performed all of the analysis in this paper with guidance and data from the other authors. I would also like to acknowledge members of the Durbin lab (particularly Shane McCarthy, Haynes Heaton, Dengfeng Guan, Erik Garrison, Markus Klarqvist, and Shilpa Garg) for discussions about using this method for genome assembly. This work has not yet been published; the software is available on https://github.com/edawson/fanthasm.

4.2 A MinHash toolkit for analyzing HPV coinfections

4.2.1 A MinHash toolkit for viral coinfection analysis

`rkmh` is a toolkit developed to help characterize HPV coinfections at the type and lineage level. `rkmh` makes use of the MinHash locality-sensitive hashing scheme, a technique developed for detecting similarity in webpages that has been previously applied in metagenomics [95]. Tools are included for classifying reads and removing contaminating sequences. A pipeline specifically for analyzing HPV16 lineage coinfections is also included. `rkmh` can classify a deep-sequenced HPV16 sample in minutes on a laptop computer. While all results here are from `rkmh` as applied to HPV, the tools are genome

agnostic and could be applied to other genomes of interest and read technologies without requiring any modifications.

4.2.2 Implementation

I developed `rkmh` based on methods introduced in [95], extending their algorithm to use various filters at the per-read level which improve classification performance (described in section 4.2.2). `rkmh` maintains information about type and lineage assignment on a per-read basis to enable estimation of relative abundances in a mixed infection.

`rkmh` is written in C++ and is threaded with OpenMP. It is freely available under the MIT open source software license at github.com/edawson/rkmh.

Hashing reads with `rkmh`

Much like Mash [95] and sourmash [255], `rkmh` relies on MinHash to transform reads for similarity comparison. Briefly, the algorithm works by generating all consecutive overlapping kmers of the read and hashing them with MurmurHash3 (Austin Appleby, https://github.com/aappleby/smhasher) to 64-bit integers. These integers are then sorted. A subset of size N of these hashes, usually the lowest N according to standard numerical ordering, are then chosen as a signature or "sketch" of the read. This effectively represents a sample of the kmers present in a read.

MinHash is locality-sensitive at the sketch level: reads which are more similar will share more kmers. By comparing only N integers, the number of comparisons per reference is reduced by $L - k - N$ where L is the length of the genome and k is the kmer size.

Classifying reads

Reads are classified by first generating the MinHash sketches for the reference sequences. A MinHash sketch is then generated for each read. All sketches use a single, fixed kmer size k and sketch size N. Abundance and uniqueness filters are optionally applied at this stage. Each read's sketch is then compared to each reference sketch. The intersection of the two sketches is calculated in $O(N)$ time where N is the sketch size. The read is then labeled as the reference with which the read shares the largest number of hashes.

Filtering kmers to improve classifications of individual reads

To improve specificity I implemented a set of kmer- and read-level filters in `rkmh` that are not offered by other MinHash-based classifiers. The `classify`, `stream`, and `filter` commands support four filters. The first is a floor for kmer abundance in reads ($-M$). As the reads are hashed the number of times each hash is seen is stored. Any hashes that do not meet the threshold for abundance are then excluded from a read's MinHash sketch. Reference [95] implemented this filter to remove sequencing errors in sketches of read sets. Here I have simply extended it to remove them in individual read sketches.

The second available filter is a ceiling on the number of times a hash may occur in the reference sequence set ($-I$). This filter is designed to remove repetitive kmers or those shared among many references, making them uninformative.

I also implemented a minimum difference filter ($-D$) that flags read sketches if the difference between the first- and second-best classifications is less than the desired threshold. This removes reads that cannot be given a unique classification because they come from genomic regions shared among references. Finally, a minimum number of shared hashes may be set so that reads that do not match well to any reference are flagged ($-N$).

Filtering reads

I initially tried assessing the performance of our type classifier on raw data but found that its performance was very poor, with high rates of supposedly false negatives. I performed a BLASTN [133] search on some of these reads to find that many of their top hits were in the human genome. I implemented a filter to deal with this at the classification level.

Such a feature is also useful in filtering a FASTQ file to find only reads which come from the organism of interest. The `rkmh filter` command implements the filters used in classification to filter reads. The `rkmh stream` command also implements an option for this, allowing real-time filtering of FASTQ reads during analysis.

Quantifying lineage and sublineage prevalence within a sample

Lineage and sublineage strains are differentiated mostly by SNVs and small INDELs. These polymorphisms alter the kmers of the sequence. If these kmers are unique among the reference sequence they can be used as a way of quantifying the strain they define. I implemented an exact kmer matching strategy in `rkmh` by removing all kmers that

appear in multiple references. This creates a minimal sketch that contains kmers unique to each reference sequence.

Each read is kmerized, hashed, and then compared against these reduced sketches. Reads that match well to a given reference sketch can be used to estimate the reference strain's abundance in that set of reads. This process has been wrapped in the `rkmh hpv16` command. When run in the `rkmh` directory, all reads in a fastq file can be labeled with their HPV type and HPV16 lineage/sublineage by running:

```
rkmh hpv16 -f <fastq.fq> > out.rk
```

The read classifications can be converted to lineage/sublineage prevalence estimates by running:

```
python scripts/score_real_classification.py < out.rk > out.cls
```

This will produce a file that contains a single line listing the estimated lineage and sublineage frequencies.

Run time performance of `rkmh`

`rkmh` was designed to scale to millions of reads and genomes megabases in size. Classifying over 400,000 Ion Torrent reads against all 182 HPV type references in PAVE requires less than one gigabyte of RAM and runs on a quad-core Intel desktop in 1 minute 16 seconds. In general, `rkmh` can process around 250,000 basepairs per core-second and scales well to increasing numbers of cores. Run times are dominated by sketch size and the number of reads as these two parameters affect the total number of comparisons to be made. Memory usage is dominated by the size and number of the reference genomes, meaning that there is not a major penalty for using long reads and that memory usage remains relatively constant over time. I have tested `rkmh` on ONT minION reads from genomes as large as 4.5 Mbp (*Escherichia coli* strain K-12) in under 16 GB of RAM using sketch sizes in the tens of thousands (data not shown).

`rkmh` output formats

There are three main output formats produced by `rkmh`. The outputs of the `stream` and `classify` commands are a tab-separated classification description similar to that produced by [95]. This format is easily manipulated using command line tools such as `grep`, `cut`, and `sed`, making analysis on any Unix system simple and portable. Additionally, the `rkmh hash` command can output sketches in JSON or the vowpal-wabbit vector format, a tab-separated format used by the vowpal-wabbit machine

learning package [256]. The version used by `rkmh` needs only to be labeled with its correct class by replacing a single sentinel string using `sed`. Sketches and vw-vectors may be computed for individual reads in a FASTA/FASTQ file or for the entire file.

Generation of simulated data

To assess the performance of `rkmh` I generated simulated read sets of coinfected and non-coinfected samples at known mixture proportions. I simulated reads at extremely high depth from 62 manually-prepared HPV16 sublineage reference genomes using DWGSIM (Nils Homer, https://github.com/nh13/DWGSIM). I set DWGSIM to create 225 basepair reads using the Ion Torrent error profile and flow order. This produced a set of large FASTQ files, one for each sublineage.

I generated random coinfections using the scripts at https://github.com/edawson/siminf. Briefly, `siminf` randomly selects an overall coverage to simulate along with a list of infecting strains and their relative proportion. A minimum of 5% strain abundance is required. `siminf` then samples our large sublineage FASTQ files to generate a FASTQ containing reads from the chosen sublineages in the desired proportions. Fifty of these simulated coinfections are available in https://github.com/edawson/rkmh_sim_data.

4.2.3 HPV typing performance across sequencing technologies is sensitive to kmer and sketch size

I assessed the HPV typing performance of `rkmh` on three datasets: simulated 100bp paired end Illumina reads based on the PAVE database of HPV reference genomes [257]; a real HPV16 sample sequenced on the Ion Torrent Proton platform (typical read length 250bp); and a set of 3,660 Oxford Nanopore minION reads generated from two HPV16 reference strains (typical read length over 6500bp). The minION reads typically cover the majority of the 7-8kb HPV genome, but have a relatively high error rate of 10% or more, comparable to the difference between HPV types and greater than that between lineages (they were collected in 2015 using the R7 pore).

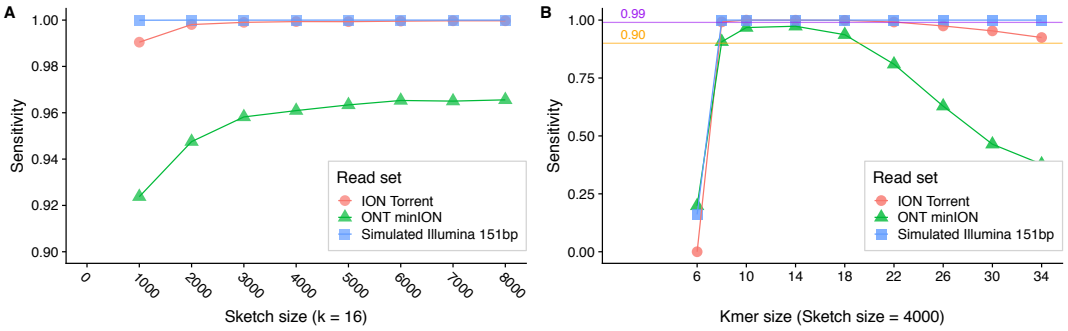


Fig. 4.1 Sensitivity of `rkmh` with respect to sketch size (A) and kmer size (B). There are diminishing returns to increasing sketch size above roughly 4000, regardless of read length. (B) shows that kmers are not sufficiently unique to classify reads with $k \leq 10$. Above $k = 18$, sensitivity begins to drop, likely due to the effects of incorporating sequencing errors into kmers. This is especially noticeable for ONT minION reads, which have a much higher error rate (above 12% per base for the R7.4 pore) compared to ION Torrent and Illumina ($< 0.1\%$ per base).

MinHash-based methods depend on a “sketch" which is a characteristic subset of kmers from a set of input sequences. Even at a low sketch size of 1000, `rkmh` correctly classifies more than 99% of the short reads and more than 90% of the nanopore reads (Figure 4.1A). As sketch size increases to 4000, per-read accuracy approaches 100% for short reads and 96% for ONT minION reads, with negligible improvements for sketch sizes higher than 4000. Sketch sizes below 1000 are not sufficiently sensitive for classifying HPV types, showing per-read accuracies well below 90%.

Kmer size is the main determinant of MinHash classification performance when errors are present. For HPV type classification we find that performance is diminished above $k = 18$ for our Ion Torrent reads and above $k = 14$ for our ONT minION reads (Figure 4.1B). This is due to the introduction of kmers containing one or more sequencing errors. The high per-base error rate of the ONT minION R7.4 pore (12% total per base [258]) means that as kmer size increases there is a rapid accumulation of kmers that do not match the reference because of incorporated errors, to the extent that for some reads no diagnostic kmer is found.

I compared the performance of `rkmh` to Taxonomer [259], a tool commonly used for metagenomic classification but which is not specifically designed for viral classification. On the set of 3,660 HPV16 minION reads, Taxonomer reported that 42.4% were of viral origin and 8.3% were from HPV16. It also reported 1,177 bacterial reads and 304 human reads; 398 reads were unclassified. `rkmh` reported 3,381 (92.4%) as HPV16. When I ran Taxonomer on a simulated 250bp ION Torrent HPV16 coinfection data set (discussed

further below), it reported that 29.2% of reads were HPV16, whereas `rkmh` reported that 94% of reads came from HPV16. In summary, Taxonomer has substantially lower sensitivity and specificity than rkmh for this type of data and analysis – this is not surprising since taxonomer is a general purpose metagenomics classification tool, which is not designed for medium to long read length viral sequence analysis.

4.2.4 Kmer pruning improves classification performance

It is possible to increase the type classification rate for minION reads by decreasing the kmer size, at the cost of introducing false positive assignments to other HPV types. However, this effect can be counteracted by removing kmers that are rare in the read set or enriching for those that distinguish between reference genomes. Such filters have been previously applied across read sets but not for individual reads.

I call this sketch modification process "pruning" and describe the individual filters in more detail in the section 4.2.2. Figure 4.2 shows the effect of pruning readset kmers on the ability of `rkmh` to classify Ion Torrent and minION reads. Increasing read pruning via the M parameter has a negligible effect on Ion Torrent reads as they have a low error rate ($<< 1\%$) and are relatively short; the majority of information available in them is acquired using just the default `rkmh` settings. MinION reads, while possessing a higher error rate, also possess many more kmers, meaning that dropping an erroneous kmer from the read sketch makes room for a possibly informative one. By dropping the kmer size from $k = 16$ to $k = 10$ and increasing the readset pruning threshold, both precision and recall of read classification improve by roughly 2% (Figure 4.2C).

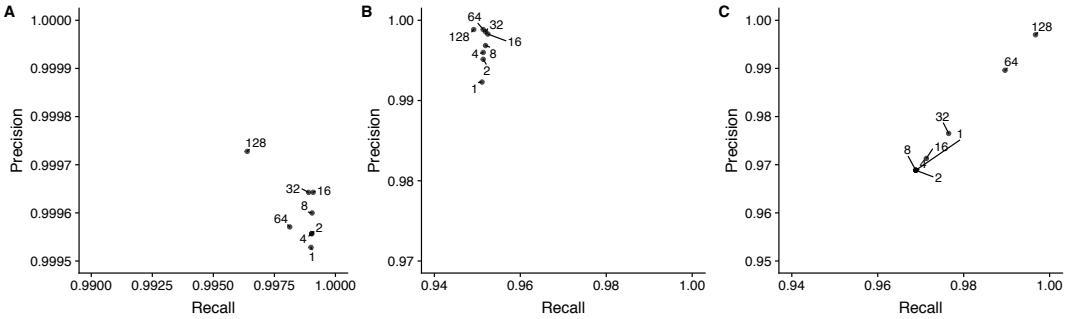


Fig. 4.2 Precision/recall plots for type classification of 70,000 Ion Torrent reads from an HPV16 amplicon sequencing reaction (A) and 3,660 ONT minION reads derived from two HPV16 isolates (B, C) at various read sketch pruning levels M indicated by the label attached to each point. Read sketch pruning removes rare kmers in the read sketch which might be random sequencing errors. (A, B) were classified using a kmer size of 16 and (C) was classified using a kmer size of 10. Ion Torrent reads have low substitution error rates, so pruning removes few kmers and the precision boost is small (<0.001%) (A). ONT minION reads have a much higher error rate approaching 10% per-base. For minION reads, pruning is able to improve precision to roughly 99.8% when using a kmer size of 16 (B). A smaller kmer size of 10 combined with high levels of pruning lead to an increase in both precision and recall, with precision and recall increasing from slightly more than 97.0% to over 99% (C).

These results demonstrate that `rkmh` is suitable for HPV typing. More than 90% of the individual reads match their known correct HPV type across Ion Torrent, ONT minION, and simulated Illumina datasets. Kmer pruning can further improve classification performance for long, noisy reads. From these per-read classifications one can determine the proportions of the infecting types by tallying the number of reads that support each type.

4.2.5 Accurate read classifications enable accurate percent composition estimates of HPV types

I next simulated a coinfection of HPV16, 18, and 31 by combining at equal proportions Ion Torrent reads from known samples of a single HPV type. I also examined the same sample after removing reads which did not map to the HPV genome(s), of which there are many (Figure 4.3A).

I summed the number of reads classified by `rkmh` to each HPV type with more than 5 kmers and divided each sum by the total number of reads classified to estimate the percent prevalence. `rkmh` is able to detect all three HPV types, though their proportions

are off by 5-15% (Figure 4.3B). Most of the reads are unclassified. It is to be expected that many of the unclassified reads may contain bits of human sequence and that our HPV18 sample appears over-reported simply because it had the most HPV DNA of the three. When restricting to reads that map to the HPV16, HPV18 or HPV31 genomes, `rkmh` accurately classifies over 99% of the reads into the correct type at the default settings (Figure 4.7). `rkmh` produces essentially perfect estimates of percent composition on this filtered subset.

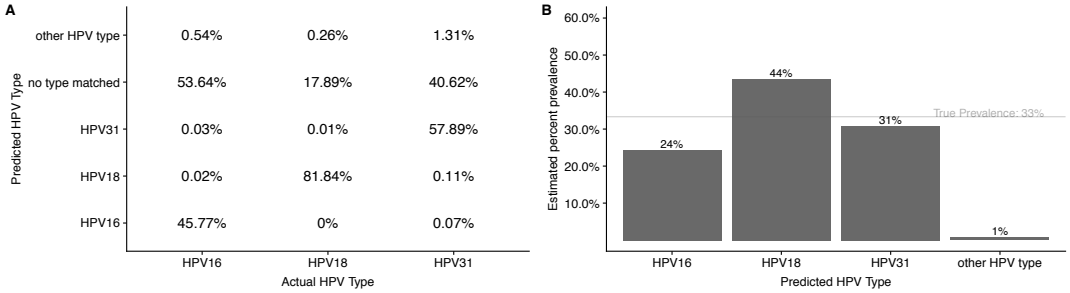


Fig. 4.3 (A) The performance of `rkmh` on a simulated HPV type coinfection. Summing the rows of this matrix gives percent prevalence estimates for each type (B).

I then applied `rkmh` to ten real samples amplified using a universal HPV primer scheme, sequenced on the ION Torrent and annotated with infecting HPV types by manual review. In eight out of the ten samples, `rkmh` correctly identified all of the manually annotated types using the default parameters (k = 16, s = 1000, threshold $\geq$ 1% or $\geq$ 1000 reads) (Supplementary Table 1).

