A Practical Guide for Structural Variation Detection in the Human Genome

Lixing Yang[1,2]

[1]Ben May Department for Cancer Research, Department of Human Genetics, University of Chicago, Chicago, Illinois
[2]Corresponding author: *lixingyang@uchicago.edu*

Profiling genetic variants—including single nucleotide variants, small insertions and deletions, copy number variations, and structural variations (SVs)—from both healthy individuals and individuals with disease is a key component of genetic and biomedical research. SVs are large-scale changes in the genome and involve breakage and rejoining of DNA fragments. They may affect thousands to millions of nucleotides and can lead to loss, gain, and reshuffling of genes and regulatory elements. SVs are known to impact gene expression and potentially result in altered phenotypes and diseases. Therefore, identifying SVs from the human genomes is particularly important. In this review, I describe advantages and disadvantages of the available high-throughput assays for the discovery of SVs, which are the most challenging genetic alterations to detect. A practical guide is offered to suggest the most suitable strategies for discovering different types of SVs including common germline, rare, somatic, and complex variants. I also discuss factors to be considered, such as cost and performance, for different strategies when designing experiments. Last, I present several approaches to identify potential SV artifacts caused by samples, experimental procedures, and computational analysis. © 2020 Wiley Periodicals LLC.

Keywords: chromothripsis • genomic rearrangements • next-generation sequencing • single-molecule sequencing

How to cite this article:
Yang, L. (2020). A practical guide for structural variation detection in the human genome. *Current Protocols in Human Genetics*, *107,* e103.
doi: 10.1002/cphg.103

INTRODUCTION

Structural variations (SVs) are an important class of genetic variants in the human genome (Feuk, Carson, & Scherer, 2006; Stankiewicz & Lupski, 2010). They include deletions, duplications, insertions, inversions, translocations, and other more complex forms (Fig. 1). Some SVs, such as deletions and duplications, change the dosage of DNA and are considered copy number variations (CNVs), while others, such as inversions and balanced translocations, do not change the DNA dosage. The major difference between SVs and CNVs is that SVs always involve breakage and rejoining of DNA fragments. Hence, events like whole chromosomal gains and losses are not considered SVs. In a given human genome, germline SVs typically affect an order of magnitude more nucleotides than single nucleotide polymorphisms (SNPs; Mills et al., 2011). The size of an SV can range from dozens to millions of base pairs. Conventionally, SVs <50 bp in size are called small insertion/deletions (indels). The strategies of indel discovery are very different from large SVs (Mullaney, Mills, Pittard, & Devine, 2010; Xu, 2018). In this guide, I only discuss SVs that are >50 bp.

Other than the simple forms of SVs, more complex SVs are also frequent in normal individuals as well as in patients with genetic disorders, such as duplication-inverted triplication-duplication (Carvalho et al.,

Current Protocols in Human Genetics e103, Volume 107
Published in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cphg.103

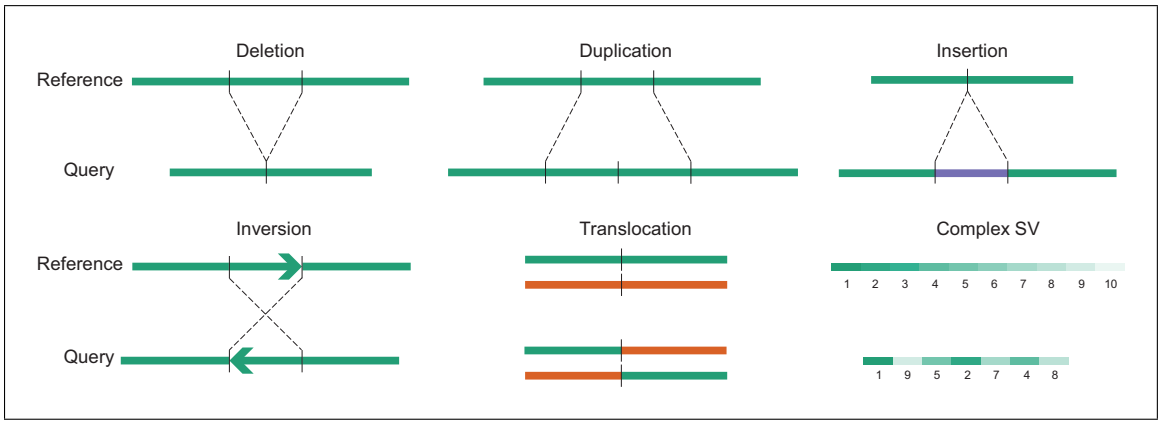

Figure 1 Schematic of different types of structural variations (SVs).

2011), insertion or inversion inside deletion (Kidd et al., 2010; Yang et al., 2013), and templated insertion (Li et al., 2020), among others. Recently, an extremely complex form of SVs was described in cancer called chromothripsis, in which dozens to hundreds of breakpoints on one or a few chromosomes are involved (Stephens et al., 2011). Such event is considered to occur at one time rather than many simple SVs being accumulated over a long period of time. Chromothripsis was originally reported in many different types of cancers (Molenaar et al., 2012; Rausch et al., 2012a; Stephens et al., 2011) and was also found in germline genomes causing developmental and neuronal disorders (Chiang et al., 2012; Liu et al., 2011). Similar complex events have been named chromoanasythesis (Liu et al., 2011), chromoanagenesis (Holland & Cleveland, 2012), and chromoplexy (Baca et al., 2013) in various contexts.

SVs are usually caused by erroneous DNA replication and DNA damage repair, as well as activities of repetitive elements (Cordaux & Batzer, 2009; Lee et al., 2012). A number of molecular mechanisms are known to form SVs in germline and somatic cells (Bunting & Nussenzweig, 2013; Hastings, Lupski, Rosenberg, & Ira, 2009). These include nonhomologous end joining, alternative end joining, nonallelic homologous recombination, single-strand annealing, break-induced replication, and fork stalling and template switching. Chromothripsis is known to be induced by micronuclei formed via chromosomal segregation error (CZ Zhang et al., 2015), chromatin bridges due to telomere attrition (Maciejowski, Li, Bosco, Campbell, & de Lange, 2015), template switching at stalled replication forks (Liu et al., 2011; Yang et al., 2013), and breakage-fusion-bridge cycles (Li et al., 2014).