The two samples where the classifications differ involved marginal decisions. For one sample, a type that had not been previously annotated was reported with 1.4% of reads assigned to it. For another sample a previously annotated type only received 942 reads, just below the reporting threshold of 1000. This was still greater than 20 times more than the next highest type (41 reads), and so could have been examined as a borderline case without generating noise. Based on the performance of `rkmh` on both our simulated set and ten real samples, we believe it is providing reliable type estimates in line with previous annotations.

4.2.6 Classification and quantification of HPV16 lineage coinfections

HPV16 lineages and sublineages differ by less than 10% of L1 sequence. HPV16A and HPV16D differ the most among HPV16's lineages but still share more than 97%

identity. Within the A lineage the A1, A2, A3, and A4 sublineages differ by less than 1% (Figure 4.4). MinHash similarity estimates and nucleotide similarity are highly correlated ($r = 0.9947$), but MinHash estimates show a bigger spread than nucleotide similarity because a single base change affects the k adjacent kmers. In essence, MinHash (and kmer-based methods in general) exaggerate differences between sequences, compared to direct string comparison.

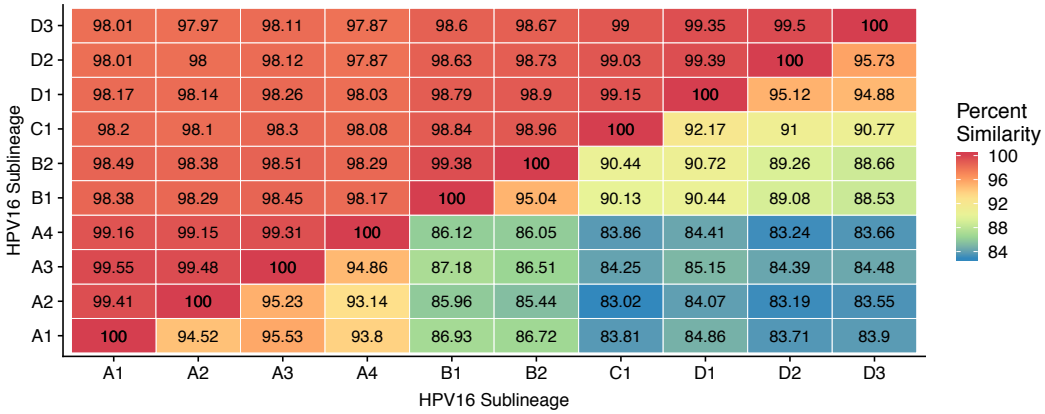


Fig. 4.4 Percent similarity for HPV sublineage; numbers above the diagonal are nucleotide similarity. Numbers under the diagonal are similarity estimates based on the number of shared hashes from `rkmh` .

To assess `rkmh`'s ability to discriminate coinfecting lineages using sketch pruning, I simulated a coinfection of HPV16 A4 / C / D3 in a 54:26:20 ratio. The per read performance as well as `rkmh`'s estimated percent composition of our sample at various parameterizations are shown in Figure 4.5. At the default settings (i.e. the standard MinHash algorithm, k = 16, s = 1000) there is a large amount of noise in the lineage classifications and the estimated percent compositions are similarly affected. Sublineage A1 is estimated to be the dominant sublineage even though no reads from sublineage A1 are present.

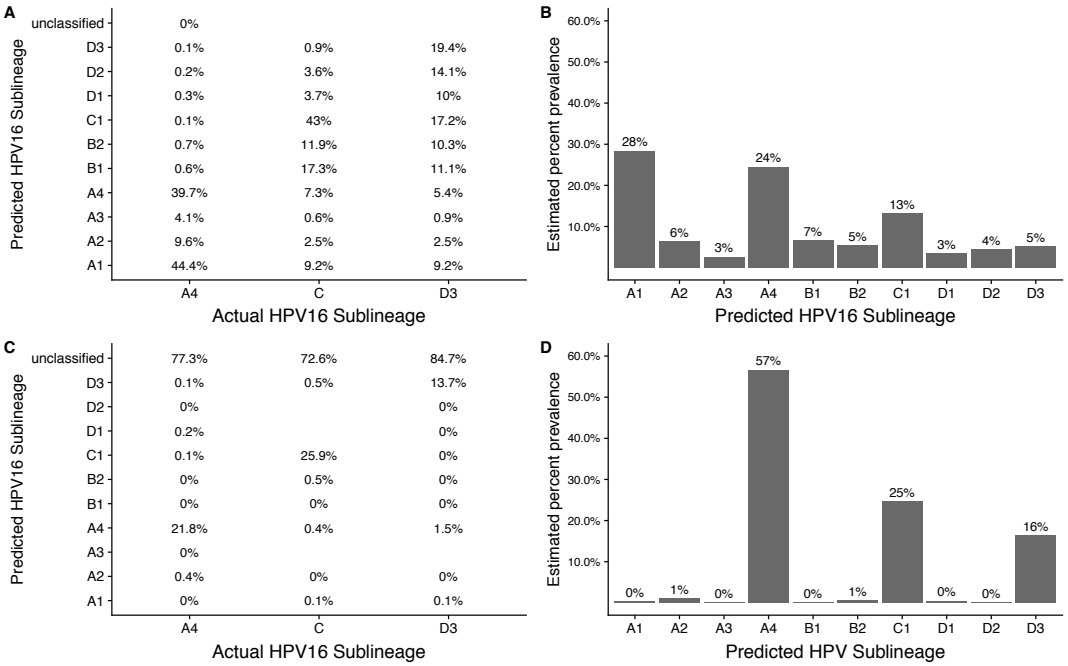


Fig. 4.5 (A) The percentage of reads from a simulated coinfection classified by `rkmh` to each of the HPV16 sublineages, at default settings (k = 16, s = 1000, no pruning, no difference filter). Summing each row of (A), with the exception of reads that couldn't be classified, gives the percent prevalence estimate of each sublineage (B). (C) The percent of reads classified to each sublineage by `rkmh` at pruning level M = 100 and I = 1. This significantly improves the prevalence estimates (D).

I applied sketch pruning to remove kmers that are shared among sublineages (see section 4.2.2 for details). At I = 1 each kmer in a reference sketch will be unique to a single sublineage. This effectively removes shared portions of the genome and reduces the MinHash procedure to exact kmer matching.

Raising the pruning level to I = 1 is sufficient to reduce erroneous read classifications from approximately 30% of reads misclassified to less than 5%; this comes at the expense of 60-90% of reads from each sublineage being removed from analysis (Figure 4.5C). This leads to much better estimates of sublineage prevalence (Figure 4.5D). Pruning is more effective at removing false classifications than simply requiring a minimum number of differences between a read's two best classifications (a filter implemented in other MinHash packages) ($s = 8000$, $D = 20$; not shown). Sketch pruning at $I = 1$ does not meaningfully affect type classification (not shown).

For the HPV16 specific workflow, I use the set differences of sublineage hashes to strictly remove kmers that appear across multiple sublineages. This enforces that each kmer appears in only one sublineage sketch; this provides only a minor improvement

over the standard pruning implementation (Figure 4.6), which is much faster. These results are representative of repeated tests on simulated coinfections (data available at https://github.com/edawson/rkmh_sim_data). The overall correlation between `rkmh` estimated prevalence and the true sublineage prevalence is 0.95.

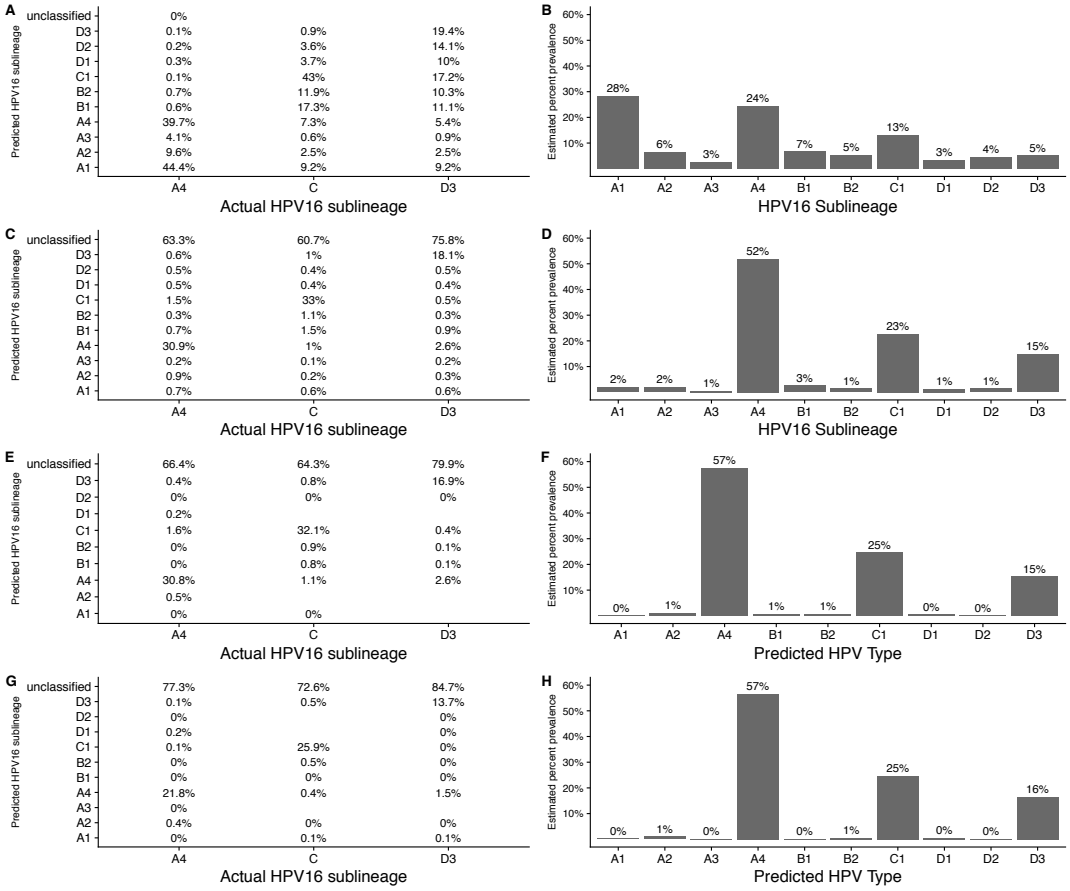


Fig. 4.6 Sublineage classification of simulated reads and corresponding prevalence estimates using different rkmh runtime parameters. (A) Per-read classification rates at the default settings (s = 1000, k = 16, no pruning) are poor, with many off-target matches. (B) This is reflected in the prevalence estimates, where a high proportion of sublineage A1 is reported though no A1 reads were present. (C, D) Read classifications and prevalence estimates at (s = 8000, k = 16, I = 1 and M = 5) are significantly improved, though still somewhat noisy. (E, F) Performance at (s = 8000, k = 16, I = 1, M = 100) is further improved; pruning of the read sketches leads to better read classifications. (G, H) Classifications using the hpv16 pipeline, which is equivalent (s = 8000, k = 16, I = 1, M = 5) but using a strategy that removes all kmers shared across references, rather than the approximate technique used by the sketch command.

I next performed a systematic analysis of the effects of divergence, read length, and error rate on read classification performance. I simulated three lineage references A, B, C

with random divergence rates 0.5%, 1%, 2.5% from the HPV reference. Then I simulated 3 sublineages A1, A2, A3, B1, B2 etc. at random divergence distances 0.05%, 0.1%, 0.25% from each of their lineage references. Then, for each reference set I simulated a million reads, selected evenly from these sublineages for each of the following sequence models, chosen to reflect the range of different read lengths and error rates available in practice:

75bp 0.1% error (short Illumina)
150bp 0.5% error (long Illumina)
250bp 1% error (IonTorrent)
5000bp 10% error (long read single pass)
5000bp 1% error (long read multi-pass)

The design of three potential references at both lineage and sublineage level facilitates evaluating false positive rates in terms of assignment to the lineage and sublineage not present in the data, as well as sensitivity in terms of correct assignment. For reads 250bp or longer, $> 80\%$ of reads are correctly classified to their known lineage, and pruning could reduce false positive assignments to almost zero (Figure 4.7). We therefore expect `rkmh` to produce accurate lineage quantifications for ION Torrent data.

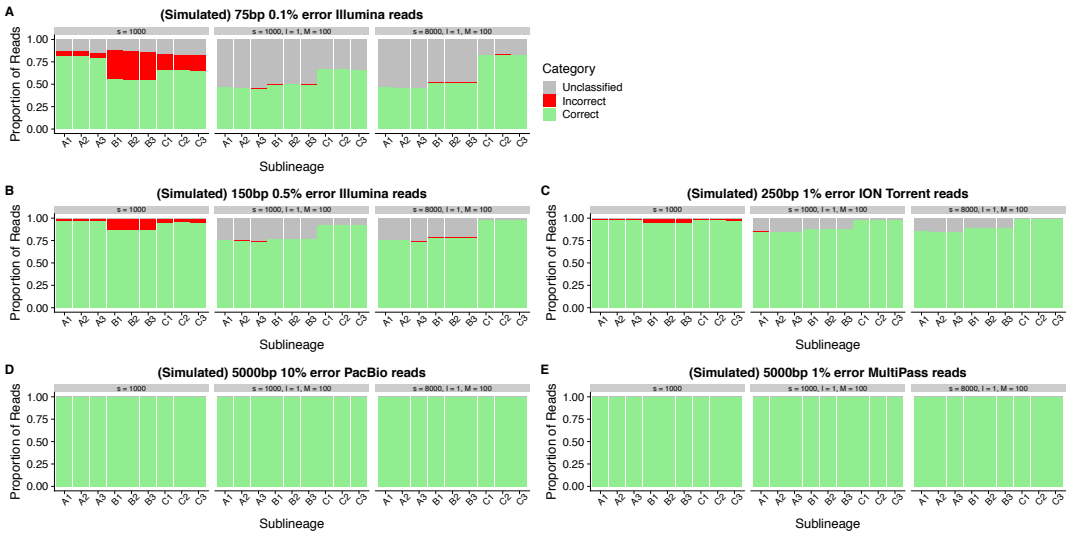


Fig. 4.7 Per-read lineage classification performance on different simulated HPV16 sublineage read sets. Lineage classification performance increases with read length and divergence. Short (75bp) Illumina reads show the worst classification performance, likely because a read may not be long enough to capture a lineage-defining SNP (A). Performance on 150bp reads is much better, and false-positive assignments are almost completely removed using kmer pruning (B). For simulated 250bp ION Torrent reads, rkmh correctly assigns over 80% of reads to their lineage at all parameter combinations tested (C). For 5000bp reads, rkmh is 100% accurate at lineage classifications across the spectrum of genome divergence and error rates (D, E). For 250bp and longer reads, we expect relatively accurate quantifications of infecting lineages given that most reads are correctly classified.

At the sublineage level, `rkmh` performed poorly at default parameters across read types (as expected) but kmer pruning could reduce the false-positive sublineage assignments to less than 0.1% of reads (Figure 4.8). Sublineage sensitivity was largely determined by divergence from the reference, with two-fold differences in the percentage of reads correctly classified between 0.05% and 0.25% divergence. While this can bias estimated proportions for sublineages, individual read classifications using kmer pruning are highly specific, indicating that `rkmh` can still detect the presence or absence of sublineages based on the presence of high-confidence read assignments.

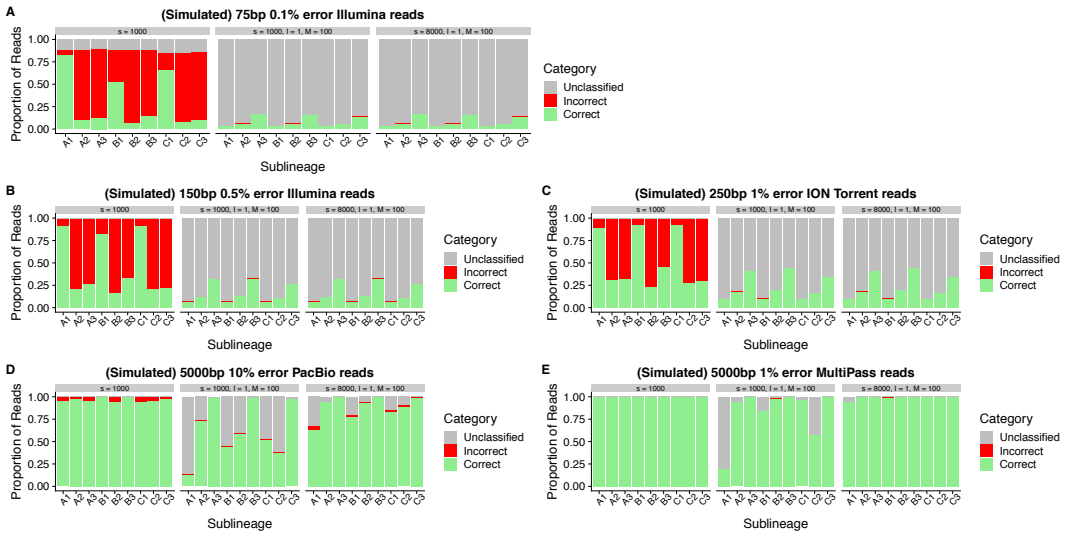


Fig. 4.8

Since `rkmh` can characterize simulated coinfections adequately, I assessed its performance on real coinfections identified in samples from Mirabello *et al.* 2016 [4].

N = 34 manually annotated samples	Agrees with annotations	Disagrees with annotation	Concordance
Primary Lineage	32	2	95%
Primary Sublineage	31	3	91%
Secondary Lineage	24	10	71%
Secondary Sublineage	12	22	35%
Coinfection status, lineage	27	7	79%
Coinfection status, sublineage	24	10	70%

Table 4.1 Performance of `rkmh` on samples from [4] which were manually reviewed for their infecting sublineages and coinfection status.

In roughly 90% of real cases I examined `rkmh` agreed with the manually annotated predominant infecting lineage and sublineage (Table 4.1). There was good concordance (70% or more) with manual annotations for coinfection status, where I consider a sample coinfected if a second lineages/sublineage is represented in at least 1% of reads. `rkmh` can identify a coinfected secondary lineage with similar accuracy.

However, the performance of `rkmh` to identify any secondary sublineage(s) is only 35%. Further review of samples for which `rkmh` did not agree with the manual annotations indicated that many had characteristics which make them difficult or impossible to

correctly classify. In some samples, the two dominant sublineages had frequencies that were close to equal and `rkmh` correctly predicted the infecting sublineages but not their order. When a sample possessed a sublineage not in the reference set, `rkmh` often predicted the correct lineage but assigned reads evenly among the sublineages in the family. This sometimes falsely indicated a coinfection was present at the sublineage level. Lastly, a small proportion of samples I examined were of low coverage or quality and had no reads that could be used for classification.

4.2.7 Pitfalls and improvements

There are various factors that can lead to biases or incompleteness in the application of `rkmh` . In the unique kmer matching sketches, each sublineage is defined by between 145 and 440 unique kmers. HPV sublineages with more available unique kmers may be more detectable, biasing results toward more divergent sublineages. It is also important to note that the amplicon sequencing scheme used to sequence the Ion Torrent samples does not produce consistent depth across the genome. If mutations are not randomly distributed, and regions of diversity are not evenly sequenced, this difference in depth could reduce the correlation between kmer prevalence and strain prevalence.

All of the real data used in this work were produced by amplicon approaches, and so should not include fusions with host DNA. However, if such sequences were present due to other enrichment approaches, they might increase noise and reduce signal for some reads. Fusions should not lead to biases, assuming multiple integration sites. Long reads from single-molecule sequencing should provide more specific per-read classifications and therefore better estimates of sublineage prevalence once the technology becomes cost efficient. MinHash, while a viable method when strain prevalences are high, may not be a viable estimator of very low-prevalence ($\leq$5%) coinfecting lineages and sublineages.

It can be expected that not all HPV16 sublineage isolates to perfectly match our reference genomes as the virus continues to evolve, albeit slowly. Many of the secondary sublineage classifications which were labeled "incorrect" may well be isolates harboring mutations present in multiple sublineages. This highlights the fact that classifications are only as good as the reference panel. In an early run of the pipeline, the sequence for sublineage A2 was mistakenly left out. This had a significant impact on sensitivity for non-A lineage reads as many reads were discarded in A2-infected samples. The upside of this is that future domain knowledge may yield even better classifications. Accounting for the phylogenetic nesting of sublineages within lineages can somewhat improve results in such situations. While `rkmh` incorrectly labeled reads from the A2 sublineage, it was able

to accurately label the reads as originating from the HPV16A lineage. In general, `rkmh` is better able to assign lineages than sublineages because the amount of discriminating information is greater. A postprocessing script is included in the repository for scrubbing incorrect sublineage labels based on lineage assignments.

It should also be noted that our reference set is based on annotations that were performed by hand in IGV and may contain mistakes and differences in opinion. In particular, some of the errors at the level of secondary lineage / sublineage may be affected by variation in reference classification. As each read is independently classified, this may indicate that some of samples require further manual review.

With respect to possible future improvements to `rkmh` , Ondov *et al.* discuss possible performance improvements to the MinHash scheme in [95]. Sequence Bloom Trees are data structures that would allow MinHash sketch comparison in logarithmic rather than linear time. This could make `rkmh` both more time- and energy-efficient.

An alternative to the Sequence Bloom Tree would be to use the minimizer database described in [110] to assign genus-level labels to reads in metagenomic samples, though the kmer sizes we use for HPV16 classification may be too small to make this sensible.

Additionally, many existing packages support pre-hashing sequences, which amortizes the expense of this procedure over later comparisons. `rkmh` will implement this in a future release. `rkmh` also removes the p-value defined in [95], which becomes harder to interpret on a per-read basis and which is affected in complex ways by the various filters in `rkmh` .

Several modifications to the sketching procedure might improve classification performance. Skip-grams (kmers generated from genomic substrings length $\frac{k}{2}$ separated by a small, fixed distance) would improve classification if genomes share rearrangement patterns. Using minimizers, where sketches are composed of hashes sampled from rolling genomic windows (rather than randomly sampling the entire sequence as in MinHash) would provide more even coverage of the reference sequences, possibly improving the chances of a read matching. Dynamic sketch sizes based on the length of the query sequence (rather than a fixed sketch size) might provide a slight improvement in runtime. Classification might be improved by introducing machine learning techniques trained on full sketches, as the current supervised approach may overlook cryptic but important features. Finally, improvements in data quality from long, high-fidelity reads should yield a large improvement in results when such data becomes available, and could be instrumental in advancing scientific inquiry and eventually developing effective public health measures to address HPV infection.

4.2.8 Summary and future directions

HPV is a common sexually-transmitted agent, and a small subset of HPV infections become chronic and can lead to cervical, anogenital or oropharyngeal cancer. Twelve of at least 170 known HPV viral types are currently associated with cancer risk, and sublineages within these carcinogenic types are further associated with variable risks.

Confounding proper classification of HPV infections is the prevalence of multiple types, lineages, and sublineages in individual infections. Thus, the accurate detection of HPV types, as well as HPV16 lineages and sublineages, could have important pleiotropic implications for public health measures.

I developed a computational toolkit to classify coinfected HPV samples, as in [4]. This method, `rkmh`, is a collection of tools that addresses some of the challenges associated with analyzing mixtures of biological sequences.

To implement `rkmh` I extended existing work utilizing the MinHash locality-sensitive hashing scheme [95], resulting in a tool that provides accurate classifications of individual reads. Accurate classification of the infecting viral types, lineages and sublineages is critical given the vast differences in disease risk between HPV types and even closely related HPV16 sublineages. The toolset demonstrates that accurate classification of individual reads and estimation of type and lineage prevalence is possible with current sequencing practices, but that sensitive sublineage detection may require improvements in technique.

While applied here to HPV, `rkmh` could be used in any context where quantification of specific sequences within a mixture and selection for or removal of such sequences might be useful. MinHash has previously been applied to larger metagenomic datasets with striking success. Ondov *et al.* demonstrate MinHash's ability to work on genomes several megabases in size and scale to billions of reads in [95]. Other viruses show significantly more intra-host variation than HPV; notably, Human Immunodeficiency Virus (HIV) evolves during infection and in response to treatment [260]. Zika and Ebola are urgent public health threats, have been shown to evolve over the course of outbreaks, and have been successfully sequenced in the field on the ONT minION [261–263]. The ability to generate per-read classifications using `rkmh` on a standard laptop could be a useful addition to the current pipelines employed by these studies. Lightweight algorithms such as `rkmh` may also be of interest in areas with strict computing power limitations such as space genomics or developing nations.

4.3 Potential frontiers for lightweight algorithms and long reads

Reference [252] provides a clear definition of lightweight algorithms as those that use data sparingly, consider and attempt to reduce constant factors in algorithmic complexity, and utilize parallel computing hardware effectively. `rkmh` is able to deconvolute individual reads from HPV16 coinfections because the core algorithms are aware of lineage-distinguishing variants *a priori*. As only 1-2% of the genome is sufficient to distinguish subtypes, the total amount of information that must be processed is significantly reduced compared to performing alignment. From individual read classifications, a picture of the overall mixture of many components can be gleaned. Kmers are selected for HPV16 classifications based upon their uniqueness across strains. The same approach used in `rkmh` has potential applications in the field of genome assembly, especially in the separation of reads from different haplotypes.

4.3.1 Separating haplotypes for individual assembly using kmer methods

There are strong similarities between genome assembly and metagenomic classification. Reads which are much shorter than the chromosomes are mixed together and must be separated to reconstruct the original genomic sequence. The extra overlap information contained in long reads is useful in both.

Linear references are haploid, but most eukaryotes have a ploidy of at least two. For most non-asexual organisms, each copy of the genome will differ from the other due to mutations accumulated since their last common ancestor. For humans, this percent difference (i.e., the heterozygosity rate) is approximately 0.1, meaning that on average every 1000 basepairs there is a single difference in the genome.

To produce a haploid reference assembly the two (or more) slightly different haplotypic sequences should be separated.

However, assembly algorithms typically merge haplotypes together. When they do this they can arbitrarily switch between haplotypes (i.e., incorrectly phase the haplotypes), or, if the haplotypes are sufficiently different, incorporate them both into the primary assembly as duplicate sequence. This effectively introduces repetitive structures with minor variations into the reference where none are present. It can also greatly increase the size of the final assembly.

Unique kmers which span heterozygous sites in the genome are present at one-half the frequency of kmers which span homozygous sites because one kmer comes from each of the two alleles. We can extract kmers that span heterozygous sites then by counting all kmers in the genome and selecting those which are present at a frequency $\frac{1}{2}$ of the homozygous frequency. This approach is analogous to the one used in [254] to bin reads by parental haplotype; their approach required data from both parents, however, which is not always available.

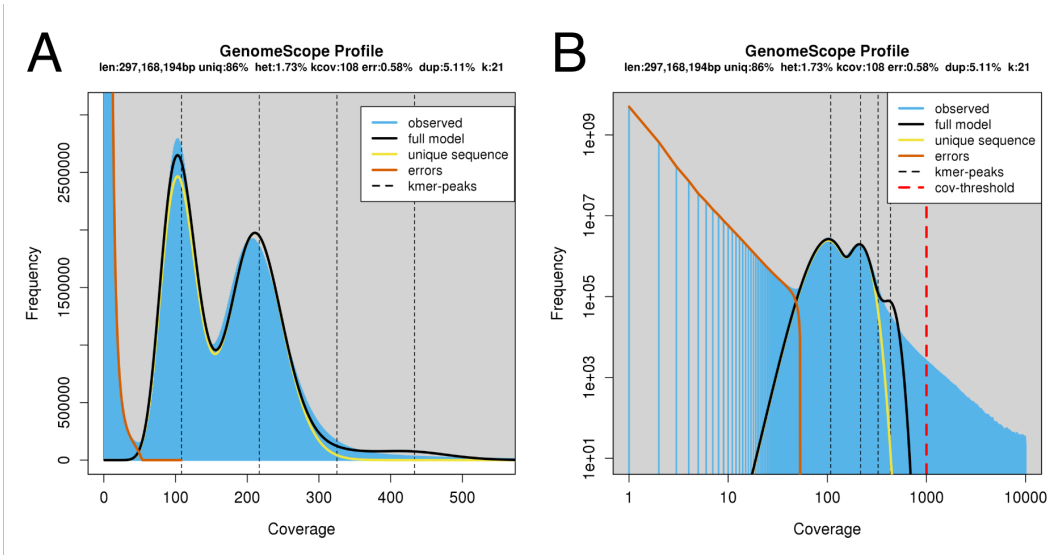


Fig. 4.9 Kmer frequency spectra from GenomeScope for deep Illumina data of a *Heliconius sara* specimen. (A) Raw frequencies show the two clear frequency peaks of kmers at heterozygous (app. 100X) and homozygous (app. 200X) sites. (B) Shows the same data with log-transformed frequencies, showing that kmers below approximately 50X are likely due to errors.

We examined the kmer spectra of deep Illumina reads from a *Heliconius sara* specimen, a butterfly that lacks a reference genome and for which we did not have parental sequencing data (Figure 4.9). *Heliconius sara* has an estimated heterozygosity rate of 1.73%, nearly twenty times that in humans. A clear kmer frequency peak is seen at 200X, representing kmers spanning homozygous regions, and another peak is seen at 100X (kmers spanning heterozygous sites).

To generate normalized representative kmers of heterozygous sites, we assembled unitigs for 21-mers with a frequency between 50X and 150X using BCALM2 [264]. Kmer frequencies were calculated with jellyfish [265]. For any heterozygous site, there will be two separate unitigs, one containing each of the two alleles. Any heterozygous sites within 21bp of each other will be merged into a single pair of unitigs.

I next labeled a deep PacBio read set with these unitigs using fanthasm. `fanthasm` takes a file containing unitigs and a set of PacBio reads. Unitigs are stored in a map of their constituent 21-mers (hashed to an integer) to the unitig identifier. Each read is then kmerized, hashed, and labeled with any unitigs with which it shares kmers. A MinHash sketch of each read is produced as well since the sorted list of kmer hashes is required for the heterozygous kmer search. MinHash has been used previously for pre-assembly overlapping of reads in MashMap [134]. Manual review of several of the longest reads showed that they contained mutually exclusive heterozygous unitigs. Ideally, reads would next be examined for overlap by MinHash and haplotype of origin by heterozygous unitigs before being assembled, however I did not implement these processes.

4.3.2 Assembly of individual tumor subclones

We are able to locate putative haplotig-defining kmers in the *Heliconius sara* genome because they are present at an allele fraction of fifty percent. This makes sentinel kmers for heterozygous sites observable as a lower-peaked distribution in Figure 4.9. Tumors often have a more complex allele frequency spectra confounded by alternative ploidies, loss of heterozygosity, and heterogeneity in the form of subclonal mutations and normal cell contamination.

Reference [46] demonstrated the ability to detect subclones in tumor samples by ultra-deep (300X) whole genome sequencing using the SciClone algorithm [266]. This algorithm was initially validated on a 188X whole genome sequenced multiple myeloma sample. Such sequencing is financially impractical but demonstrates that it is possible to partition subclonal mutations by variant allele frequency. A list of such mutations, combined with long reads from the tumor, should be sufficient to generate kmer frequencies and unitigs to partition long reads by tumor subclone as was performed in subsection 4.3.1.

4.3.3 Point-of-care testing

A major advantage of both alignment-free and variant-aware algorithms is their speed. Because much less information needs to be processed, relevant results are available more quickly. These algorithms tend to also make efficient use of computational hardware. In the case of `rkmh` , as little as 1% of the genome is sufficient to classify HPV sublineages. The data structures necessary to perform classification can fit in well under one Gigabyte of RAM.

There were two primary datasets that motivated the development of `rkmh` . The first was a large sample set of over 200,000 Pap smear samples, some of which we examined in [96]. While we have not yet applied `rkmh` to the full dataset we did calculate read classifications for 700 of these samples and 2,000 in-silico simulated coinfections (data not shown). In general, `rkmh` can classify a single sample using one core in less than five minutes. This reduced the estimated cost of analyzing these samples significantly compared to the existing pipline, making processing of all 200,000 feasible and affordable. The second dataset was the National Cancer Institute's HPV vaccination and point-of-care testing initiatives in Costa Rica. While there are still many steps before clinical sequencing of HPV infections is routine, `rkmh` could provide real-time offline analysis at point of care in this setting.

4.3.4 Lightweight algorithms improve analysis

References [267] and [252] revolutionized RNA sequencing analysis by greatly reducing the cost of the associated bioinformatics using lightweight algorithms. Combined with a drop in sequencing cost, this has greatly expanded the number of labs that can afford to sequence large numbers of samples. Sequencing has become increasingly portable as well, enabling applications such as real-time Ebola and Zika outbreak observation [268, 269].

As our catalog of known variation and reference-quality assemblies grows, these tools will further improve. Improvements in DNA sequencing analysis are following similar patterns. Reference [267] uses pseudoalignment of RNA reads to a transcriptome de Bruijn graph. `vg` and [98] provide graph-based alignment of reads to graph genomes for DNA. Alignment of DNA sequencing reads has long relied on sparse seeds, and research into better seeding strategies is ongoing. MiniMap2 [270] uses a seeding index based on minimizers (rather than maximal exact matches as in [271]) to align both long and short reads in much less time and with comparable accuracy to previous approaches. These approaches are being actively explored for graph alignment as well (Jouni Síren and the vgteam, github). The application of lightweight principles to graph genome implementations will make these approaches practical, greatly expanding our ability to interrogate genomic variation using more sensitive approaches.

Chapter 5

Conclusions

In this thesis, I describe an analysis of several hundred radiation-exposed tumors from the Chernobyl Tissue Bank (CTB). I also present new tools for variation-aware analysis using graphs and lightweight algorithms. Such algorithms can improve both the speed and sensitivity with which we find variants by incorporating information from previous studies.

Tumors from the Chernobyl Tissue Bank show dose-dependent signs of radiation exposure in their genomes. Erroneous repair of radiation-induced double strand breaks leads to an excess of small deletions and large, balanced inversions and translocations. These findings are significant in that they describe a genomic marker for radiation exposure as well as evidence for the mechanism by which ^{131}I exposure increases the risk of thyroid cancer development.

These tumors are driven exclusively by activating mutations in genes involved in the $MAPK$ signaling pathway, highlighting its role in thyroid tumorigenesis. Genes essential to normal thyroid function (TG and $TSHR$) are mutated above background frequency but only occur in the presence of mutations in $MAPK$. The number of gene fusion driven tumors increases with increasing exposure to radiation. The genes involved in these fusions are $MAPK$ signaling genes (especially RET) which become constitutively active. Fusion breakpoints are highly recurrent. This is most likely attributable to the need for these fusions to preserve specific domains in the fusion product. RET fusions always begin at the eleventh exon because they must preserve the kinase domain to be functional.

While we will always require algorithms that can find novel variants, we can save significant amounts of time by utilizing variation aware algorithms. This is especially true when we are interested in specific variants such as driver mutations or the lineage-defining

sites of HPV16. In the CTB dataset, just five mutations — $BRAF^{V600E}$, *RET-CCDC6*, *ETV6-NTRK3*, $NRAS^{Q61R}$, and $HRAS^{Q61R}$ — account for 67% of the driver mutations found, with $BRAF^{V600E}$ alone acting as the primary driver in 47% of tumors. Combining low-pass sequencing and variation-aware algorithms in a digital gene panel may one day provide sensitive genotyping at lower cost and with less bias than current state-of-the-art targeted sequencing approaches.