There are two main impacts of SVs: change of DNA dosage and change of DNA order. First, gains and losses of important genes and regulatory elements owing to SVs can impact phenotype and cause diseases (Stankiewicz & Lupski, 2010; Weischenfeldt, Symmons, Spitz, & Korbel, 2013). Dosage changes of genes and their contributions to diseases have been extensively studied (Stankiewicz & Lupski, 2010; Tang & Amon, 2013; Zhang, Gu, Hurles, & Lupski, 2009). Recently, it has been reported that duplications or deletions of enhancers and super-enhancers lead to misregulation of target genes (e.g., *MYC*, *AR*, *SOX9* and *KLF5*) and cause diseases such as cancer and sex development disorders (Croft et al., 2018; Takeda et al., 2018; X Zhang et al., 2015, 2018). Second, SVs can reorganize DNA content and connect two distal fragments together. This leads to gene fusions and chimeric proteins when two distinct genes are joined into one. Gene fusions are often major cancer-driving events, especially in pediatric cancers and liquid tumors (Mertens, Johansson, Fioretos, & Mitelman, 2015). Moreover, the interaction between genes and regulatory elements can also be altered by SVs. A number of oncogenes, such as *MYC*, *BCL2*, *EVI1*, *TERT*, and *GFI1*, are activated by distal enhancers through somatic SVs (Boxer & Dang, 2001; Davis et al., 2014; Gröschel et al., 2014; Northcott et al., 2014; Valentijn et al., 2015). When enhancer–gene interactions are rewired by various types of SVs such as deletions, duplications, or inversions around the WNT6/IHH/EPHA4/PAX3 locus, the misregulated genes can lead to different forms of limb malformation (Lupiáñez et al., 2015). Duplications of different regions near *SOX9* can cause sex reversal or limb malformation depending on the type of newly formed gene–enhancer interactions (Franke

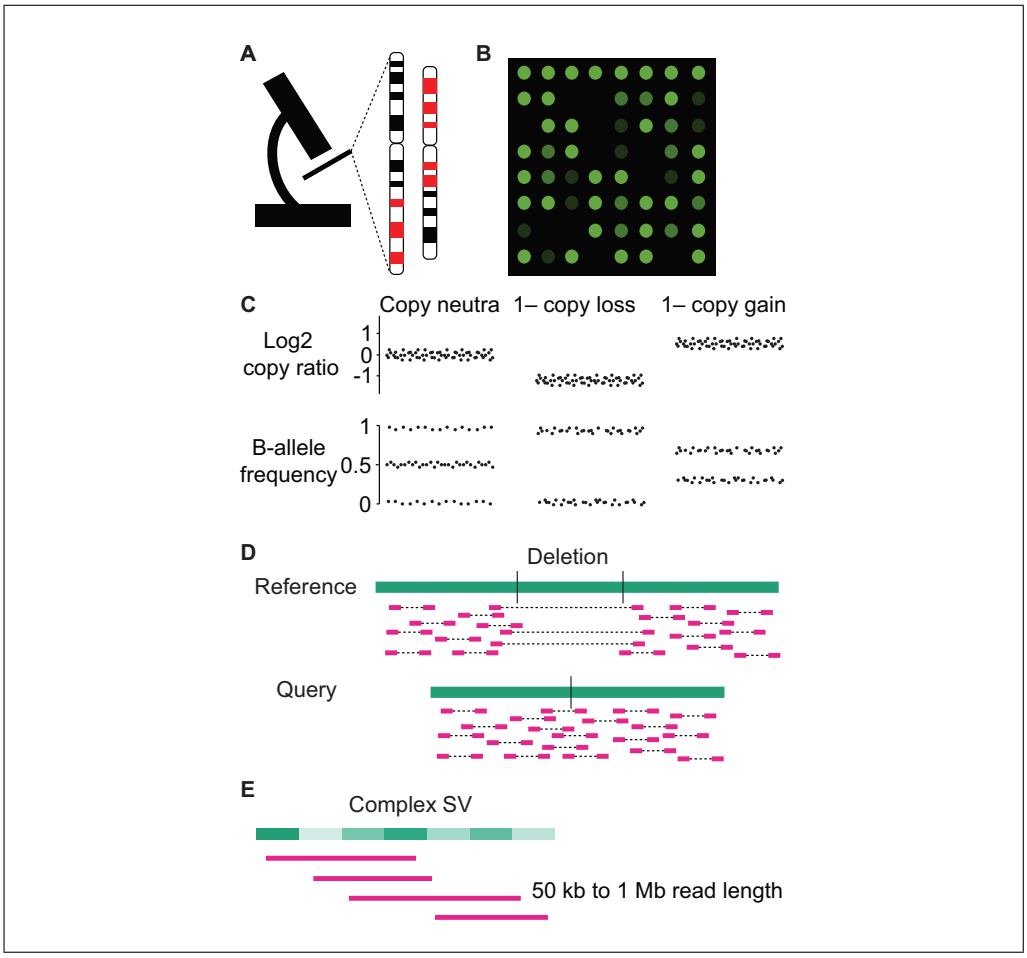

Figure 2 Structural variation (SV) detection platforms. (**A**) Karyotyping; (**B**) DNA microarray; (**C**) copy ratio and B-allele frequency profiles for copy neutral, one-copy loss, and one-copy gain regions; (**D**) short-read sequencing; and (**E**) long-read sequencing.

et al., 2016). Furthermore, inherited rare SVs in cis-regulatory elements are found to be associated with autism (Brandler et al., 2018).

METHODS OVERVIEW

Here, I briefly describe methods used for SV discovery. The main goal for these methods is to detect unknown SVs. Methods to confirm if a particular SV exists in a genome are described in the Validation and Genotyping section.

Cytogenetics

Cytogenetic testing used to be routinely performed for diagnosis and screening of genetic diseases. Recurrent translocations were found in cancer by cytogenetic analysis decades ago. Karyotyping (Fig. 2A) is the most common cytogenetic testing technique (Wan, 2014). Dividing cells are required to view condensed chromosomes at metaphase. Chromosomes are stained, or colored probes are hybridized to chromosomes. Chromosomal banding patterns are visualized with a microscope. Only large SVs visible under the microscope can be detected, such as translocations and deletions, duplications, and inversions that are at least 5 Mb in size. Mosaic events (where SVs exist in a subset of cells but not all cells) are less likely to be detected. SV breakpoints are not precise. Complex SVs are typically not detectable. Nowadays, karyotyping is rarely used for SV discovery owing to its low sensitivity and precision.

Microarray

DNA microarray (Fig. 2B) is an advanced version of florescence in situ hybridization (FISH), described in the Validation and Genotyping section, where probes are designed to hybridize with DNA, and the readout is a florescent signal (see Current Protocols article: Bumgarner, 2013; Heller, 2002). On a microarray, thousands to millions of probes are printed on a very dense surface. The intensity of the florescent signal represents the amount of DNA that can hybridize to the

probes. This feature can be used to quantify the copy number of DNA (Fig. 2C, top panel). Microarray-based comparative genomic hybridization is particularly designed to detect CNVs (Lockwood, Chari, Chi, & Lam, 2006; Pinkel & Albertson, 2005). SNP genotyping arrays can also be used to measure DNA copy number (see Current Protocols article: Lin, Naj, & Wang, 2013; Schaaf, Wiszniewska, & Beaudet, 2011). The current cost for human DNA microarray ranges from $100 to $500 per sample depending on the probe density. For SNP arrays, in addition to probe hybridization intensity, the minor allele (B-allele) frequencies can also be used to infer CNVs. For example, in copy neutral regions, germline heterozygous SNPs should have B-allele frequencies of 0.5. In one-copy loss regions, B-allele frequencies will become 0 or 1. Similarly, they will be 0.33 and 0.67 in one-copy gain regions. Since there is always noise in the florescent signal, CNVs are typically called when there are several consecutive probes supporting the events. CNVs <50 kb in size are usually undetectable with microarray. Breakpoints cannot be precisely determined. In addition, balanced SVs are not detectable because there is no DNA dosage change. Due to its high-throughput nature, microarray was widely used to study CNVs in normal populations (Conrad, Andrews, Carter, Hurles, & Pritchard, 2005; McCarroll et al., 2008), as well as in individuals with disease (Bochukova et al., 2010; Sebat et al., 2007; Zack et al., 2013). Although microarray has been mostly replaced by second-generation sequencing in scientific research, it is still commonly used for clinical diagnosis of genetic disorders.

Second-Generation Sequencing

More than 10 years ago, a number of second-generation sequencing technologies became available, including pyrosequencing (Roche 454), sequencing by synthesis (Illumina/Solexa), sequencing by ligation (Life Technologies SOLiD), nanoball sequencing (Complete Genomics/BGI), and Ion Torrent sequencing (Life Technologies), among others. Over the years, sequencing quality has improved significantly while the cost continues to drop. Whole-genome sequencing (WGS) has become quite affordable and has enabled SV detection at the base pair resolution (Fig. 2D). The current cost for library preparation and Illumina sequencing (150 bp paired-end) of a 30× human genome is between $800 and $1300. In second-generation sequencing experiments, genomic DNA is first shredded into small fragments and then sequenced from both ends of the DNA molecules (paired-end sequencing). Sequencing reads are typically short (<500 bp). With high-coverage WGS, since all genomic regions are sequenced, theoretically all SVs may be detected, including balanced SVs and complex events. However, due to the limitation of short read length and the repetitive nature of the human genome, SVs in repetitive regions and segmental duplicated regions remain difficult to identify. Furthermore, sampling bias and gaps in the reference genome will also affect SV detection. Fresh and fresh frozen samples are recommended to study SVs for both second- and third-generation platforms. In formalin-fixed paraffin-embedded (FFPE) samples, DNA is highly degraded, and chimeric molecules are abundant. There will be many artifact SVs detected. Therefore, FFPE samples are generally not recommended for SV discovery.

Third-Generation Sequencing and Imaging

Third-generation sequencing technologies feature long reads. PacBio single-molecule real-time (SMRT) sequencing passes single-stranded DNA through an immobilized DNA polymerase to detect florescence. Oxford Nanopore detects bases via an ion current when DNA or RNA passes through the protein nanopores. Both technologies can produce kb to Mb sequences at the single-molecule level without amplifying the templates. The cost for a 30× human genome sequenced by both PacBio continuous long reads (CLR) and Nanopore PromethION, including library preparation, ranges between $3000 and $5000. A major drawback of long-read sequencing is the high error rates (15% for PacBio and 30% for Nanopore). Hence, long-read sequencing technologies by themselves are great for SV discovery but not optimal for detection of single nucleotide variants (SNVs) or small indels. To overcome the high error rate, the latest HiFi reads from PacBio substantially improved the base quality by sequencing the same molecules multiple times. The cost of a 30× human genome with PacBio HiFi sequencing is about $15,000. An alternative approach is linked-read sequencing offered by 10X Genomics. Long DNA molecules are embedded in individual microfluidic droplets, called gel-bead in emulsion (GEMs), containing unique barcodes. DNA fragments are sequenced by the Illumina short-read sequencing platform, and the barcodes can be used to link short reads into longer contigs. The cost of

linked-read sequencing is about $2000 for a 30× human genome. In addition, microscopic imaging can provide genomic information through very long range as well. The genome mapping technology offered by BioNano Genomics labels specific sequence motifs in the genome and scans these labels by imaging. The maps of sequence motifs can then be used for SV detection and genome assembly. However, the resolution of SV breakpoints in optical mapping is much lower than sequencing-based approaches. The cost of BioNano Optical Mapping is about $1500 for a human genome. All of the third-generation sequencing and imaging technologies can overcome the major limitation of second-generation short-read sequencing. The long-range information is ideal to resolve complex SVs and SVs in repetitive regions (Fig. 2E). Note that third-generation sequencing and imaging platforms depend on high molecular weight (HMW) DNA so that the long DNA molecules can be sequenced or imaged. HMW DNA extraction costs another $500 to $1000 per sample on top of library preparation and sequencing costs. Fresh or fresh frozen samples are recommended for third-generation sequencing and imaging.

Validation and Genotyping

Validation and genotyping are often performed to confirm the presence of SVs in query samples or new samples. Usually, validation needs to be performed using an orthogonal method. The widely used methods are PCR with Sanger sequencing, FISH, or another sequencing platform that differs from the variant discovery platform. Genotyping can be done by PCR with Sanger sequencing, low-coverage WGS, or targeted sequencing.

STUDY DESIGN

Both karyotyping and microarray are rarely used for SV discovery any more due to their cost inefficiency, labor intensity, and low throughput. Sequencing and imaging have become the go-to choices. As noted above, third-generation sequencing and imaging technologies are superior but often cost more. Therefore, in a specific study, researchers always need to find the balance between cost and performance. Performance often depends on the choice of platform and sequencing coverage. Data analysis will certainly play an important role as well. Here, I provide some recommendations on different types of variants.

CNV Detection

If detecting CNVs is the major goal, and balanced SVs and the precise breakpoints do not matter, low-coverage short-read WGS will be the most cost-efficient method. At 2× coverage, Illumina WGS can achieve comparable or better accuracy and precision than SNP array (Muzny et al., 2012). The sequencing cost will be much less than microarray. However, with such coverage, other types of variants such as SNVs cannot be detected. In addition, breakpoints are not at the base pair resolution.

Common Germline Variants

SVs including CNVs are known to have more drastic effects on gene expression (Chiang et al., 2017; Stranger et al., 2007). Microarray has been used for genome-wide association studies on common germline CNVs (Craddock et al., 2010; Glessner et al., 2009). Common SVs can be used to study their associations with diseases as well. Low-coverage short-read WGS is again a cost-efficient choice for common variant detection. The rationale is that common variants are shared in the population. When sequencing many individuals, the signal can be combined for variant discovery. The 1000 Genomes Project has used this strategy to interrogate SVs, and the software Genome STRiP was designed for this particular purpose (Handsaker, Korn, Nemesh, & McCarroll, 2011). Once the variants are called, the same sequencing data can be used to genotype the variants in all individuals.