Genome graphs provide a powerful variation-aware implementation for genotyping both small and large variants. Graph-based approaches consistently outperform structural variant genotypers which rely on the linear reference. Their supremacy in genotyping inversions makes them especially promising in cancer genome analysis. Though such variants are rare in the general population they are common and clinically important in cancer, particularly radiation-associated papillary thyroid carcinoma cases.

`vg` provides methods for calling structural variants *de novo*, but such approaches are still nascent. Graph deconstruction and snarl-based variant calling have relied on augmenting the graph with reads, though there is active work to remove this limitation. In subsection 3.4.2 I describe methods for locating structural variant signatures in reads mapped to a graph. I hope that these approaches will enable development of downstream variant callers analogous to those developed a decade ago for linear reference alignments. With further development, breakpoint homogenization may be used to further refine these calls. Graphs hold significant promise to improve our knowledge of structural variation in healthy and morbid populations.

While our variant calling pipelines used in analyzing the CTB samples should benefit from the application of graph-based regenotyping, doing so is prohibitively expensive with current techniques. Many studies have avoided realigning to updated linear references for this reason, and read-to-graph alignment currently takes 3-5 times as long as alignment to a linear reference. As graph alignment algorithms continue to improve it will become affordable to reanalyze our data. I describe areas for improvement in graph-based structural variant calling and genotyping in Chapter 3, but even with these changes alignment will remain the most expensive operation. Improvements in alignment speed will be necessary to make graph-based analysis routine, and research in this area is highly active.

MinHash and other lightweight algorithms can significantly speed up specific tasks and reduce costs. Sparse seeds, particularly minimizers, are used extensively in alignment and are one method being explored for improving graph-based alignment. I show in Chapter 4 how a MinHash-based toolkit for viral analysis improved HPV16 coinfection detection

while simultaneously making analysis practical on any modern compute hardware. This approach generalizes well to data from other species, new reference panels, and reads from many modern sequencing platforms. Adaptations may one day be useful in genome assembly and analysis of subclonal architecture.

Both graph genomes and the MinHash toolkit `rkmh` are examples of variation-aware algorithms. These have been applied extensively in population studies but much less often in cancer genomics. Notable exceptions to this generalization include the extensive use of previously-observed mutations for annotation and the use of panel-of-normals filtering to remove germline and passenger mutations from somatic variant calls. In this thesis I have demonstrated the power of variation-aware algorithms, both graph- and MinHash-based, for structural variant genotyping and HPV16 coinfection analysis. However many more applications can be imagined, some of which I have described in Chapters 3 and 4. The recurrent nature of mutations in cancer and the relatively small number of mutations of interest (compared to the total number often analyzed in population studies) make variation-aware approaches especially promising for cancer genome analysis. This is especially important as study sizes approach two orders of magnitude larger size than those of just a decade ago. The use of lightweight, variation-aware algorithms will help bring genomics into the era of million-genome studies and provide invaluable insights into human genetics and disease.

References

[1] Kijong Yi and Young Seok Ju. Patterns and mechanisms of structural variations in human cancer. *Experimental and Molecular Medicine*, 50(8), 2018.

[2] Lorenzo Tattini, Romina D'Aurizio, and Alberto Magi. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Frontiers in Bioengineering and Biotechnology*, 3(June):1–8, 2015.

[3] Glenn Hickey, David Heller, Jean Monlong, Jonas Andreas Sibbesen, and Adam Novak. Genotyping structural variants in pangenome graphs using the vg toolkit. *bioRxiv*, 2019.

[4] Lisa Mirabello, Meredith Yeager, Michael Cullen, Joseph F. Boland, Zigui Chen, Nicolas Wentzensen, Xijun Zhang, Kai Yu, Qi Yang, Jason Mitchell, David Roberson, Sara Bass, Yanzi Xiao, Laurie Burdett, Tina Raine-Bennett, Thomas Lorey, Philip E. Castle, Robert D. Burk, and Mark Schiffman. HPV16 Sublineage Associations with Histology-Specific Cancer Risk Using HPV Whole-Genome Sequences in 3200 Women. *Journal of the National Cancer Institute*, 108(9):1–9, 2016.

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

[6] Steven I. Hajdu, Manjunath Vadmal, and Ping Tang. A note from history: Landmarks in history of cancer, part 7. *Cancer*, 121(15):2480–2513, 2015.

[7] Samantha Hansford and David G. Huntsman. Boveri at 100: Theodor Boveri and genetic predisposition to cancer. *Journal of Pathology*, 234(2):142–145, 2014.

[8] Theodor Boveri. *Zur frage der entstehung maligner tumoren*. Gustav Fischer, 1914.

[9] Nowel PC and Hungerford DA. A minute chromosome in human chronic granulocytic leukemia. *Science*, 132:1497–1501, 1960.

[10] J Lejuene, M Gautier, and Turpin R. Etude des chromosomes somatiques de nuf enfants mongoliens. *Comptes rendus hebdomaires des seances de l'Academie de sciences*, 248:1721–1722, 1959.

[11] A. M. Maxam and W. Gilbert. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2):560–564, 1977.

[12] F Sanger and AR Coulson. A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase. *Journal of Molecular Biology*, 94(3):441–448, 1975.

[13] The Human Genome Project Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(February), 2001.

[14] Sara Ballouz, Alexander Dobin, and Jesse Gillis. Is it time to change the reference genome? *bioRxiv*, 465, 2019.

[15] The International and Hapmap Consortium. The International HapMap Project. *Nature*, 426(6968):789–96, 2003.

[16] The International and Hapmap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005.

[17] C. C. Buchanan, E. S. Torstenson, W. S. Bush, and M. D. Ritchie. A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. *Journal of the American Medical Informatics Association*, 19(2):289–294, 2012.

[18] The 1000 Genome Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 135(V):0–9, 2012.

[19] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[20] Peter H. Sudmant, Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, Markus Hsi-Yang Fritz, Miriam K. Konkel, Ankit Malhotra, Adrian M. Stütz, Xinghua Shi, Francesco Paolo Casale, Jieming Chen, Fereydoun Hormozdiari, Gargi Dayama, Ken Chen, Maika Malig, Mark J. P. Chaisson, Klaudia Walter, Sascha Meiers, Seva Kashin, Erik Garrison, Adam Auton, Hugo Y. K. Lam, Xinmeng Jasmine Mu, Can Alkan, Danny Antaki, Taejeong Bae, Eliza Cerveira, Peter Chines, Zechen Chong, Laura Clarke, Elif Dal, Li Ding, Sarah Emery, Xian Fan, Madhusudan Gujral, Fatma Kahveci, Jeffrey M. Kidd, Yu Kong, Eric-Wubbo Lameijer, Shane McCarthy, Paul Flicek, Richard a. Gibbs, Gabor Marth, Christopher E. Mason, Androniki Menelaou, Donna M. Muzny, Bradley J. Nelson, Amina Noor, Nicholas F. Parrish, Matthew Pendleton, Andrew Quitadamo, Benjamin Raeder, Eric E. Schadt, Mallory Romanovitch, Andreas Schlattl, Robert Sebra, Andrey a. Shabalin, Andreas Untergasser, Jerilyn a. Walker, Min Wang, Fuli Yu, Chengsheng Zhang, Jing Zhang, Xiangqun Zheng-Bradley, Wanding Zhou, Thomas Zichner, Jonathan Sebat, Mark a. Batzer, Steven a. McCarroll, Ryan E. Mills, Mark B. Gerstein, Ali Bashir, Oliver Stegle, Scott E. Devine, Charles Lee, Evan E. Eichler, and Jan O. Korbel. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, 2015.

[21] The Simons Genome Diversity Project Consortium. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206, 2016.

[22] Konrad J. Karczewski, Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, Kristen M. Laricchia, Andrea Ganna, Daniel P. Birnbaum, Laura D. Gauthier, Harrison Brand, Matthew Solomonson, Nicholas A. Watts, Daniel Rhodes, Moriel Singer-Berk, Eleanor G. Seaby, Jack A. Kosmicki, Raymond K. Walters, Katherine Tashman, Yossi Farjoun, Eric Banks, Timothy Poterba, Arcturus Wang, Cotton Seed, Nicola Whiffin, Jessica X. Chong, Kaitlin E. Samocha, Emma Pierce-Hoffman, Zachary Zappala, Anne H. O'Donnell-Luria, Eric Vallabh Minikel, Ben Weisburd, Monkol Lek, James S. Ware, Christopher Vittal, Irina M. Armean, Louis Bergelson, Kristian Cibulskis, Kristen M. Connolly, Miguel Covarrubias, Stacey Donnelly, Steven Ferriera, Stacey Gabriel, Jeff Gentry, Namrata Gupta, Thibault Jeandet, Diane Kaplan, Christopher Llanwarne, Ruchi Munshi, Sam Novod, Nikelle Petrillo, David Roazen, Valentin Ruano-Rubio, Andrea Saltzman, Molly Schleicher, Jose Soto, Kathleen Tibbetts, Charlotte Tolonen, Gordon Wade, Michael E. Talkowski, The Genome Aggregation Database Consortium, Benjamin M. Neale, Mark J. Daly, and Daniel G. MacArthur. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*, 2019.

[23] Ryan L. Collins, Harrison Brand, Konrad J. Karczewski, Xuefang Zhao, Jessica Alföldi, Amit V. Khera, Laurent C. Francioli, Laura D. Gauthier, Harold Wang, Nicholas A. Watts, Matthew Solomonson, Anne O'Donnell-Luria, Alexander Baumann, Ruchi Munshi, Chelsea Lowther, Mark Walker, Christopher Whelan, Yongqing Huang, Ted Brookings, Ted Sharpe, Matthew R. Stone, Elise Valkanas, Jack Fu, Grace Tiao, Kristen M. Laricchia, Christine Stevens, Namrata Gupta, Lauren Margolin, The Genome Aggregation Database (gnomAD) Production Team, The gnomAD Consortium, John A. Spertus, Kent D. Taylor, Henry J. Lin, Stephen S. Rich, Wendy 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. An open resource of structural variation for medical and population genetics. *bioRxiv*, page 578674, 2019.

[24] ST Sherry, MH Ward, M Kholodov, J Baker, L Phan, EM Smigielski, and K Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic Acids Research*, 29:308–11, 2001.

[25] Rachel M. Sherman, Juliet Forman, Valentin Antonescu, Daniela Puiu, Michelle Daya, Nicholas Rafaels, Meher Preethi Boorgula, Sameer Chavan, Candelaria Vergara, Victor E. Ortega, Albert M. Levin, Celeste Eng, Maria Yazdanbakhsh, James G. Wilson, Javier Marrugo, Leslie A. Lange, L. Keoki Williams, Harold Watson, Lorraine B. Ware, Christopher O. Olopade, Olufunmilayo Olopade, Ricardo R. Oliveira, Carole Ober, Dan L. Nicolae, Deborah A. Meyers, Alvaro Mayorga, Jennifer Knight-Madden, Tina Hartert, Nadia N. Hansel, Marilyn G. Foreman, Jean G. Ford, Mezbah U. Faruque, Georgia M. Dunston, Luis Caraballo, Esteban G. Burchard, Eugene R. Bleecker, Maria I. Araujo, Edwin F. Herrera-Paz, Monica Campbell, Cassandra Foster, Margaret A. Taub, Terri H. Beaty, Ingo Ruczinski, Rasika A. Mathias, Kathleen C. Barnes, and Steven L. Salzberg. Assembly of

a pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics*, 51(January), 2018.

[26] David A. Wheeler and and Linghua Wang. From human genome to cancer genome: The first decade. *Genome Research*, 23(7):1954, 2013.

[27] Timothy J. Ley, Elaine R. Mardis, Li Ding, Bob Fulton, Michael D. McLellan, Ken Chen, David Dooling, Brian H. Dunford-Shore, Sean McGrath, Matthew Hickenbotham, Lisa Cook, Rachel Abbott, David E. Larson, Dan C. Koboldt, Craig Pohl, Scott Smith, Amy Hawkins, Scott Abbott, Devin Locke, Ladeana W. Hillier, Tracie Miner, Lucinda Fulton, Vincent Magrini, Todd Wylie, Jarret Glasscock, Joshua Conyers, Nathan Sander, Xiaoqi Shi, John R. Osborne, Patrick Minx, David Gordon, Asif Chinwalla, Yu Zhao, Rhonda E. Ries, Jacqueline E. Payton, Peter Westervelt, Michael H. Tomasson, Mark Watson, Jack Baty, Jennifer Ivanovich, Sharon Heath, William D. Shannon, Rakesh Nagarajan, Matthew J. Walter, Daniel C. Link, Timothy A. Graubert, John F. DiPersio, and Richard K. Wilson. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, 456(7218):66–72, 2008.

[28] E Mardis. Genome sequencing and cancer. *Current Opinion in Genetics and Development*, 22:245–250, 2012.

[29] International Cancer Genome Consortium. International network of cancer genome projects. *Nature*, 464(7291):993–998, 2010.

[30] John N Weinstein, Eric a Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad a Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–20, 2013.

[31] Peter J. Campbell, Gad Getz, Joshua M. Stuart, Jan O. Korbel, Lincoln D. Stein, and ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Net. Pan-cancer analysis of whole genomes. *bioRxiv*, 3(June):162784, 2017.

[32] John G. Tate, Sally Bamford, Harry C. Jubb, Zbyslaw Sondka, David M. Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G. Cole, Celestino Creatore, Elisabeth Dawson, Peter Fish, Bhavana Harsha, Charlie Hathaway, Steve C. Jupe, Chai Yin Kok, Kate Noble, Laura Ponting, Christopher C. Ramshaw, Claire E. Rye, Helen E. Speedy, Ray Stefancsik, Sam L. Thompson, Shicai Wang, Sari Ward, Peter J. Campbell, and Simon A. Forbes. COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, 47(D1):D941–D947, 2019.

[33] Vagheesh M. Narasimhan, Raheleh Rahbari, Aylwyn Scally, Arthur Wuster, Dan Mason, Yali Xue, John Wright, Richard C. Trembath, Eamonn R. Maher, David A. Van Heel, Adam Auton, Matthew E. Hurles, Chris Tyler-Smith, and Richard Durbin. Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. *Nature Communications*, 8(1):1–6, 2017.

[34] Rosa Ana Risques and Scott R. Kennedy. Aging and the rise of somatic cancer-associated mutations in normal tissues. *PLoS Genetics*, 14(1):1–12, 2018.

[35] Peter J Campbell and Iñigo Martincorena. Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489, 2015.

[36] David Malkin, Judy E Garber, Louise C Strong, Stephen H Friend, F. P. Li, J. F. Fraumeni, D. Malkin, S. H. Friend, D. I. Linzer, A. J. Levine, D. P. Lane, L. V. Crawford, M. Mowat, F. P. Li, S. J. Baker, A. Lavigueur, A. Shlien, D. A. Hill, M. Hisada, and A. Villani. CANCER. The cancer predisposition revolution. *Science (New York, N.Y.)*, 352(6289):1052–3, 2016.

[37] Nazneen Rahman. Realizing the promise of cancer predisposition genes. *Nature*, 505(7483):302–8, 2014.

[38] Shirley Hodgson. Mechanisms of inherited cancer susceptibility. *Journal of Zhejiang University SCIENCE B*, 9(1):1–4, 2008.

[39] Melissa S. Cline et al. BRCA Challenge: BRCA Exchange as a global resource for variants in BRCA1 and BRCA2. *PLoS Genetics*, 14(12):1–17, 2018.

[40] Kari Hemminki and Kamila Czene. Age specific and attributable risks of familial prostate carcinoma from the family-cancer database. *Cancer*, 95(6):1346–1353, 2002.

[41] Saud H. AlDubayan, Marios Giannakis, Nathanael D. Moore, G. Celine Han, Brendan Reardon, Tsuyoshi Hamada, Xinmeng Jasmine Mu, Reiko Nishihara, Zhirong Qian, Li Liu, Matthew B. Yurgelun, Sapna Syngal, Levi A. Garraway, Shuji Ogino, Charles S. Fuchs, and Eliezer M. Van Allen. Inherited DNA-Repair Defects in Colorectal Cancer. *The American Journal of Human Genetics*, 102(3):401–414, 2018.

[42] A G Knudson. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America*, 68(4):820–3, 1971.

[43] Douglas Hanahan and Robert A Weinberg. Hallmarks of Cancer : The Next Generation. *Cell*, 144(5):646–674, 2011.

[44] Ludmil B Alexandrov, Jaegil Kim, Nicholas J Haradhvala, Mi Ni Huang, Alvin WT Ng, Arnoud Boot, Kyle R Covington, Dmitry A Gordenin, Erik Bergstrom, Nuria Lopez-Bigas, Leszek J Klimczak, John R McPherson, Sandro Morganella, Radhakrishnan Sabarinathan, David A Wheeler, Ville Mustonen, Gad Getz, Steven G Rozen, Michael R Stratton, on behalf of the PCAWG Mutational Signatures Working Group Network, and the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes. The Repertoire of Mutational Signatures in Human Cancer. *bioRxiv*, 2018.

[45] Susanne N. Gröbner et al. The landscape of genomic alterations across childhood cancers. *Nature*, 555(7696):321–327, 2018.

[46] Malachi Griffith, Christopher A. Miller, Obi L. Griffith, Kilannin Krysiak, Zachary L. Skidmore, Avinash Ramu, Jason R. Walker, Ha X. Dang, Lee Trani, David E. Larson, Ryan T. Demeter, Michael C. Wendl, Joshua F. McMichael, Rachel E. Austin, Vincent Magrini, Sean D. McGrath, Amy Ly, Shashikant Kulkarni, Matthew G.

Cordes, Catrina C. Fronick, Robert S. Fulton, Christopher A. Maher, Li Ding, Jeffery M. Klco, Elaine R. Mardis, Timothy J. Ley, and Richard K. Wilson. Optimizing Cancer Genome Sequencing and Analysis. *Cell Systems*, 1(3):210–223, 2015.

[47] Natalia L. Komarova and Dominik Wodarz. Drug resistance in cancer: Principles of emergence and prevention. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9714–9719, 2005.

[48] 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*, 171(5):1029–1041.e21, 2017.

[49] Radhakrishnan Sabarinathan, Oriol Pich, Iñigo Martincorena, Carlota Rubio-Perez, Malene Juul, Jeremiah Wala, Steven Schumacher, Ofer Shapira, Nikos Sidiropoulos, Sebastian Waszak, David Tamborero, Loris Mularoni, Esther Rheinbay, Henrik Hornshoj, Jordi Deu-Pons, Ferran Muinos, Johanna Bertl, Qianyun Guo, Joachim Weischenfeldt, Jan O Korbel, Gad Getz, Peter J Campbell, Jakob S Pedersen, Rameen Beroukhim, Abel Perez-Gonzalez, Nuria Lopez-Bigas, PCAWG Drivers Group, Functional Interpretation, and the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Network. The whole-genome panorama of cancer drivers. *bioRxiv*, page 190330, 2017.

[50] H Lodish, A Berk, LS Zipursky, P Matsudaira, D Baltimore, and J Darnell. *Molecular Cell Biology*. W.H. Freeman, New York, 4 edition, 2000. Accessed through NCBI bookshelf.

[51] The ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[52] Sahar Pakneshan, Ali Salajegheh, Robert Anthony Smith, and Alfred King Yin Lam. Clinicopathological relevance of BRAF mutations in human cancer. *Pathology*, 45(4):346–356, 2013.

[53] The Cancer Genome Atlas Research Network. Integrated Genomic Characterization of Papillary Thyroid Carcinoma. *Cell*, 159(3):676–690, 2014.

[54] Sam Behjati, Gunes Gundem, David C. Wedge, Nicola D. Roberts, Patrick S. Tarpey, Susanna L. Cooke, Peter Van Loo, Ludmil B. Alexandrov, Manasa Ramakrishna, Helen Davies, Serena Nik-Zainal, Claire Hardy, Calli Latimer, Keiran M. Raine, Lucy Stebbings, Andy Menzies, David Jones, Rebecca Shepherd, Adam P. Butler, Jon W. Teague, Mette Jorgensen, Bhavisha Khatri, Nischalan Pillay, Adam Shlien, P. Andrew Futreal, Christophe Badie, Colin S. Cooper, Rosalind A. Eeles, Douglas Easton, Christopher Foster, David E. Neal, Daniel S. Brewer, Freddie Hamdy, Yong-Jie Lu, Andrew G. Lynch, Charlie E. Massi, Anthony Ng, Hayley C. Whitaker, Yongwei Yu, Hongwei Zhang, Elizabeth Bancroft, Dan Berney, Niedzica Camacho, Cathy Corbishley, Tokhir Dadaev, Nening Dennis, Tim Dudderidge, Sandra Edwards, Cyril Fisher, Jilur Ghori, Vincent J. Gnanapragasam, Christopher Greenman, Steve Hawkins, Steven Hazell, Will Howat, Katalin Karaszi, Jonathan

Kay, Zsofia Kote-Jarai, Barbara Kremeyer, Pardeep Kumar, Adam Lambert, Daniel Leongamornlert, Naomi Livni, Hayley Luxton, Lucy Matthews, Erik Mayer, Susan Merson, David Nicol, Christopher Ogden, Sarah O'Meara, Gill Pelvender, Nimish C. Shah, Simon Tavare, Sarah Thomas, Alan Thompson, Claire Verrill, Anne Warren, Jorge Zamora, Ultan McDermott, G. Steven Bova, Andrea L. Richardson, Adrienne M. Flanagan, Michael R. Stratton, Peter J. Campbell, K. J. Soderlind, P. J. Campbell, S. C. Baca, R. Abbott, C. Palmieri, S. Behjati, H. Li, R. Durbin, P. Van Loo, N. Bolli, and S. Nik-Zainal. Mutational signatures of ionizing radiation in second malignancies. *Nature Communications*, 7:12605, 2016.

[55] Avantika Lal, Daniele Ramazzotti, Ziming Weng, Keli Liu, James M. Ford, and Arend Sidow. Comprehensive genomic characterization of breast tumors with BRCA1 and BRCA2 mutations. *BMC Medical Genomics*, 12(1):1–13, 2019.

[56] Yosef E Maruvka, Kent W Mouw, Rosa Karlic, Prasanna Parasuraman, Atanas Kamburov, Paz Polak, Nicholas J Haradhvala, Julian M Hess, Esther Rheinbay, Yehuda Brody, Amnon Koren, Lior Z Braunstein, Alan D Andrea, Michael S Lawrence, Adam Bass, and Andre Bernards. Analysis of somatic microsatellite indels identifies driver events in human tumors. *Nature Biotechnology*, (August), 2017.

[57] Qingbo Wang, Emma Pierce-hoffman, Beryl B Cummings, Konrad J Karczewski, Jessica Alföldi, Laurent C Francioli, Laura D Gauthier, Andrew J Hill, and H Anne. Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. *bioRxiv*, 2019.

[58] Peter Priestley et al. Pan-cancer whole genome analyses of metastatic solid tumors. *bioRxiv*, 2018.

[59] Joanna Kaplanis, Nadia Akawi, Giuseppe Gallone, Jeremy F McRae, Elena Prigmore, Caroline F Wright, David R Fitzpatrick, Helen V Firth, Jeffrey C Barrett, Matthew E Hurles, and on behalf of the Deciphering Developmental Disorders Study. Exome-wide assessment of the functional impact and pathogenicity of multi-nucleotide mutations. *bioRxiv*, page 258723, 2018.

[60] 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 Langerod, Andrew Tutt, John W.M. Martens, Samuel A.J.R. Aparicio, Åke Borg, Anne Vincent Salomon, Gilles Thomas, Anne Lise Borresen-Dale, Andrea L. Richardson, Michael S. Neuberger, P. Andrew Futreal, Peter J. Campbell, and Michael R. Stratton. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993, 2012.

[61] Jaegil Kim, Kent W Mouw, Paz Polak, Lior Z Braunstein, Atanas Kamburov, Grace Tiao, David J Kwiatkowski, Jonathan E Rosenberg, Eliezer M Van Allen, Alan D D'Andrea, and Gad Getz. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nature Genetics*, 48(6):600–606, 2016.

[62] Jill E. Kucab, Xueqing Zou, Sandro Morganella, Madeleine Joel, A. Scott Nanda, Eszter Nagy, Celine Gomez, Andrea Degasperi, Rebecca Harris, Stephen P. Jackson, Volker M. Arlt, David H. Phillips, and Serena Nik-Zainal. A Compendium of Mutational Signatures of Environmental Agents. *Cell*, 177(4):821–836.e16, 2019.

[63] Joachim Weischenfeldt, Orsolya Symmons, François Spitz, and Jan O. Korbel. Phenotypic impact of genomic structural variation: Insights from and for human disease. *Nature Reviews Genetics*, 14(2):125–138, 2013.

[64] Andy W. Pang, Jeffrey R. MacDonald, Dalila Pinto, John Wei, Muhammad A. Rafiq, Donald F. Conrad, Hansoo Park, Matthew E. Hurles, Charles Lee, J. Craig Venter, Ewen F. Kirkness, Samuel Levy, Lars Feuk, and Stephen W. Scherer. Towards a comprehensive structural variation map of an individual human genome. *Genome Biology*, 11(5), 2010.

[65] 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*, 2017.

[66] Ian Michael Cartwright. Investigations Of Radiation-Induced And Spontaneous Chromosomal Inversion Formation And Characteristics, 2014.

[67] Franck Pellestor. Chromoanagenesis: Cataclysms behind complex chromosomal rearrangements. *Molecular Cytogenetics*, 12(1):1–12, 2019.

[68] Fouad Yousif, Stephenie Prokopec, Ren X Sun, Fan Fan, Christopher M Lalansingh, David H Park, Lesia Szyca, PCAWG Network, and Paul C Boutros. The Origins and Consequences of Localized and Global Somatic Hypermutation. *bioRxiv*, 2018.

[69] 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(3):666–677, 2013.

[70] Peter J. Campbell, Shinichi Yachida, Laura J. Mudie, Philip J. Stephens, Erin D. Pleasance, Lucy A. Stebbings, Laura A. Morsberger, Calli Latimer, Stuart McLaren, Meng Lay Lin, David J. McBride, Ignacio Varela, Serena A. Nik-Zainal, Catherine

Leroy, Mingming Jia, Andrew Menzies, Adam P. Butler, Jon W. Teague, Constance A. Griffin, John Burton, Harold Swerdlow, Michael A. Quail, Michael R. Stratton, Christine Iacobuzio-Donahue, and P. Andrew Futreal. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature*, 467(7319):1109–1113, 2010.

[71] B McClintock. The Stability of Broken Ends of Chromosomes in Zea Mays. *Genetics*, 26(2):234–82, 1941.

[72] Tom Walsh, Jon M McClellan, Shane E McCarthy, Anjené M Addington, Sarah B Pierce, Greg M Cooper, Alex S Nord, Mary Kusenda, Dheeraj Malhotra, Abhishek Bhandari, Sunday M Stray, Caitlin F Rippey, Patricia Roccanova, Vlad Makarov, B Lakshmi, Robert L Findling, Linmarie Sikich, Thomas Stromberg, Barry Merriman, Nitin Gogtay, Philip Butler, Kristen Eckstrand, Laila Noory, Peter Gochman, Robert Long, Zugen Chen, Sean Davis, Carl Baker, Evan E Eichler, Paul S Meltzer, Stanley F Nelson, Andrew B Singleton, Ming K Lee, Judith L Rapoport, Mary-Claire King, and Jonathan Sebat. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*, 320(5875):539–43, 2008.

[73] Maria Karayiorgou, Tony J Simon, and Joseph a Gogos. 22q11.2 microdeletions: linking DNA structural variation to brain dysfunction and schizophrenia. *Nature reviews Neuroscience*, 11(6):402–416, 2010.

[74] Ira M. Hall and Aaron R. Quinlan. Detection and interpretation of genomic structural variation in mammals. *Methods in Molecular Biology*, 838:225–248, 2012.

[75] Joshua M. Shulman. Structural Variation and the Expanding Genomic Architecture of Parkinson Disease. *JAMA Neurology*, 70(11):1355, 2013.

[76] Paweł Stankiewicz and James R. Lupski. Structural Variation in the Human Genome and its Role in Disease. *Annual Review of Medicine*, 61(1):437–455, 2010.

[77] Pierotti MA, Sozzi G, and Croce CM. Mechanisms of oncogene activation. In Kufe DW, Pollock RE, and Weichselbaum RR, editors, *Holland-Frei Cancer Medicine 6th Edition*. BC Decker, 2003.

[78] Douglas B. Johnson, Merrida A. Childress, Zachary R. Chalmers, Garrett M. Frampton, Siraj M. Ali, Samuel M. Rubinstein, David Fabrizio, Jeffrey S. Ross, Sohail Balasubramanian, Vincent A. Miller, Philip J. Stephens, Jeffrey A. Sosman, and Christine M. Lovly. BRAF internal deletions and resistance to BRAF/MEK inhibitor therapy. *Pigment Cell and Melanoma Research*, 31(3):432–436, 2018.

[79] George T. Rudkin and B. D. Stollar. High resolution detection of DNA-RNA hybrids in situ by indirect immunofluorescence [29]. *Nature*, 265(5593):472–473, 1977.

[80] M Prakash Hande, Tamara V Azizova, Charles R Geard, Ludmilla E Burak, Catherine R Mitchell, Valentin F Khokhryakov, Evgeny K Vasilenko, and David J Brenner. Past exposure to densely ionizing radiation leaves a unique permanent signature in the genome. *The American Journal of Human Genetics*, 72(5):1162–1170, 2003.

[81] Michael N. Cornforth and Bradford D. Loucas. A Cytogenetic Profile of Radiation Damage. *Radiation Research*, 191(1):1, 2018.

[82] Ken Chen, John W Wallis, Michael D McLellan, David E Larson, Joelle M Kalicki, Craig S Pohl, Sean D McGrath, Michael C Wendl, Qunyuan Zhang, Devin P Locke, Xiaoqi Shi, Robert S Fulton, Timothy J Ley, Richard K Wilson, Li Ding, and Elaine R Mardis. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods*, 6(9):677–81, sep 2009.

[83] 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. *Nature Methods*, 8(8):652–654, 2011.

[84] 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, 2009.

[85] 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):333–339, 2012.

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

[87] Xiaotong Yao, Cheng-Zhong Zhang, Noah F. Greenwald, Gad Getz, Joachim Weischenfeldt, Yilong Li, Chip Stewart, Chad Nusbaum, Ted Sharpe, Pratiti Bandopadhayay, Ryan O'Rourke, Steve Schumacher, Marcin Imielinski, Matthew Meyerson, Peter Campbell, Rameen Beroukhim, and Jeremiah A. Wala. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Research*, 28(4):581–591, 2018.

[88] Fritz J. Sedlazeck, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt Von Haeseler, and Michael C. Schatz. Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, 15(6):461–468, 2018.

[89] Aaron M Wenger, Paul Peluso, William J Rowell, Pi-Chuan Chang, Richard J Hall, Gregory T Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D Olson, Armin Toepfer, Chen-Shan Chin, Michael Alonge, Medhat Mahmoud, Yufeng Qian, Adam M Phillippy, Michael C Schatz, Gene Myers, Mark A DePristo, Jue Ruan, Tobias Marschall, Fritz J Sedlazeck, Justin M Zook, Heng Li, Sergey Koren, Andrew Carroll, David R Rank, and Michael W Hunkapiller. Highly-accurate long-read sequencing improves variant detection and assembly of a human genome. *bioRxiv*, page 519025, 2019.

[90] Alex Bishara, Yuling Liu, Ziming Weng, Dorna Kashef-haghighi, and Daniel E Newburger. Read Clouds Uncover Variation in Complex Regions of the Human Genome. *Genome Research*, pages 1–35, 2015.

[91] Mark J.P. Chaisson, Ashley D. Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, 10(1):1–16, 2019.

[92] Geraldine A. Van der Auwera, Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, Eric Banks, Kiran V. Garimella, David Altshuler, Stacey Gabriel, and Mark A. DePristo. *From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline*. 2013.

[93] Samantha Zarate, Andrew Carroll, Olga Krashenina, Fritz J Sedlazeck, Goo Jun, William Salerno, Eric Boerwinkle, and Richard Gibbs. Parliament2: Fast Structural Variant Calling Using Optimized Combinations of Callers. *bioRxiv*, page 424267, 2018.

[94] Lincoln D. Stein, Bartha M. Knoppers, Peter Campbell, Gad Getz, and Jan O. Korbel. Data analysis: Create a cloud commons. *Nature*, 523(7559):149–151, 2015.

[95] Brian D. Ondov, Todd J. Treangen, Pall Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. Mash: fast genome and metagenome distance estimation using minhash. *Genome Biology*, 17:1–14, 2016.

[96] Eric T Dawson, Sarah Wagner, David Roberson, Meredith Yeager, Joseph Boland, Erik Garrison, Stephen Chanock, Mark Schiffman, Tina Raine-bennett, Thomas Lorey, Phillip E Castle, Lisa Mirabello, and Richard Durbin. Viral coinfection analysis using a MinHash toolkit. *BMC Bioinformatics*, pages 1–10, 2019.

[97] Simon Gog, Timo Beller, Alistair Moffat, and Matthias Petri. From theory to practice: Plug and play with succinct data structures. In *13th International Symposium on Experimental Algorithms, (SEA 2014)*, pages 326–337, 2014.

[98] Jouni Sirén, Niko Valimaki, and Veli Makinen. Indexing Graphs for Path Queries with Applications in Genome Research. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(2):375–388, 2014.

[99] 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. *Nature Biotechnology*, 36(9):875–881, 2018.

[100] Markus Hsi Yang Fritz, Rasko Leinonen, Guy Cochrane, and Ewan Birney. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Research*, 21(5):734–740, 2011.

[101] Ravi Vijaya Satya, Nela Zavaljevski, and Jaques Reifman. A new strategy to reduce allelic bias in RNA-Seq readmapping. *Nucleic Acids Research*, 40(16):1–9, 2012.

[102] Jacob Pritt, Nae-Chyun Chen, and Ben Langmead. FORGe: prioritizing variants for graph genomes. *Genome Biology*, 19(1):1–16, 2018.

[103] Jan Schröder, Santhosh Girirajan, Anthony T. Papenfuss, and Paul Medvedev. Improving the Power of Structural Variation Detection by Augmenting the Reference. *Plos One*, 10(8):e0136771, 2015.

[104] Daehwan Kim, Ben Langmead, and Steven L. Salzberg. HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, 12(4):357–360, 2015.

[105] HiSat2 Authors. Hisat-genotype tutorial. http://ccb.jhu.edu/hisat-genotype/index.php/Main:Tutorial, 2018.

[106] J. Kim, Sangwoo Kim, Hojung Nam, Sangwoo Kim, and Doheon Lee. SoloDel: A probabilistic model for detecting low-frequent somatic deletions from unmatched sequencing data. *Bioinformatics*, (June):1–9, 2015.