De Novo Variants and Rare Germline Variants

De novo variants are the ones not inherited from parents and likely occur during germ cell formation of the parents. Rare germline variants are inherited from parents, but the allele frequencies are very low in the general population. Both of these variants may be associated with diseases (Bochukova et al., 2010; Georgieva et al., 2014; Redin et al., 2017). Since rare variants and de novo variants are typically not shared between individuals, their discovery must rely on deep sequencing. High-coverage (>20×) short-read WGS is a very powerful method to detect SVs in non-repetitive regions. However, resolving SVs in repetitive regions is very challenging. A recent study using the very high–coverage (average depth 65×) PacBio long-read sequencing platform reported that as much as 87% of germline SVs discovered by long-read sequencing cannot be detected by short-read sequencing (Audano et al., 2019). Most of the SVs missed

by short-read sequencing are associated with variable number tandem repeats. Larger-scale efforts to elucidate the full spectrum of SVs in the human genome by integrating multiple sequencing platforms are ongoing. For individual studies, such as to identify disease-associated variants, it is recommended to perform second-generation sequencing initially. If there are reasons to believe pathogenic SVs are located in repetitive regions, third-generation sequencing technologies may be considered to extend the search.

Mosaic Variants and Somatic Variants

Mosaic variants and somatic variants both refer to variants that occur during cell division after the formation of fertilized eggs. Therefore, in an individual, only a subset of his/her cells carry such variants in contrast to germline variants and de novo variants, which are present in all cells. It is widely accepted that all somatic cells carry somatic mutations to some extent because DNA polymerase cannot replicate DNA 100% accurately. In every cell cycle, a small number of point mutations are accumulated. Sometimes, some cells will acquire growth advantage, expand the cell population (Blokzijl et al., 2016; Lodato et al., 2018; Martincorena et al., 2015), and even become cancerous. Tumor cells from a patient with cancer are mostly derived from a common ancestor cell, although in some cases it is possible for tumors to develop with multiple independent origins. Somatic variants acquired by the common ancestor cell are shared by all tumor cells. These variants are called clonal variants. During clonal expansion, additional somatic variants occur. The ones that occur after major clonal expansion and that are not shared by all tumor cells are subclonal variants, which may play important roles in tumor evolution and drug response (Schmitt, Loeb, & Salk, 2016). Mosaic variants and somatic variants are almost always private and not present in other individuals. Even for the highly frequent pathogenic SVs, such as translocations leading to *BCR-ABL1* fusion, the precise locations of translocation junctions are clustered but still differ between patients (Groffen et al., 1984). Therefore, pooling data from multiple individuals will not be helpful for SV discovery.

The difficulty in SV detection caused by repetitive regions described previously also applies to mosaic and somatic variants. The rationale to choose between short-read and long-read technologies presented above also applies here. A major issue that is specific to mosaic and somatic variants is sequencing depth. Typically, 30× sequencing coverage is sufficient to confidently detect 99% of the heterozygous variants that are present in all cells. However, for variants present in half of the cells, 60× coverage will be needed to achieve the same sensitivity. As a proof-of-concept, my group performed a simple simulation to test how sequencing coverage affects SV detection (Fig. 3). Various types of SVs, as well as Illumina sequencing reads, were simulated, and SVs were called by Meerkat (Yang et al., 2013). As expected, at 30× coverage, almost all SVs were detectable. At 20× coverage, about 80% of the SVs remained detectable. When the coverage was <10×, more than half of the SVs were missed. The undetected SVs were mostly due to lack of supporting reads. A similar trend was observed in SNV calling by down-sampling a 410× WGS dataset (Kishikawa et al., 2019). The simulation reflects the best-case scenario. Other factors such as repeats, sequencing bias, sequencing error, and chimeric molecules formed during library preparation are not considered. When deciding sequencing depth, aneuploidy, which is a hallmark of cancer, should be taken into account as well. Many chromosomes in tumor cells have more than two copies. With the same sequencing coverage, reads spanning the variants in aneuploid chromosomes will be less if the variants are present in only one copy of the chromosomes. Due to the existence of subclonal variants, in tumor sequencing studies more somatic variants can be identified with higher sequencing depth. The choice of sequencing coverage will depend on the importance of subclonal variants. If the goal is to find cancer-driving events and actionable variants, it will not be necessary to sequence very deeply, since the variants present in a very small fraction of cells are probably not the major drivers of the disease. It is recommended to select a coverage that allows 80% of the variants to be detectable. For example, if a tumor has a purity of 50% and is known to be aneuploid, 60× coverage would offer enough depth to detect the majority of clonal SVs and a good portion of subclonal variants. If the main goal is to interrogate the very low–frequency variants, it is recommended to test very high coverage, such as 200×, in one or two samples and down-sample the reads to see if a lower coverage can achieve satisfactory results.

Complex Variants

Mildly complex SVs can be reconstructed by integrating copy number profiles and

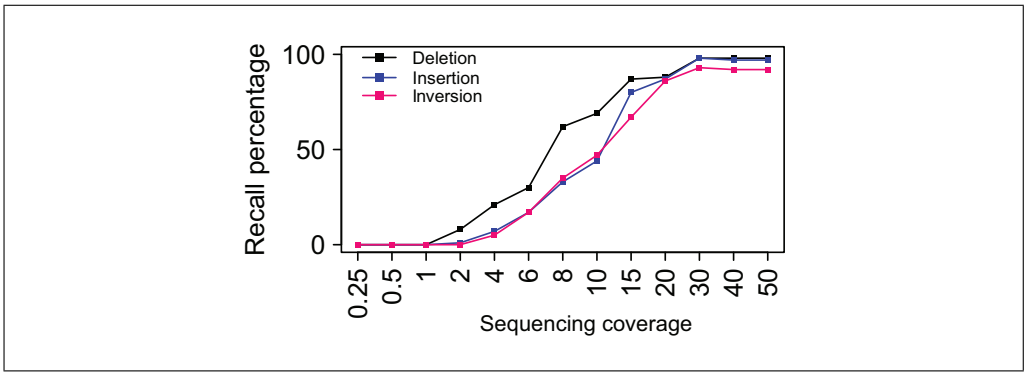

Figure 3 The effect of sequencing coverage on structural variation (SV) detection sensitivity. All simulated SVs are 1 kb in size. Synthetic heterozygous SVs are placed at uniquely mappable regions on a pseudo diploid chromosome. Illumina sequencing reads (paired-end 75 bp) are simulated.