[107] 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. *Nature Biotechnology*, 36:875, 8 2018.

[108] Goran Rakocevic, Vladimir Semenyuk, Wan-Ping Lee, James Spencer, John Browning, Ivan J Johnson, Vladan Arsenijevic, Jelena Nadj, Kaushik Ghose, Maria C Suciu, Sun-Gou Ji, Gülfem Demir, Lizao Li, Berke Ç Toptaş, Alexey Dolgoborodov, Björn Pollex, Iosif Spulber, Irina Glotova, Péter Kómár, Andrew L Stachyra, Yilong Li, Milos Popovic, Morten Källberg, Amit Jain, and Deniz Kural. Fast and accurate genomic analyses using genome graphs. *Nature Genetics*, 51(2):354–362, 2019.

[109] Hannes P. Eggertsson, Hakon Jonsson, Snaedis Kristmundsdottir, Eirikur Hjartarson, Birte Kehr, Gisli Masson, Florian Zink, Kristjan E. Hjorleifsson, Aslaug Jonasdottir, Adalbjorg Jonasdottir, Ingileif Jonsdottir, Daniel F. Gudbjartsson, Pall Melsted, Kari Stefansson, and Bjarni V. Halldorsson. Graphtyper enables population-scale genotyping using pangenome graphs. *Nature Genetics*, 49(11):1654–1660, 2017.

[110] Derrick E. Wood and Steven L. Salzberg. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3), 2014.

[111] Dirk D. Dolle, Zhicheng Liu, Matthew Cotten, Jared T. Simpson, Zamin Iqbal, Richard Durbin, Shane A. McCarthy, and Thomas M. Keane. Using reference-free compressed data structures to analyze sequencing reads from thousands of human genomes. *Genome Research*, 27(2):300–309, 2017.

[112] Phelim Bradley, Henk den Bakker, Eduardo Rocha, Gil McVean, and Zamin Iqbal. Real-time search of all bacterial and viral genomic data. *bioRxiv*, 37(February):234955, 2017.

[113] Adam M Novak, Erik Garrison, Heng Li, Glenn Hickey, Deniz Kural, Richard Durbin, Benedict Paten, David Haussler, and Gilean McVean. Genome Graphs. *bioRxiv*, pages 1–26, 2017.

[114] Daniel R. Zerbino and Ewan Birney. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–829, 2008.

[115] Eugene W. Myers. The fragment assembly string graph. *Bioinformatics*, 21(SUPPL. 2):79–85, 2005.

[116] Bahlul Haider, Tae-Hyuk Ahn, Brian Bushnell, Juanjuan Chai, Alex Copeland, and Chongle Pan. Omega: an Overlap-graph de novo Assembler for Metagenomics. *Bioinformatics (Oxford, England)*, 30(19):1–6, 2014.

[117] Jared T. Simpson and Richard Durbin. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics*, 26(12):367–373, 2010.

[118] Giuseppe Narzisi, André Corvelo, Kanika Arora, Ewa A Bergmann, Minita Shah, Rajeeva Musunuri, Anne-katrin Emde, Nicolas Robine, Vladimir Vacic, and Michael C Zody. Genome-wide somatic variant calling using localized colored de Bruijn graphs. *Communications Biology*, (2018):1–9, 2018.

[119] Benedict Paten, Dent Earl, Ngan Nguyen, Mark Diekhans, Daniel Zerbino, and David Haussler. Cactus: Algorithms for genome multiple sequence alignment. *Genome Research*, 21(9):1512–1528, 2011.

[120] Isaac Turner, Kiran V. Garimella, Zamin Iqbal, and Gil McVean. Integrating long-range connectivity information into de Bruijn graphs. *Bioinformatics*, 34(15):2556–2565, 2018.

[121] A. Dilthey, C. J. Cox, Z. Iqbal, M. R. Nelson, and G. McVean. Improved genome inference in the MHC using a population reference graph. *bioRxiv*, 47(6):006973, 2014.

[122] Erik Garrison. *Graphical pangenomics*. PhD thesis, University of Cambridge, 2018.

[123] Lia Chappell, Andrew J C Russell, and Thierry Voet. Single-Cell (Multi) omics Technologies. 2018.

[124] Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev, and Steven A. McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[125] Sergey Koren, Arang Rhie, Brian P. Walenz, Alexander T. Dilthey, Derek M. Bickhart, Sarah B. Kingan, Stefan Hiendleder, John L. Williams, Timothy P.L. Smith, and Adam Phillippy. Complete assembly of parental haplotypes with trio binning. *bioRxiv*, page 271486, 2018.

[126] Milan Malinsky, Jared T Simpson, and Richard Durbin. Trio-Sga : facilitating de novo assembly of highly heterozygous genomes with parent-child trios. *bioRxiv*, pages 1–6, 2016.

[127] Dengfeng Guan, Shane A Mccarthy, Jonathan Wood, Kerstin Howe, Yadong Wang, and Richard Durbin. Identifying and removing haplotypic duplication in primary genome assemblies. *bioRxiv*, 2019.

[128] Michael J Roach, Simon A Schmidt, and Anthony R Borneman. Purge Haplotigs: Synteny Reduction for Third-gen Diploid Genome Assemblies. *bioRxiv*, page 286252, 2018.

[129] Jun Yu, Qiang Feng, Sunny Hei Wong, Dongya Zhang, Qiao Yi Liang, Youwen Qin, Longqing Tang, Hui Zhao, Jan Stenvang, Yanli Li, Xiaokai Wang, Xiaoqiang Xu, Ning Chen, William Ka Kei Wu, Jumana Al-Aama, Hans Jorgen Nielsen, Pia Kiilerich, Benjamin Anderschou Holbech Jensen, Tung On Yau, Zhou Lan, Huijue Jia, Junhua Li, Liang Xiao, Thomas Yuen Tung Lam, Siew Chien Ng, Alfred Sze Lok Cheng, Vincent Wai Sun Wong, Francis Ka Leung Chan, Xun Xu, Huanming Yang, Lise Madsen, Christian Datz, Herbert Tilg, Jian Wang, Nils Brünner, Karsten Kristiansen, Manimozhiyan Arumugam, Joseph Jao Yiu Sung, and Jun Wang. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut*, 66(1):70–78, 2017.

[130] Steven J. Biller, Paul M. Berube, Keven Dooley, Madeline Williams, Brandon M. Satinsky, Thomas Hackl, Shane L. Hogle, Allison Coe, Kristin Bergauer, Heather A. Bouman, Thomas J. Browning, Daniele De Corte, Christel Hassler, Debbie Hulston, Jeremy E. Jacquot, Elizabeth W. Maas, Thomas Reinthaler, Eva Sintes, Taichi Yokokawa, and Sallie W. Chisholm. Data descriptor: Marine microbial metagenomes sampled across space and time. *Scientific Data*, 5:1–7, 2018.

[131] Nicholas A. Be, Aram Avila-Herrera, Jonathan E. Allen, Nitin Singh, Aleksandra Checinska Sielaff, Crystal Jaing, and Kasthuri Venkateswaran. Whole metagenome profiles of particulates collected from the International Space Station. *Microbiome*, 5(1):81, 2017.

[132] Xi Liu, Shu Yang, Yangqing Wang, He Ping Zhao, and Liyan Song. Metagenomic analysis of antibiotic resistance genes (ARGs) during refuse decomposition. *Science of the Total Environment*, 634(266):1231–1237, 2018.

[133] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–10, 1990.

[134] Chirag Jain, Alexander Dilthey, Sergey Koren, Srinivas Aluru, and Adam M Phillippy. A fast approximate algorithm for mapping long reads to large reference databases. In Cham, editor, *International Conference on Research in Computational Molecular Biology*. Springer, 2017.

[135] Global Buden of Disease Cancer Collaboration. Europe PMC Funders Group The Global Burden of Cancer 2013. *JAMA Oncol.*, 1(January 2014):505–527, 2015.

[136] Mark Schiffman, John Doorbar, Nicolas Wentzensen, Silvia de Sanjosé, Carole Fakhry, Bradley J. Monk, Margaret A. Stanley, and Silvia Franceschi. Carcinogenic human papillomavirus infection. *Nature Reviews Disease Primers*, 2:16086, 2016.

[137] Peng Guan, Rebecca Howell-Jones, Ni Li, Laia Bruni, Silvia De Sanjosé, Silvia Franceschi, and Gary M. Clifford. Human papillomavirus types in 115,789 HPV-positive women: A meta-analysis from cervical infection to cancer. *International Journal of Cancer*, 131(10):2349–2359, 2012.

[138] Mark Schiffman, Rolando Herrero, Rob Desalle, Allan Hildesheim, Sholom Wacholder, Ana Cecilia Rodriguez, Maria C. Bratti, Mark E. Sherman, Jorge Morales, Diego Guillen, Mario Alfaro, Martha Hutchinson, Thomas C. Wright, Diane Solomon, Zigui Chen, John Schussler, Philip E. Castle, and Robert D. Burk. The carcinogenicity of human papillomavirus types reflects viral evolution. *Virology*, 337(1):76–84, 2005.

[139] Salvatore Vaccarella, Anna Söderlund-Strand, Silvia Franceschi, Martyn Plummer, and Joakim Dillner. Patterns of Human Papillomavirus Types in Multiple Infections: An Analysis in Women and Men of the High Throughput Human Papillomavirus Monitoring Study. *PLoS ONE*, 8(8):e71617, 2013.

[140] Mark Schiffman, Philip E. Castle, Jose Jeronimo, Ana C. Rodriguez, and Sholom Wacholder. Human papillomavirus and cervical cancer. *Lancet*, 370(9590):890–907, 2007.

[141] Anil K. Chaturvedi, Hormuzd A. Katki, Allan Hildesheim, Ana Cecilia Rodríguez, Wim Quint, Mark Schiffman, Leen Jan Van Doorn, Carolina Porras, Sholom Wacholder, Paula Gonzalez, Mark E. Sherman, and Rolando Herrero. Human papillomavirus infection with multiple types: Pattern of coinfection and risk of cervical disease. *Journal of Infectious Diseases*, 203(7):910–920, 2011.

[142] Marcos P. Freire, Daniel Pires, Raphael Forjaz, Sergio Sato, Ismael Cotrim, Monica Stiepcich, Bruno Scarpellini, and Jose C. Truzzi. Genital prevalence of HPV types and co-infection in men. *International Braz J Urol*, 40(1):67–71, 2014.

[143] Robert D. Burk, Ariana Harari, and Zigui Chen. Human papillomavirus genome variants. *Virology*, 445(1-2):232–243, 2013.

[144] Robert D. Burk, Ariana Harari, and Zigui Chen. Human papillomavirus genome variants. *Virology*, 445(1-2):232–243, 2013.

[145] Donavan T. Cheng, Talia N. Mitchell, Ahmet Zehir, Ronak H. Shah, Ryma Benayed, Aijazuddin Syed, Raghu Chandramohan, Zhen Yu Liu, Helen H. Won, Sasinya N. Scott, A. Rose Brannon, Catherine O'Reilly, Justyna Sadowska, Jacklyn Casanova, Angela Yannes, Jaclyn F. Hechtman, Jinjuan Yao, Wei Song, Dara S. Ross, Alifya Oultache, Snjezana Dogan, Laetitia Borsu, Meera Hameed, Khedoudja Nafa, Maria E. Arcila, Marc Ladanyi, and Michael F. Berger. Memorial sloan kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT): A hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *Journal of Molecular Diagnostics*, 17(3):251–264, 2015.

[146] J. Grinfeld, J. Nangalia, E. J. Baxter, D. C. Wedge, N. Angelopoulos, R. Cantrill, A. L. Godfrey, E. Papaemmanuil, G. Gundem, C. MacLean, J. Cook, L. O'Neil, S. O'Meara, J. W. Teague, A. P. Butler, C. E. Massie, N. Williams, F. L. Nice, C. L. Andersen, H. C. Hasselbalch, P. Guglielmelli, M. F. McMullin, A. M. Vannucchi,

C. N. Harrison, M. Gerstung, A. R. Green, and P. J. Campbell. Classification and personalized prognosis in myeloproliferative neoplasms. *New England Journal of Medicine*, 379(15):1416–1430, 2018.

[147] Stefan C. Dentro, Ignaty Leshchiner, Kerstin Haase, Maxime Tarabichi, Jeff Wintersinger, Amit G. Deshwar, Kaixian Yu, Yulia Rubanova, Geoff Macintyre, Ignacio Vazquez-Garcia, Kortine Kleinheinz, Dimitri G. Livitz, Salem Malikic, Nilgun Donmez, Subhajit Sengupta, Jonas Demeulemeester, Pavana Anur, Clemency Jolly, Marek Cmero, Daniel Rosebrock, Steven Schumacher, Yu Fan, Matthew Fittall, Ruben M. Drews, Xiaotong Yao, Juhee Lee, Matthias Schlesner, Hongtu Zhu, David J. Adams, Gad Getz, Paul C. Boutros, Marcin Imielinski, Rameen Beroukhim, S. Cenk Sahinalp, Yuan Ji, Martin Peifer, Iñigo Martincorena, Florian Markowetz, Ville Mustonen, Ke Yuan, Moritz Gerstung, Paul T. Spellman, Wenyi Wang, Quaid D. Morris, David C. Wedge, Peter Van Loo, PCAWG Evolution Group, Heterogeneity Working, and PCAWG Network. Portraits of genetic intratumour heterogeneity and subclonal selection across cancer types. *bioRxiv*, page 312041, 2018.

[148] Henry Lee-Six, Nina Friesgaard Øbro, Mairi S. Shepherd, Sebastian Grossmann, Kevin Dawson, Miriam Belmonte, Robert J. Osborne, Brian J. P. Huntly, Iñigo Martincorena, Elizabeth Anderson, Laura O'Neill, Michael R. Stratton, Elisa Laurenti, Anthony R. Green, David G. Kent, and Peter J. Campbell. Population dynamics of normal human blood inferred from somatic mutations. *Nature*, page 1, 2018.

[149] Iñigo Martincorena. Somatic mutation and clonal expansions in human tissues. *Genome Medicine*, 11(1):11–13, 2019.

[150] Florian Zink, Simon N. Stacey, Gudmundur L. Norddahl, Michael L. Frigge, Olafur T. Magnusson, Ingileif Jonsdottir, Thorgeir E. Thorgeirsson, Asgeir Sigurdsson, Sigurjon A. Gudjonsson, Julius Gudmundsson, Jon G. Jonasson, Laufey Tryggvadottir, Thorvaldur Jonsson, Agnar Helgason, Arnaldur Gylfason, Patrick Sulem, Thorunn Rafnar, Unnur Thorsteinsdottir, Daniel F. Gudbjartsson, Gisli Masson, Augustine Kong, and Kari Stefansson. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood*, 130(6):742–752, 2017.

[151] Siddhartha Jaiswal, Pierre Fontanillas, Jason Flannick, Alisa Manning, Peter V. Grauman, Brenton G. Mar, R. Coleman Lindsley, Craig H. Mermel, Noel Burtt, Alejandro Chavez, John M. Higgins, Vladislav Moltchanov, Frank C. Kuo, Michael J. Kluk, Brian Henderson, Leena Kinnunen, Heikki A. Koistinen, Claes Ladenvall, Gad Getz, Adolfo Correa, Benjamin F. Banahan, Stacey Gabriel, Sekar Kathiresan, Heather M. Stringham, Mark I. McCarthy, Michael Boehnke, Jaakko Tuomilehto, Christopher Haiman, Leif Groop, Gil Atzmon, James G. Wilson, Donna Neuberg, David Altshuler, and Benjamin L. Ebert. Age-related clonal hematopoiesis associated with adverse outcomes. *New England Journal of Medicine*, 371(26):2488–2498, 2014.

[152] V. H. Teixeira, C. P. Pipinikas, A. Pennycuick, H. Lee-Six, D. Chandrasekharan, J. Beane, T. J. Morris, A. Karpathakis, A. Feber, C. E. Breeze, P. Ntolios, R. E. Hynds, M. Falzon, A. Capitanio, B. Carroll, P. F. Durrenberger, G. Hardavella,

J. M. Brown, A. G. Lynch, H. Farmery, D. S. Paul, R. C. Chambers, N. McGranahan, N. Navani, R. Thakrar, C. Swanton, S. Beck, P. J. George, A. Spira, P. J. Campbell, C. Thirlwell, and S. M. Janes. Deciphering the genomic, epigenomic and transcriptomic landscapes of pre-invasive lung cancer lesions. *Nature Medicine*, Accepted, 2018.

[153] Peter C. Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

[154] P. Vineis, L. B. Alexandrov, M. R. Stratton, P. J. Campbell, S. Nik-Zainal, A. Fujimoto, D. H. Phillips, H. Nakagawa, P. Van Loo, T. Shibata, K. Haase, I. Martincorena, Y. S. Ju, and Y. Totoki. Mutational signatures associated with tobacco smoking in human cancer. *Science*, 354(6312):618–622, 2016.

[155] Cristian Tomasetti and Bert Vogelstein. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, 347(6217):78–81, 2015.

[156] Per Strand. Radioactive fallout in Norway from the Chernobyl accident Radioactive fallout in Norway from the Chernobyl accident, 1994.

[157] N. A. Beresford, S. Fesenko, A. Konoplev, L. Skuterud, J. T. Smith, and G. Voigt. Thirty years after the Chernobyl accident: What lessons have we learnt? *Journal of Environmental Radioactivity*, 157:77–89, 2016.

[158] United Nations Scientific Committee on the Effects of Atomic Radiation. Sources and effects of ionizing radiation, 2008.

[159] Vasili Kazakov, Evgeni Demidchik, and Larisa Astakhova. Thyroid cancer after Chernobyl. *Nature*, 359(September):21, 1992.