SV junctions (Greenman et al., 2011). The structure of chromothripsis events can still be inferred from short-read sequencing data (Yang et al., 2013), but it is an extremely challenging task. Therefore, long-read sequencing and imaging technologies will be very powerful for uncovering the fine structures of very complex SVs. For example, linked-read sequencing was used to resolve haplotypes and somatic SVs in metastatic castration-resistant prostate cancers (Viswanathan et al., 2018). Short-read WGS, high-throughput chromosome conformation capture, and optical mapping were integrated to resolve complex SVs and phase multiple SVs to single haplotype (Dixon et al., 2018), and the structures of circular extra-chromosomal DNA in glioblastoma cell lines were determined by a combination of short-read WGS, optical mapping, and super-resolution microscopy (Wu et al., 2019). When the main goal is to characterize complex SVs, I recommend combining short- and long-read platforms in order to determine the long-range structure and the precise breakpoints at the same time. A combination of 30× short reads and 10× long reads will perform well. If clonality and aneuploidy cannot be ignored, higher sequencing depth will be required as described above.

Data Recycling

Although WGS is preferred for comprehensive SV detection, other types of sequencing data can still be used to identify SVs. For example, short read–based whole exome sequencing (WES) data are commonly used for CNV detection. In addition, if SV breakpoints are present in the enriched regions being sequenced, such as exons, they may be identified using standard SV-calling algorithms. It can be particularly fruitful if data are readily available on a large number of samples even with limited sensitivity. For examples, rare germline CNVs were studied using nearly 60,000 exomes (Ruderfer et al., 2016), and somatic SVs in cancer were discovered in 5000 tumor exomes (Yang et al., 2016). Although only 1% of the somatic SV breakpoints were detectable in WES data, a large portion of complex SVs such as chromothripsis could still be detected (Yang et al., 2016). The large sample size in those studies (Ruderfer et al., 2016; Yang et al., 2016) enabled meaningful biological inferences without generating any new data.

VARIANT CALLING

By far, there is no gold standard for how to call SVs from sequencing data. Numerous predicting algorithms are available, and the calls made by different algorithms on the same data may differ substantially. No algorithm can outperform others in all types of SVs and across all size ranges. Numerous published papers and reviews have compared and benchmarked the available tools extensively. However, benchmarking results also differ in different studies. The main reason is again the lack of gold standard. In this section, I briefly describe computational tools for variant detection and strategies to determine which tool(s) to use.

Many algorithms have been developed to detect CNVs in short-read WGS data based on read depth such as BIC-seq (Xi et al., 2011), CNVnator (Abyzov, Urban, Snyder, & Gerstein, 2011), and Control-FREEC (Boeva et al., 2012). A comprehensive list of CNV detection software can be found at *https://bioinformaticshome.com/tools/cnv/cnv.html*. Many tools have been reviewed (Hehir-Kwa, Pfundt, & Veltman, 2015; Pirooznia, Goes, & Zandi, 2015; Zhao, Wang, Wang, Jia, & Zhao,

2013) and benchmarked (Trost et al., 2018; Zhang, Bai, Yuan, & Du, 2019). Evaluation of CNV detection tools in exome sequencing and targeted sequencing data are also available (Kadalayil et al., 2014; Yao, Yu, Qing, Wang, & Shen, 2019; Zare, Dow, Monteleone, Hosny, & Nabavi, 2017).

For SV detection using short-read sequencing data, it is recommended to use BWA-MEM (Li, 2013) to align the reads to the reference genome because it can partially align reads. This function is particularly useful to detect reads spanning the SV breakpoints since they are marked as clipped reads. Several strategies can be used to predict SVs from short-read sequencing data (Alkan, Coe, & Eichler, 2011). Read depth can be used to interrogate SVs with copy number change. The presence of discordant read pairs (reads in a pair mapped to different chromosomes, in incompatible orientations, or not within the size limit of sequencing library) often suggests the presence of SVs. Reads spanning the SV breakpoint junction (split read) can be used to refine the precise locations of breakpoints. Reads around breakpoints can be assembled for SV detection. Dozens of computational algorithms have been developed using one or several of the above strategies, such as Meerkat (Yang et al., 2013), DELLY (Rausch et al., 2012b), Manta (Chen et al., 2016), and novoBreak (Chong et al., 2016), among others. A comprehensive list of SV detection software can be found at *https://omictools.com/structural-variant-detection-category*. Several benchmarking studies have tested the performances of many of these tools (Alaei-Mahabadi, Bhadury, Karlsson, Nilsson, & Larsson, 2016; Cameron, Di Stefano, & Papenfuss, 2019; Kosugi et al., 2019; Lee et al., 2018). If choosing only one SV caller, it is recommended to choose one that implements several SV detection strategies, such as Meerkat and Manta, because such software often performs better than ones that use only one SV discovery strategy. To achieve the highest accuracy, the best practice is to use several tools and integrate their SV calls to minimize caller-specific bias (Campbell et al., 2020; Sudmant et al., 2015). It will significantly increase the computational burden for large datasets. When SVs are called by multiple algorithms, one needs to identify the overlapping calls and merge them into a unified call set. Theoretically, the same SVs should have the same genomic coordinates called by different algorithms. However, read depth and read pair strategies cannot provide the precise location of breakpoints. In addition, SV breakpoints are often not blunt ends but carry homologous sequences or insertions (Yang et al., 2013). Different algorithms handle homology and insertion sequences differently and provide slightly different coordinates. Practically, if two SVs have the same breakpoint orientations at both breakpoint junctions and the distances between the corresponding breakpoints are within 50 bp, they can be considered the same SV. Breakpoint orientation is determined by read mapping orientation. For example, for the deletion shown in Figure 2D, the breakpoint on the left is typically marked with an orientation of "1" or "+" because the supporting reads near this breakpoints are mapped to the forward strand, while the breakpoint on the right is marked as "–1" or "–" (see Yang et al. 2013 for more examples). One widely used strategy to assemble a unified SV call set is that once the overlapping SVs are determined, caller-specific SVs can be removed. The remaining SVs supported by at least two callers are usually of high quality. Researchers can also consider different ways of combining SVs called by multiple algorithms, such as based on individual SV scores provided by callers. Different SV callers may be weighted differently. For example, if one algorithm is known to produce very few false calls, all SVs detected by this algorithm can be included in the final call set. Some tools, such as Parliament2 (Zarate et al., 2018) and FusorSV (Becker et al., 2018), run a suite of individual SV callers and provide an ensemble call set.

For long-read sequencing and imaging platforms, there is a number of vendor-provided software and open source tools for SV detection, such as SMRT Link (PacBio), EPI2ME (Nanopore), LongRanger (10X Genomics), Bionano Access (Bionano), NanoSV (Cretu Stancu et al., 2017), Picky (Gong et al., 2018), Sniffles (Sedlazeck et al., 2018), SMRT-SV (Chaisson et al., 2015), GROC-SVs (Spies et al., 2017), and LinkedSV (Fang et al., 2019). Benchmarking studies have shown significant differences between different platforms and computational tools (Audano et al., 2019; Chaisson et al., 2019; De Coster et al., 2019; Luan, Zhang, Zhu, Chen, & Xie, 2020; Zook et al., 2020). Therefore, it is recommended to combine multiple tools for SV detection to achieve the best sensitivity and accuracy. For SV detection combining multiple platforms, although most pipelines

are developed in house, a few of them are streamlined and publicly available such as Multibreak-SV (Ritz et al., 2014) and HySA (Fan, Chaisson, Nakhleh, & Chen, 2017).

QUALITY CONTROL

Quality control after SV calling is as important as choosing sequencing platforms and deciding SV calling algorithms in order to achieve satisfactory performance in SV detection. Poor-quality samples, suboptimal library preparation, sequencing, and SV calling algorithms may produce artifactual SV calls. Read-level quality control such as FastQC (Andrews, 2010) may not pick up issues that affect SV calling. These issues can be summarized as follows: (1) Whole-genome amplification can produce artifactual duplications (Yang et al., 2019). (2) An unknown source during library preparation induces small inversion-like SVs with microhomology (Campbell et al., 2020; Yang et al., 2016). (3) Random ligation of DNA fragments may produce chimeric molecules. (4) Germline SVs may not be properly removed when calling somatic and de novo SVs due to poor-quality control samples or inadequate sequencing depth of control samples. (5) Repetitive elements can lead to artifactual SV calls if not handled well. Much of the experience in identifying artifacts came from comparisons of high-coverage ($>30\times$) short-read WGS (Li et al., 2020), low-pass ($6\times$ to $8\times$) short-read WGS (Y Zhang et al., 2018), and short read–based WES (Yang et al., 2016) performed on the same samples of The Cancer Genome Atlas (TCGA) cohort. My group compared somatic SVs detected in each approach and showed that small duplications and small inversion-like events are often artifacts because they are not observed in the same samples profiled by other sequencing approaches. Detailed procedures for identification and filtering of SV artifacts in WES data were described previously (Yang et al., 2016).

Here, I offer several strategies to identify problematic samples and SVs:

1. The number of SVs detected in all samples should be inspected to identify outliers. Samples with an extremely high number of SVs require close attention. For example, in one of my group's studies of somatic SVs in 98 tumors, one sample (Fig. 4A, left) had many more somatic SVs than other samples in the same cohort. After detailed investigation using other strategies described in this section, this sample turned out to be fine. In the somatic SV study based on WES data, outlier samples were all discarded (Yang et al., 2016) due to artifactual amplification-induced small duplications.
2. SV composition should be inspected. In the same cohort shown in Figure 4A, most samples had a mixture of deletions, duplications, inversions, and translocations. However, a few samples predominantly carried deletions, which require further investigation. In fact, these deletions were artifacts (see the next paragraph for more details).
4. SV size, microhomology at breakpoints, number of supporting reads, and location of breakpoints, among other parameters, should be inspected. For the deletions in Figure 4A, when my group plotted size, microhomology at breakpoints, and number of supporting reads, we found that a large number of deletions were around 300 bp in size with a microhomology of >10 bp (Fig. 4B). Compared with other deletions, these small deletions all had very few supporting reads. If these deletions were true somatic SVs, their unique size range and microhomology suggested they were likely generated via a distinct mutational process. Few supporting reads suggested they were subclonal events that were only present in a subset of tumor cells. It is unlikely that a mutational process only operates in a subset of the cells. We then used a different SV caller and did not detect any of these deletions. Therefore, we concluded that the artifactual deletions were caused by the SV caller used initially. Similarly, the small inversion-like events identified in WGS and WES in the TCGA cohort were also <1 kb in size with large homology and few supporting reads. For germline SVs, the number of supporting reads is also a very useful measure. The allele fractions of homozygous and heterozygous SVs are 1 and 0.5, respectively. They can be calculated as the numbers of reads supporting SVs divided by read depth at the SV breakpoint junctions or as the number of discordant read pairs divided by the sum of discordant and concordant read pairs. The allele fractions of germline heterozygous SVs should follow a binomial distribution with a mean of 0.5. If the allele fractions of SVs detected in a particular sample and/or type deviate substantially from the binomial distribution, the SVs are very likely to be artifacts.