[160] Peter Jacob, Tatiana I. Bogdanova, Elena E. Buglova, Jacov E. Kenigsberg, and Nikolay D. Tronko. Comparison of thyroid cancer incidence after the Chernobyl accident in Belarus and in Ukraine. *International Congress Series*, 1234(C):215–219, 2002.

[161] Elaine Ron and Arthur B. Schneider. Thyroid Cancer. In *Cancer Epidemiology and Prevention*, pages 975–994. 2006.

[162] George Contis and Thomas P Foley. Depression , Suicide Ideation , and Thyroid Tumors Among Ukrainian Adolescents Exposed as Children to Chernobyl Radiation. 7(5):332–338, 2015.

[163] Steven L. Simon and André Bouville. Health effects of nuclear weapons testing. *The Lancet*, 386(9992):407–409, 2015.

[164] Leonard Wartofsky. Increasing world incidence of thyroid cancer: Increased detection or higher radiation exposure? *Hormones*, 9(2):103–108, 2010.

[165] International Agency for Research on Cancer. *World Cancer Report 2014*. 2014.

[166] Maria E Cabanillas, David G Mcfadden, and Cosimo Durante. Thyroid cancer. *The Lancet*, 388(10061):2783–2795, 2016.

[167] Elaine Ron, Jay H Lubin, Roy E Shore, Kiyohiko Mabuchi, Baruch Modan, Linda M Pottern, Arthur B Schneider, Margaret A Tucker, John D Boice, No Mar, Elaine Ron, Jay H Lubin, Roy E Shore, Kiyohiko Mabuchi, Baruch Modan, and Linda M Pottern. Thyroid Cancer after Exposure to External Radiation : A Pooled Analysis of Seven Studies. *Radiation Research*, 141(3):259–277, 1995.

[168] Michael B. Zimmermann and Valeria Galetti. Iodine intake as a risk factor for thyroid cancer: a comprehensive review of animal and human studies. *Thyroid Research*, 8(1):8, 2015.

[169] International Agency for Research on Cancer. A review of human carcinogens (radiation), 2012.

[170] Anuja Mehta and James E. Haber. Sources of DNA double-strand breaks and models of recombinational DNA repair. *Cold Spring Harbor Perspectives in Biology*, 6(9), 2014.

[171] Susan L. Tucker and Howard D. Thames Jr. Flexure dose: The low-dose limit of effective fractionation. *International Journal of Radiation Oncology Biology*, 9:1373–1983.

[172] Eric Grégoire, Laurence Roy, Valérie Buard, Martine Delbos, Valérie Durand, Cécile Martin-Bodiot, Pascale Voisin, Irène Sorokine-Durm, Aurélie Vaurijoux, Philippe Voisin, Céline Baldeyron, and Joan Francesc Barquinero. Twenty years of FISH-based translocation analysis for retrospective ionizing radiation biodosimetry. *International Journal of Radiation Biology*, 94(3):248–258, 2018.

[173] 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. *Nature Biotechnology*, 30(5):413–421, 2012.

[174] Fan Zhang, Matthew Flickinger, Goncalo Abecasis, Michael Boehnke, and Hyun Min Kang. Ancestry-agnostic estimation of DNA sample contamination from sequence reads. *bioRxiv Bioinformatics*, pages 1–42, 2018.

[175] Mark A. Depristo, Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, Guillermo Del Angel, Manuel A. Rivas, Matt Hanna, Aaron McKenna, Tim J. Fennell, Andrew M. Kernytsky, Andrey Y. Sivachenko, Kristian Cibulskis, Stacey B. Gabriel, David Altshuler, and Mark J. Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–501, 2011.

[176] Kristian Cibulskis, Aaron McKenna, Tim Fennell, Eric Banks, Mark DePristo, and Gad Getz. ContEst: Estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics*, 27(18):2601–2602, 2011.

[177] Paul P.S. Wang, Wendy T. Parker, Susan Branford, and Andreas W. Schreiber. BAM-matcher: A tool for rapid NGS sample matching. *Bioinformatics*, 32(17):2699–2701, 2016.

[178] Cronin KA (eds). Howlader N, Noone AM, Krapcho M, Miller D, Bishop K, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ. National Cancer Institute SEER Cancer Statistics Review 1975-2013, 2016.

[179] Kate Voss, Jeff Gentry, and Geraldine Van Der Auwera. Full-stack genomics pipelining with GATK4+ WDL+ Cromwell [version 1; not peer reviewed]. *F1000Research*, page 4, 2017.

[180] Matteo Chiara, Silvia Gioiosa, Giovanni Chillemi, Mattia D'Antonio, Tiziano Flati, Ernesto Picardi, Federico Zambelli, David Stephen Horner, Graziano Pesole, and Tiziana Castrignanò. CoVaCS: A consensus variant calling system. *BMC Genomics*, 19(1):1–9, 2018.

[181] 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. *Nature Biotechnology*, 31(3):213–219, 2013.

[182] 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, 2012.

[183] Sangtae Kim, Konrad Scheffler, Aaron Halpern, Mitchell Bekritsky, Eunho Noh, Xiaoyu Chen, Doruk Beyter, Peter Krusche, and Christopher Saunders. Strelka2: Fast and accurate variant calling for clinical sequencing applications. *bioRxiv*, 2017.

[184] 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.

[185] Alex H. Ramos, Lee Lichtenstein, Manaswi Gupta, Michael S. Lawrence, Trevor J. Pugh, Gordon Saksena, Matthew Meyerson, and Gad Getz. Oncotator: Cancer variant annotation tool. *Human Mutation*, 36(4):E2423–E2429, 2015.

[186] Maura Costello, Trevor J. Pugh, Timothy J. Fennell, Chip Stewart, Lee Lichtenstein, James C. Meldrim, Jennifer L. Fostel, Dennis C. Friedrich, Danielle Perrin, Danielle Dionne, Sharon Kim, Stacey B. Gabriel, Eric S. Lander, Sheila Fisher, and Gad Getz. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Research*, 41(6):1–12, 2013.

[187] Cyriac Kandoth, Michael D McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, Qunyuan Zhang, Joshua F McMichael, Matthew a Wyczalkowski, Mark D M Leiserson, Christopher a Miller, John S Welch, Matthew J Walter, Michael C Wendl, Timothy J Ley, Richard K Wilson, Benjamin J Raphael, and Li Ding. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–9, oct 2013.

[188] Michael S Lawrence et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–8, jul 2013.

[189] Dang Vu-Phan and Ronald J. Koenig. Genetics and epigenetics of sporadic thyroid cancer. *Molecular and Cellular Endocrinology*, 386(1-2):55–66, 2014.

[190] The role of thyroglobulin in thyroid hormonogenesis. *Nature Reviews Endocrinology*, 15:323–338, 2019.

[191] Marcin Imielinski, Guangwu Guo, and Matthew Meyerson. Insertions and deletions target lineage-defining genes in human cancers. *Cell*, 3(168):460–472, 2017.

[192] Kirti Mittal, Muhammad A. Rafiq, Rafiullah Rafiullah, Ricardo Harripaul, Hazrat Ali, Muhammad Ayaz, Muhammad Aslam, Farooq Naeem, Muhammad Amin-ud din, Ahmed Waqas, Joyce So, Gudrun A. Rappold, John B. Vincent, and Muhammad Ayub. Mutations in the genes for thyroglobulin and thyroid peroxidase cause thyroid dyshormonogenesis and autosomal-recessive intellectual disability. *Journal of Human Genetics*, 61(10):867–872, 2016.

[193] Héctor M. Targovnik, Cintia E. Citterio, and Carina M. Rivolta. Thyroglobulin gene mutations in congenital hypothyroidism. *Hormone Research in Paediatrics*, 75(5):311–321, 2011.

[194] Ileana G.S. Rubio and Geraldo Medeiros-Neto. Mutations of the thyroglobulin gene and its relevance to thyroid disorders. *Current Opinion in Endocrinology, Diabetes and Obesity*, 16(5):373–378, 2009.

[195] Bruno Di Jeso and Peter Arvan. Thyroglobulin from molecular and cellular biology to clinical endocrinology. *Endocrine Reviews*, 37(1):2–36, 2016.

[196] Sann Y. Mon, Gregory Riedlinger, Collette E. Abbott, Raja Seethala, N. Paul Ohori, Marina N. Nikiforova, Yuri E. Nikiforov, and Steven P. Hodak. Cancer risk and clinicopathological characteristics of thyroid nodules harboring thyroid-stimulating hormone receptor gene mutations. *Diagnostic Cytopathology*, 46(5):369–377, 2018.

[197] Jake C. Robertson, Cheryl L. Jorcyk, and Julia Thom Oxford. DICER1 syndrome: DICER1 mutations in rare cancers. *Cancers*, 10(5):1–17, 2018.

[198] Kris Ann P. Schultz, Gretchen M. Williams, Junne Kamihara, Douglas R. Stewart, Anne K. Harris, Andrew J. Bauer, Joyce Turner, Rachana Shah, Katherine Schneider, Kami Wolfe Schneider, Ann Garrity Carr, Laura A. Harney, Shari Baldinger, A. Lindsay Frazier, Daniel Orbach, Dominik T. Schneider, David Malkin, Louis P. Dehner, Yoav H. Messinger, and D. Ashley Hill. Dicer1 and associated conditions: Identification of at-risk individuals and recommended surveillance strategies. *Clinical Cancer Research*, 24(10):2251–2261, 2018.

[199] Kyung H Yi, Jossette Axtmayer, John P Gustin, Anandita Rajpurohit, and Josh Lauring. Functional analysis of non-hotspot AKT1 mutants found in human breast cancers identifies novel driver mutations: implications for personalized medicine. *Oncotarget*, 4(1):29–34, 2013.

[200] John D. Carpten, Andrew L. Faber, Candice Horn, Gregory P. Donoho, Stephen L. Briggs, Christiane M. Robbins, Galen Hostetter, Sophie Boguslawski, Tracy Y. Moses, Stephanie Savage, Mark Uhlik, Aimin Lin, Jian Du, Yue Wei Qian, Douglas J. Zeckner, Greg Tucker-Kellogg, Jeffrey Touchman, Ketan Patel, Spyro Mousses, Michael Bittner, Richard Schevitz, Mei Huei T. Lai, Kerry L. Blanchard, and James E. Thomas. A transforming mutation in the pleckstrin homology domain of AKT1 in cancer. *Nature*, 448(7152):439–444, 2007.

[201] Ivan Adzhubei, Daniel M. Jordan, and Shamil R. Sunyaev. *Predicting functional effect of human missense mutations using PolyPhen-2*. 2013.

[202] Anand Mayakonda, De Chen Lin, Yassen Assenov, Christoph Plass, and H. Phillip Koeffler. Maftools: Efficient and comprehensive analysis of somatic variants in cancer. *Genome Research*, 28(11):1747–1756, 2018.

[203] Samuel A. Wells, Furio Pacini, Bruce G. Robinson, and Massimo Santoro. Multiple endocrine neoplasia type 2 and familial medullary thyroid carcinoma: An update. *Journal of Clinical Endocrinology and Metabolism*, 98(8):3149–3164, 2013.

[204] Livia Manzella, Stefania Stella, Maria Stella Pennisi, Elena Tirrò, Michele Massimino, Chiara Romano, Adriana Puma, Martina Tavarelli, and Paolo Vigneri. New insights in thyroid cancer and p53 family proteins. *International Journal of Molecular Sciences*, 18(6), 2017.

[205] 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, Marcin Imielinsk, 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, 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(7463):415–21, 8 2013.

[206] Y Lio et al. Human rad51c deficiency destabilizes xrcc3, impairs recombination, and radiosensitizes s/g2-phase cells. *Journal of Biological Chemistry*, 279:42313–42320, 2004.

[207] Cynthia Rothblum-Oviatt, Jennifer Wright, Maureen A Lefton-Greif, Sharon A Mcgrath-Morrow, Thomas O Crawford, and Howard M Lederman. Orphanet Journal of Rare Diseases. *Orphanet Journal of Rare Diseases*, 11:1–21, 2016.

[208] A Sigurdson and D Stram. Genetic predisposition to radiation-related cancer and potential implications for risk assessment. *Annals of the ICRP*, 41:108–116, 2012.

[209] P Bhatti et al. Novel breast cancer risk alleles and interaction with ionizing radiation among u.s. radiologic technologists. *Radiation Research*, 173:214–224, 2010.

[210] Erik Garrison and Gabor Marth. Haplotype-based variant detection from short-read sequencing. *arXiv*, 2012.

[211] Xiaoming Liu, Chunlei Wu, Chang Li, and Eric Boerwinkle. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Human Mutation*, 37(3):235–241, 2016.

[212] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, Karl Voelkerding, and Heidi L. Rehm. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17(5):405–424, 2015.

[213] Melissa J. Landrum, Jennifer M. Lee, George R. Riley, Wonhee Jang, Wendy S. Rubinstein, Deanna M. Church, and Donna R. Maglott. ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(D1):980–985, 2014.

[214] Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92, 2012.

[215] K Magaard Koldby et al. Somatically acquired structural genetic differences: A longitudinal study of elderly danish twins. *European Journal of Human Genetics*, 24:1506–1510, 2016.

[216] L Forsberg et al. Age-related somatic structural changes in the nuclear genome of human blood cells. *American Journal of Human Genetics*, 90:217–228, 2012.

[217] J Demeulemeester, M Tarabichi, M Fittall, et al. Patterns of clustered mutational processes: Pan-cancer analysis of chromothripsis, chromoplexy and kataegis. volume 3, 2018.

[218] Dora Dias-Santagata, Kirsten Kübler, Daniel Rosebrock, Samuel E. Donovan, Gilbert H. Daniels, Carrie C. Lubitz, Dimitri Livitz, Raj K. Gopal, Peter F. Arndt, A. John Iafrate, Julian M. Hess, Lori Wirth, David G. McFadden, Braidie Campbell, Sareh Parangi, Peter M. Sadow, Ignaty Leshchiner, Jiwoong Kim, Chip Stewart, Scott Mordecai, Angela R. Shih, Gad Getz, Sarah E. Calvo, Vamsi K. Mootha, Benjamin J. Gigliotti, Zenon Grabarek, Tiannan Zhan, Lior Z. Braunstein, Paz Polak, Frances Chaves, and Salma Amin. Widespread Chromosomal Losses and Mitochondrial DNA Alterations as Genetic Drivers in Hürthle Cell Carcinoma. *Cancer Cell*, 34(2):242–255.e5, 2018.

[219] Sara Ahmadi, Michael Stang, Xiaoyin Sara Jiang, and Julie Ann Sosa. Hürthle cell carcinoma: Current perspectives. *OncoTargets and Therapy*, 9:6873–6884, 2016.

[220] Yuri E Nikiforov. RET / PTC Rearrangement in Thyroid Tumors. *Endocrine Pathology*, 13(1):3–16, 2002.

[221] Claudia Be´u Volpato, Minerva Martı´nez-Alfaro, 4 Raffaella Corvi, 2 Coralie Gabus, 5 Sylvie Sauvaigo, 3 Pietro Ferrari, 6 Elena Bonora, 1 Alessandro De Grandi, and Giovanni Romoe. Enhanced Sensitivity of the RET Proto-Oncogene to Ionizing Radiation In vitro Claudia Be. (21):8986–8993, 2008.

[222] a Bounacer, R Wicker, B Caillou, a F Cailleux, a Sarasin, M Schlumberger, and H G Suárez. High prevalence of activating ret proto-oncogene rearrangements, in thyroid tumors from patients who had received external radiation. *Oncogene*, 15(11):1263–1273, 1997.

[223] Julio C. Ricarte-Filho, Sheng Li, Maria E R Garcia-Rendueles, Cristina Montero-Conde, Francesca Voza, Jeffrey a. Knauf, Adriana Heguy, Agnes Viale, Tetyana Bogdanova, Geraldine a. Thomas, Christopher E. Mason, and James a. Fagin. Identification of kinase fusion oncogenes in post-Chernobyl radiation-induced thyroid cancers. *Journal of Clinical Investigation*, 123(11):4935–4944, 2013.

[224] Jason D. Prescott and Martha A. Zeiger. The RET oncogene in papillary thyroid carcinoma. *Cancer*, 121(13):2137–2146, 2015.

[225] Christopher B. Umbricht, Motoyasu Saji, William H. Westra, Robert Udelsman, Martha A. Zeiger, and Saraswati Sukumar. Telomerase activity: A marker to distinguish follicular thyroid adenoma from carcinoma. *Cancer Research*, 57(11):2144–2147, 1997.

[226] Claire Bournaud, Françoise Descotes, Myriam Decaussin-Petrucci, Julien Berthiller, Christelle de la Fouchardière, Anne Laure Giraudet, Mireille Bertholon-Gregoire, Philip Robinson, Jean Christophe Lifante, Jonathan Lopez, and Françoise Borson-Chazot. TERT promoter mutations identify a high-risk group in metastasis-free advanced thyroid carcinoma. *European Journal of Cancer*, 108:41–49, 2019.

[227] Malori A. Lankenau, Ravi Patel, Sandya Liyanarachchi, Sophia E. Maharry, Kevin W. Hoag, Megan Duggan, Christopher J. Walker, Joseph Markowitz, William E. Carson, Ann Kathrin Eisfeld, and Albert De La Chapelle. MicroRNA-3151 inactivates TP53 in BRAF-mutated human malignancies. *Proceedings of the National Academy of Sciences of the United States of America*, 112(49):E6744–E6751, 2015.