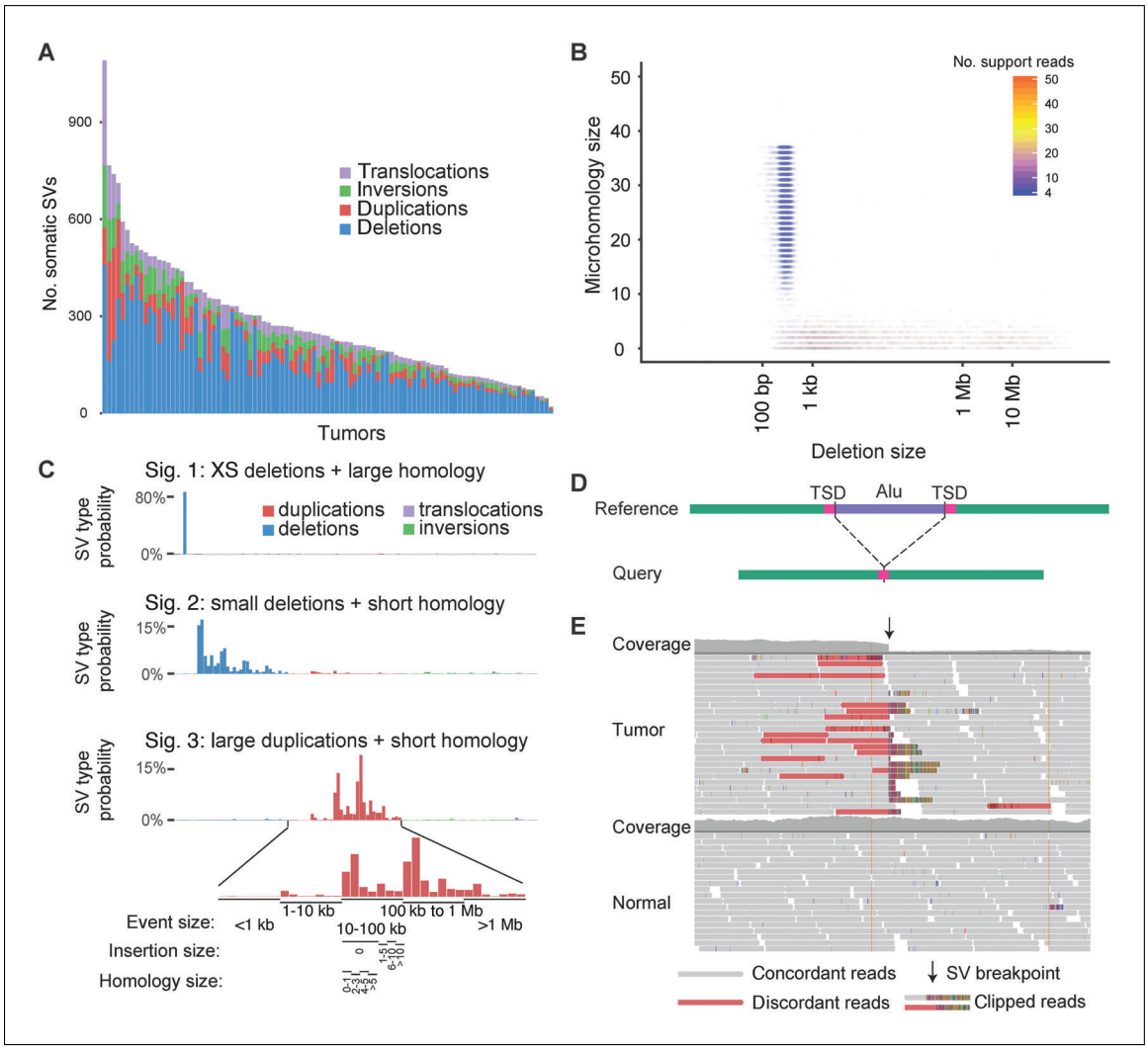

Figure 4 Quality control of structural variations (SVs) detected from sequencing data. (**A**) Numbers and composition of somatic SVs discovered in 98 tumors. (**B**) Distribution of deletion size and microhomology size color coded by number of supporting reads in 98 tumors. (**C**) Somatic SV signatures in the 98 tumors. (**D**) Schematic of germline Alu polymorphism. The Alu element shown is inserted in the reference genome but not in the query genome. TSD=target site duplication. (**E**) IGV screenshot of a true somatic SV.

4. The number and composition of SVs should be compared with previous studies of similar cohorts to identify problematic samples and SVs. For germline SVs, a recent large population-based study reported an average of 4400 germline SVs per individual (Abel et al., 2020), and the Genome Aggregation Database reported 7400 germline SVs per individual on average (Collins et al., 2020). Both studies were based on Illumina short reads. The vast majority of germline SVs were deletions and insertions, while inversions and duplications were much less common. Recent comprehensive multi-platform studies reported 13,000 SVs in a trio (Zook et al., 2020) and 27,000 SVs per germline genome (Chaisson et al., 2019). Since many germline SVs in a genome are common in a population, studies focusing on germline SVs should produce SV calls consistent with the above-mentioned large-scale comprehensive studies. For somatic SVs in cancer, the Pan-Cancer Analysis of Whole Genome reported that each tumor carries from 0 to up to 2000 somatic SVs when queried by Illumina high-coverage WGS (Campbell et al., 2020). When my group studied somatic SVs in cancer using WES data, samples with over 1000 somatic SVs immediately rang alarm (Yang et al., 2016) because WES only captures about 2% of the genome. Those samples were discarded.
5. A better approach to identify somatic SV artifacts is to perform mutational signature analysis on SVs. Somatic alterations, including SNVs and SVs, form in tumor

cells via different molecular mechanisms such as external mutagens, internal mutational processes, and defects of DNA damage repair. It is mathematically feasible to decompose the mutational signatures if a large number of tumor samples are sequenced. Methods like non-negative matrix factorization (NMF) can be used to extract such signatures (Alexandrov et al., 2013a). NMF has been applied to deconvolute signatures of somatic SNVs (Alexandrov et al., 2013b, 2020) and SVs (Li et al., 2020; Nik-Zainal et al., 2016) in cancer. If systematic artifacts are present in the variant calls, they are likely to be captured by one or more signatures (Alexandrov et al., 2020). In the cohort shown in Figure 4A, my group used signeR (Rosales, Drummond, Valieris, Dias-Neto, & da Silva, 2016) to decompose somatic SV signatures. All SVs were initially classified into deletions, duplications, inversions, and translocations. For nontranslocation events, they were then divided into five size ranges: <1 kb, 1 to 10 kb, 10 to 100 kb, 100 kb to 1 Mb, and >1 Mb. For each size range, the SVs were further divided into seven categories based on homology length and insertion sequence length at the breakpoints (Fig. 4C, bottom panel). After applying such SV classification, we obtained three signatures in the cohort. Signature 1 represented <1 kb deletions with >5 bp homology. Signatures 2 and 3 were small deletions (predominantly 1 to 10 kb in size) and large duplications (10 kb to 1 Mb in size), respectively, with short homology. If the homology and insertion sequences are not available for SVs, the SVs can be classified based on event type and size (Nik-Zainal et al., 2016). After identifying signatures, the number of supporting reads and locations of breakpoints should be inspected to identify possible artifact signatures as described previously. For example, Signature 1 corresponded to the deletion artifacts shown in Figure 4B. In another study of a different cohort of patients with cancer, my group also identified a somatic small deletion signature with large homology. The deletions were also around 300 bp; however, the homology at breakpoints was between 10 and 15 bp. The breakpoints of these deletions were all at the boundaries of Alu elements in the reference genome. These deletions were in fact germline Alu polymorphisms. The Alu insertions in the reference genome were not present in the query genome, so they appeared as deletions in the query genomes when compared with the standard reference genome (Fig. 4D). The 10 to 15 bp homology reflected the target site duplication of Alu insertions. In this particular study, the germline Alu polymorphisms were not properly filtered and remained in the somatic SV call set. In general, if the sequencing quality of matched normal tissue is poor or if the read depth is not high enough, germline SVs may be unfiltered. Therefore, in the SV caller previously developed (i.e., Meerkat; Yang et al., 2013), SVs detected in each tumor were filtered against all normal samples merged together so that the common germline variants could be filtered even if one or a few normal samples were of poor quality.

6. Read alignment should always be inspected in IGV (Robinson et al., 2011) or other genome browsers for a random subset of samples and SVs, especially suspicious samples and SVs. For short read–based sequencing, true SVs should present clear support of discordant read pairs, split reads, and sometimes changes in coverage (Fig. 4E). Note that read depth would not change for balanced SVs, such as inversions and balanced translocations. Users may need to change the insert size range setting to properly display discordant read pairs in IGV. The insert size range can be set as either percentiles (e.g., >99.9% and <0.1%) or actual base pair length (e.g., >800 bp and <200 bp). For somatic SVs, one should load the tumor and matched normal genomes at the same time. Somatic SVs should be well supported in the tumor genome but not in the normal genome (Fig. 4E). If similar supporting reads are present in the normal genome, the SV is probably a germline event. If no similar support reads are found in the normal genome but the region has many other discordant read pairs, clipped reads, and nonunique mapped reads (reads shown as white boxes), this suggests that the read alignment of this region is problematic. Somatic SVs called from this region may not be of high confidence.

CONCLUSIONS

SVs are important genetic variants to study population diversity and human diseases. Detection of SVs is particularly challenging compared with other types of variants. Various technologies and platforms are available to

interrogate SVs in the human genome. These SV discovery technologies have their own pros and cons. Most of the time, to a specific research project, the major limitation would be budget. Researchers should decide on their strategies based on their overall goal and cost. After computational detection of SVs, researchers should perform rigorous quality control to eliminate possible artifacts caused by sample quality, sequencing and imaging platforms, and computational tools.