[228] Ingrid Slade, Chiara Bacchelli, Helen Davies, Anne Murray, Fatemeh Abbaszadeh, Sandra Hanks, Rita Barfoot, Amos Burke, Julia Chisholm, Martin Hewitt, Helen Jenkinson, Derek King, Bruce Morland, Barry Pizer, Katrina Prescott, Anand Saggar, Lucy Side, Heidi Traunecker, Sucheta Vaidya, Paul Ward, P. Andrew Futreal, Gordan Vujanic, Andrew G. Nicholson, Neil Sebire, Clare Turnbull, John R. Priest, Kathryn Pritchard-Jones, Richard Houlston, Charles Stiller, Michael R. Stratton, Jenny Douglas, and Nazneen Rahman. DICER1 syndrome: Clarifying the

diagnosis, clinical features and management implications of a pleiotropic tumour predisposition syndrome. *Journal of Medical Genetics*, 48(4):273–278, 2011.

[229] Steven A. Roberts, Michael S. Lawrence, Leszek J. Klimczak, Sara A. Grimm, David Fargo, Petar Stojanov, Adam Kiezun, Gregory V. Kryukov, Scott L. Carter, Gordon Saksena, Shawn Harris, Ruchir R. Shah, Michael A. Resnick, Gad Getz, and Dmitry A. Gordenin. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nature Genetics*, 45(9):970–976, 2013.

[230] Ludmil B. Alexandrov, Philip H. Jones, David C. Wedge, Julian E. Sale, Peter J. Campbell, Serena Nik-Zainal, and Michael R. Stratton. Clock-like mutational processes in human somatic cells. *Nature Genetics*, 47(12):1402–1407, 2015.

[231] P. Armitage and R. Doll. The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer*, 8:1–12, 1954.

[232] C O Nordling. *British Journal of Cancer*, 7, 1953.

[233] Cristian Tomasetti et al. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proceedings of the National Academy of Sciences of the United States of America*, 112:118–123, 2015.

[234] Elisabeth Cardis, Ausrele Kesminiene, Victor Ivanov, Irina Malakhova, Yoshisada Shibata, Valeryi Khrouch, Vladimir Drozdovitch, Evaldas Maceika, Irina Zvonova, Oleg Vlassov, André Bouville, Guennadi Goulko, Masaharu Hoshi, Alexander Abrosimov, Jadvyga Anoshko, Larisa Astakhova, Sergey Chekin, Evgenyi Demidchik, Rosaria Galanti, Masahiro Ito, Elena Korobova, Evgenyi Lushnikov, Marat Maksioutov, Vladimir Masyakin, Alexander Neronvia, Vladimir Parshin, Evgenyi Parshkov, Nikolay Piliptsevich, Aldo Pinchera, Semyon Polyakov, Nina Shabeka, Eero Suonio, Vanessa Tenet, Anatoli Tsyb, Shunichi Yamashita, and Dillwyn Williams. Risk of thyroid cancer after exposure to 131I in childhood. *Journal of the National Cancer Institute*, 97(10):724–732, 2005.

[235] VA Stsjazhko, AF Tsyb, ND Tronko, G Souchkevitch, and KF Baverstock. Childhood thyroid cancer since accident at chernobyl. *BMJ*, 310:801, 1995.

[236] Eric T. Dawson and Richard Durbin. GFAKluge: A C++ library and command line utilities for the Graphical Fragment Assembly formats. *Journal of Open Source Software*, 4(33):1083, aug 2019.

[237] Danny Antaki, William M. Brandler, and Jonathan Sebat. SV 2 : Accurate structural variation genotyping and de novo mutation detection from whole genomes. *Bioinformatics*, 34(10):1774–1777, 2018.

[238] Sai Chen, Peter Krusche, Egor Dolzhenko, Rachel M. Sherman, Roman Petrovski, Felix Schlesinger, Michael Kirsche, David R. Bentley, Michael C. Schatz, Fritz J. Sedlazeck, and Michael A. Eberle. Paragraph: A graph-based structural variant genotyper for short-read sequence data. *bioRxiv*, page 635011, 2019.

[239] Colby Chiang, Ryan M Layer, Gregory G Faust, Michael R Lindberg, David B Rose, Erik P Garrison, Gabor T Marth, Aaron R Quinlan, and Ira M Hall. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nature Methods*, 12(10):966–968, 2015.

[240] Mark J.P. Chaisson et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv*, 2017.

[241] Benedict Paten, Adam M Novak, Erik Garrison, and Glenn Hickey. Superbubbles, Ultrabubbles and Cacti. *bioRxiv*, 1(C):1–12, 2017.

[242] Jonas Andreas Sibbesen, Lasse Maretty, and Anders Krogh. Accurate genotyping across variant classes and lengths using variant graphs. *Nature Genetics*, 50(7):1054–1059, 2018.

[243] Justin M Zook, Nancy F Hansen, Nathan D Olson, Lesley M Chapman, James C Mullikin, Stephen Sherry, Sergey Koren, Adam M Phillippy, Paul C Boutros, Sayed Mohammad E, Weichen Zhou, Ryan E Mills, and Jay M Sage. A robust benchmark for germline structural variant detection. 2019.

[244] Ljiljana Brankovic, Costas S. Iliopoulos, Ritu Kundu, Manal Mohamed, Solon P. Pissis, and Fatima Vayani. Linear-Time Superbubble Identification Algorithm for Genome Assembly. 2015.

[245] Glenn Hickey, Benedict Paten, Dent Earl, Daniel Zerbino, and David Haussler. HAL: A hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics*, 29(10):1341–1342, 2013.

[246] Heng Li. On a reference pan-genome model. https://github.com/lh3/gfatools/blob/master/doc/rGFA.md, 2019.

[247] Erik Garrison, Adam Novak, Glenn Hickey, Jouni Sirén, and Eric Dawson. vg : the variation graph toolkit. pages 1–31, 2016.

[248] Zamin Iqbal, Mario Caccamo, Isaac Turner, Paul Flicek, and Gil McVean. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*, 44(2):226–232, 2012.

[249] Yosef E Maruvka, Kent W Mouw, Rosa Karlic, Prasanna Parasuraman, Atanas Kamburov, Paz Polak, Nicholas J Haradhvala, Julian M Hess, Esther Rheinbay, Yehuda Brody, Amnon Koren, Lior Z Braunstein, Alan D Andrea, Michael S Lawrence, Adam Bass, and Andre Bernards. Analysis of somatic microsatellite indels identifies driver events in human tumors. *Nature Biotechnology*, (August), 2017.

[250] Yuri E. Dubrova. Long-term genetic effects of radiation exposure. *Mutation Research - Reviews in Mutation Research*, 544(2-3):433–439, 2003.

[251] Yuri E. Nikiforov, Marina Nikiforova, and James A. Fagin. Prevalence of minisatellite and microsatellite instability in radiation-induced post-Chernobyl pediatric thyroid carcinomas. *Oncogene*, 17(15):1983–1988, 1998.

[252] Rob Patro, Stephen M. Mount, and Carl Kingsford. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, 32(5):462–464, 2014.

[253] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, 2016.

[254] Sergey Koren, Arang Rhie, Brian P. Walenz, Alexander T. Dilthey, Derek M. Bickhart, Sarah B. Kingan, Stefan Hiendleder, John L. Williams, Timothy P.L. Smith, and Adam Phillippy. Complete assembly of parental haplotypes with trio binning. *bioRxiv*, page 271486, 2018.

[255] C. Titus Brown and Luiz Irber. sourmash: a library for MinHash sketching of DNA. *The Journal of Open Source Software*, 1(5):27, 2016.

[256] Alekh Agarwal, Olivier Chapelle, Miroslav Dudik, and John Langford. A Reliable Effective Terascale Linear Learning System. 2011.

[257] Koenraad Van Doorslaer, Qina Tan, Sandhya Xirasagar, Sandya Bandaru, Vivek Gopalan, Yasmin Mohamoud, Yentram Huyen, and Alison A. McBride. The Papillomavirus Episteme: A central resource for papillomavirus sequence data and analysis. *Nucleic Acids Research*, 41(D1):571–578, 2013.

[258] Camilla L.C. Ip, Matthew Loose, John R. Tyson, Mariateresa de Cesare, Bonnie L. Brown, Miten Jain, Richard M. Leggett, David A. Eccles, Vadim Zalunin, John M. Urban, Paolo Piazza, Rory J. Bowden, Benedict Paten, Solomon Mwaigwisya, Elizabeth M. Batty, Jared T. Simpson, Terrance P. Snutch, Ewan Birney, David Buck, Sara Goodwin, Hans J. Jansen, Justin O'Grady, and Hugh E. Olsen. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Research*, 4(1075):1–35, 2015.

[259] Steven Flygare, Keith Simmon, Chase Miller, Yi Qiao, Brett Kennedy, Tonya Di Sera, Erin H. Graf, Keith D. Tardif, Aurélie Kapusta, Shawn Rynearson, Chris Stockmann, Krista Queen, Suxiang Tong, Karl V. Voelkerding, Anne Blaschke, Carrie L. Byington, Seema Jain, Andrew Pavia, Krow Ampofo, Karen Eilbeck, Gabor Marth, Mark Yandell, and Robert Schlaberg. Taxonomer: An interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome Biology*, 17(1):1–18, 2016.

[260] José M. Cuevas, Ron Geller, Raquel Garijo, José López-Aldeguer, and Rafael Sanjuán. Extremely High Mutation Rate of HIV-1 In Vivo. *PLoS Biology*, 13(9):1–19, 2015.

[261] Joshua Quick, Nicholas J. Loman, Sophie Duraffour, Jared T. Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, Raymond Koundouno, Gytis Dudas, Amy Mikhail, Nobila Ouédraogo, Babak Afrough, Amadou Bah, Jonathan H. J. Baum, Beate Becker-Ziaja, Jan Peter Boettcher, Mar Cabeza-Cabrerizo, Álvaro Camino-Sánchez, Lisa L. Carter, Juliane Doerrbecker, Theresa Enkirch, Isabel García Dorival, Nicole Hetzelt, Julia Hinzmann, Tobias Holm, Liana Eleni Kafetzopoulou, Michel Koropogui, Abigael Kosgey, Eeva Kuisma, Christopher H. Logue,

Antonio Mazzarelli, Sarah Meisel, Marc Mertens, Janine Michel, Didier Ngabo, Katja Nitzsche, Elisa Pallasch, Livia Victoria Patrono, Jasmine Portmann, Johanna Gabriella Repits, Natasha Y. Rickett, Andreas Sachse, Katrin Singethan, Inês Vitoriano, Rahel L. Yemanaberhan, Elsa G. Zekeng, Trina Racine, Alexander Bello, Amadou Alpha Sall, Ousmane Faye, Oumar Faye, N'Faly Magassouba, Cecelia V. Williams, Victoria Amburgey, Linda Winona, Emily Davis, Jon Gerlach, Frank Washington, Vanessa Monteil, Marine Jourdain, Marion Bererd, Alimou Camara, Hermann Somlare, Abdoulaye Camara, Marianne Gerard, Guillaume Bado, Bernard Baillet, Déborah Delaune, Koumpingnin Yacouba Nebie, Abdoulaye Diarra, Yacouba Savane, Raymond Bernard Pallawo, Giovanna Jaramillo Gutierrez, Natacha Milhano, Isabelle Roger, Christopher J. Williams, Facinet Yattara, Kuiama Lewandowski, James Taylor, Phillip Rachwal, Daniel J. Turner, Georgios Pollakis, Julian A. Hiscox, David A. Matthews, Matthew K. O' Shea, Andrew McD. Johnston, Duncan Wilson, Emma Hutley, Erasmus Smit, Antonino Di Caro, Roman Wölfel, Kilian Stoecker, Erna Fleischmann, Martin Gabriel, Simon A. Weller, Lamine Koivogui, Boubacar Diallo, Sakoba Keïta, Andrew Rambaut, Pierre Formenty, Stephan Günther, and Miles W. Carroll. Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530(7589):228–232, 2016.

[262] Nuno Rodrigues Faria, Ester C. Sabino, Marcio R. T. Nunes, Luiz Carlos Junior Alcantara, Nicholas J. Loman, and Oliver G. Pybus. Mobile real-time surveillance of Zika virus in Brazil. *Genome Medicine*, 8(1):97, 2016.

[263] N. R. Faria, J. Quick, I.M. Claro, J. Thézé, J. G. de Jesus, M. Giovanetti, M. U. G. Kraemer, S. C. Hill, A. Black, A. C. da Costa, L. C. Franco, S. P. Silva, C.-H. Wu, J. Raghwani, S. Cauchemez, L. du Plessis, M. P. Verotti, W. K. de Oliveira, E. H. Carmo, G. E. Coelho, A. C. F. S. Santelli, L. C. Vinhal, C. M. Henriques, J. T. Simpson, M. Loose, K. G. Andersen, N. D. Grubaugh, S. Somasekar, C. Y. Chiu, J. E. Muñoz-Medina, C. R. Gonzalez-Bonilla, C. F. Arias, L. L. Lewis-Ximenez, S. A. Baylis, A. O. Chieppe, S. F. Aguiar, C. A. Fernandes, P. S. Lemos, B. L. S. Nascimento, H. A. O. Monteiro, I. C. Siqueira, M. G. de Queiroz, T. R. de Souza, J. F. Bezerra, M. R. Lemos, G. F. Pereira, D. Loudal, L. C. Moura, R. Dhalia, R. F. França, T. Magalhães, E. T. Marques, T. Jaenisch, G. L. Wallau, M. C. de Lima, V. Nascimento, E. M. de Cerqueira, M. M. de Lima, D. L. Mascarenhas, J. P. Moura Neto, A. S. Levin, T. R. Tozetto-Mendoza, S. N. Fonseca, M. C. Mendes-Correa, F. P. Milagres, A. Segurado, E. C. Holmes, A. Rambaut, T. Bedford, M. R. T. Nunes, E. C. Sabino, L. C. J. Alcantara, N. J. Loman, and O. G. Pybus. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature*, 546(7658):406–410, 2017.

[264] Rayan Chikhi, Antoine Limasset, and Paul Medvedev. Compacting de Bruijn graphs from sequencing data quickly and in low memory. *Bioinformatics*, 32(12):i201–i208, 2016.

[265] Guillaume Marçais and Carl Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, 2011.

[266] Christopher A. Miller, Brian S. White, Nathan D. Dees, Malachi Griffith, John S. Welch, Obi L. Griffith, Ravi Vij, Michael H. Tomasson, Timothy A. Graubert, Matthew J. Walter, Matthew J. Ellis, William Schierding, John F. DiPersio,

Timothy J. Ley, Elaine R. Mardis, Richard K. Wilson, and Li Ding. SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. *PLoS Computational Biology*, 10(8), 2014.

[267] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal RNA-Seq quantification. *arXiv*, 2015.

[268] Joshua Quick et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530(7589):228–232, 2016.

[269] N. R. Faria, J. Quick, I.M. Claro, J. Thézé, J. G. de Jesus, M. Giovanetti, M. U. G. Kraemer, S. C. Hill, A. Black, A. C. da Costa, L. C. Franco, S. P. Silva, C.-H. Wu, J. Raghwani, S. Cauchemez, L. du Plessis, M. P. Verotti, W. K. de Oliveira, E. H. Carmo, G. E. Coelho, A. C. F. S. Santelli, L. C. Vinhal, C. M. Henriques, J. T. Simpson, M. Loose, K. G. Andersen, N. D. Grubaugh, S. Somasekar, C. Y. Chiu, J. E. Muñoz-Medina, C. R. Gonzalez-Bonilla, C. F. Arias, L. L. Lewis-Ximenez, S. A. Baylis, A. O. Chieppe, S. F. Aguiar, C. A. Fernandes, P. S. Lemos, B. L. S. Nascimento, H. A. O. Monteiro, I. C. Siqueira, M. G. de Queiroz, T. R. de Souza, J. F. Bezerra, M. R. Lemos, G. F. Pereira, D. Loudal, L. C. Moura, R. Dhalia, R. F. França, T. Magalhães, E. T. Marques, T. Jaenisch, G. L. Wallau, M. C. de Lima, V. Nascimento, E. M. de Cerqueira, M. M. de Lima, D. L. Mascarenhas, J. P. Moura Neto, A. S. Levin, T. R. Tozetto-Mendoza, S. N. Fonseca, M. C. Mendes-Correa, F. P. Milagres, A. Segurado, E. C. Holmes, A. Rambaut, T. Bedford, M. R. T. Nunes, E. C. Sabino, L. C. J. Alcantara, N. J. Loman, and O. G. Pybus. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature*, 546(7658):406–410, 2017.

[270] Heng Li. Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, 32(14):2103–2110, 2016.

[271] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 00(00):1–3, 2013.

Appendix A

Related publications

I have been an author on several publications during my PhD.

- Garrison, Erik, Jouni Sirén, Adam M. Novak, Glenn Hickey, Jordan M. Eizenga, Eric T. Dawson, William Jones et al. "Variation graph toolkit improves read mapping by representing genetic variation in the reference." *Nature Biotechnology*, 36(9):875–879, (2018).
- Eric Dawson and Richard Durbin. "GFAKluge: A C++ library and command line utilities for the Graphical Fragment Assembly formats." *Journal of Open-Source Software*, 4(33):1083, (2019).
- Eric T Dawson, Sarah Wagner, David Roberson, Meredith Yeager, Joe Boland *et al.* "Viral coinfection analysis using a MinHash toolkit." *BMC Bioinformatics*, 20(389):1-10, (2019).
- Lesley M. Chapman *et al.* "SVCurator: A Crowdsourcing app to visualize evidence of structural variants for the human genome." *bioRxiv* preprint, submitted to *PLoS Computational Biology* September (2019).
- Glenn Hickey *et al.* "Genotyping structural variants in pangenome graphs using the vg toolkit." *bioRxiv* preprint (2019).

I will also be a joint author on a manuscript describing the analyses in Chapter 2, to be submitted in the near future.