ACKNOWLEDGMENTS

This work was supported by grant R03CA246228-01 from the National Cancer Institute.

AUTHOR CONTRIBUTIONS

Lixing Yang: Conceptualization; funding acquisition; investigation; methodology; project administration; supervision; visualization; writing-original draft.

LITERATURE CITED

Abel, H. J., Larson, D. E., Regier, A. A., Chiang, C., Das, I., Kanchi, K. L., … Hall, I. M. (2020). Mapping and characterization of structural variation in 17,795 human genomes. *Nature*, *583*, 83–89. doi: 10.1038/s41586-020-2371-0.

Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, *21*, 974–984. doi: 10.1101/gr.114876.110.

Alaei-Mahabadi, B., Bhadury, J., Karlsson, J. W., Nilsson, J. A., & Larsson, E. (2016). Global analysis of somatic structural genomic alterations and their impact on gene expression in diverse human cancers. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 13768–13773. doi: 10.1073/pnas.1606220113.

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., … Stratton, M. R. (2020). The repertoire of mutational signatures in human cancer. *Nature*, *578*, 94–101. doi: 10.1038/s41586-020-1943-3.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., … Stratton, M. R. (2013a). Signatures of mutational processes in human cancer. *Nature*, *500*, 415–421. doi: 10.1038/nature12477.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., Stratton, M. R., Leary, R. J., Shen, D., Boca, S. M., Barber, T., Ptak, J., et al. (2013b). Deciphering signatures of mutational processes operative in human cancer. *Cell Reports*, *3*, 246–259. doi: 10.1016/j.celrep.2012.12.008.

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, *12*, 363–376. doi: 10.1038/nrg2958.

Andrews, S. (2010). FastQC a quality control tool for high throughput sequence data. Retrieved from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. M. E., … Eichler, E. E. (2019). Characterizing the major structural variant alleles of the human genome. *Cell*, *176*, 663–675. e19. doi: 10.1016/j.cell.2018.12.019.

Baca, S. C., Prandi, D., Lawrence, M. S., Mosquera, J. M., Romanel, A., Drier, Y., … Ghandi, M. (2013). Punctuated evolution of prostate cancer genomes. *Cell*, *153*, 666–677. doi: 10.1016/j.cell.2013.03.021.

Becker, T., Lee, W. P., Leone, J., Zhu, Q., Zhang, C., Liu, S., … Malhotra, A. (2018). FusorSV: An algorithm for optimally combining data from multiple structural variation detection methods. *Genome Biology*, *19*, 38. doi: 10.1186/s13059-018-1404-6.

Blokzijl, F., de Ligt, J., Jager, M., Sasselli, V., Roerink, S., Sasaki, N., … Van Boxtel, R. (2016). Tissue-specific mutation accumulation in human adult stem cells during life. *Nature*, *538*, 260–264. doi: 10.1038/nature19768.

Bochukova, E. G., Huang, N., Keogh, J., Henning, E., Purmann, C., Blaszczyk, K., … Farooqi, I. S. (2010). Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature*, *463*, 666–670. doi: 10.1038/nature08689.

Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., … Barillot, E. (2012). Control-FREEC: A tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, *28*, 423–425. doi: 10.1093/bioinformatics/btr670.

Boxer, L. M., & Dang, C. V. (2001). Translocations involving c-myc and c-myc function. *Oncogene*, *20*, 5595–5610. doi: 10.1038/sj.onc.1204595.

Brandler, W. M., Antaki, D., Gujral, M., Kleiber, M. L., Whitney, J., Maile, M. S., … Sebat, J. (2018). Paternally inherited cis-regulatory structural variants are associated with autism. *Science*, *360*, 327–331. doi: 10.1126/science.aan2261.

Bumgarner, R. (2013). Overview of DNA microarrays: Types, applications, and their future. *Current Protocols in Molecular Biology*, *101*, 22.1.1–22.1.11. doi: 10.1002/0471142727.mb2201s101.

Bunting, S. F., & Nussenzweig, A. (2013). End-joining, translocations and cancer. *Nature Reviews Cancer*, *13*, 443–454. doi: 10.1038/nrc3537.

Cameron, D. L., Di Stefano, L., & Papenfuss, A. T. (2019). Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nature Communications*, *10*, 3240. doi: 10.1038/s41467-019-11146-4.

Campbell, P. J., Getz, G., Korbel, J. O., Stuart, J. M., Jennings, J. L., Stein, L. D., … Zhang, J. (2020). Pan-cancer analysis of whole genomes. *Nature*, *578*, 82–93. doi: 10.1038/s41586-020-1969-6.

Carvalho, C. M. B., Ramocki, M. B., Pehlivan, D., Franco, L. M., Gonzaga-Jauregui, C., Fang, P., … Lupski, J. R. (2011). Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nature Genetics*, *43*, 1074–1081. doi: 10.1038/ng.944.

Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., … Eichler, E. E. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, *517*, 608–611. doi: 10.1038/nature13907.

Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., … Lee, C. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, *10*, 1–16. doi: 10.1038/s41467-018-08148-z.

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., … Saunders, C. T. (2016). Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, *32*, 1220–1222. doi: 10.1093/bioinformatics/btv710.

Chiang, C., Jacobsen, J. C., Ernst, C., Hanscom, C., Heilbut, A., Blumenthal, I., … Talkowski, M. E. (2012). Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nature Genetics*, *44*, 390–397. doi: 10.1038/ng.2202.

Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., Kim, Y., … Hall, I. M. (2017). The impact of structural variation on human gene expression. *Nature Genetics*, *49*, 692–699. doi: 10.1038/ng.3834.

Chong, Z., Ruan, J., Gao, M., Zhou, W., Chen, T., Fan, X., … Chen, K. (2016). NovoBreak: Local assembly for breakpoint detection in cancer genomes. *Nature Methods*, *14*, 65–67. doi: 10.1038/nmeth.4084.

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., … Talkowski, M. E. (2020). A structural variation reference for medical and population genetics. *Nature*, *581*, 444–451. doi: 10.1038/s41586-020-2287-8.

Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E., & Pritchard, J. K. (2005). A high-resolution survey of deletion polymorphism in the human genome. *Nature Genetics*, *38*, 75–81. doi: 10.1038/ng1697.

Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, *10*, 691–703. doi: 10.1038/nrg2640.

Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., Robson, S., … Donnelly, P. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, *464*, 713–720. doi: 10.1038/nature08979.

Cretu Stancu, M., Van Roosmalen, M. J., Renkens, I., Nieboer, M. M., Middelkamp, S., De Ligt, J., … Kloosterman, W. P. (2017). Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nature Communications*, *8*, 1326. doi: 10.1038/s41467-017-01343-4.

Croft, B., Ohnesorg, T., Hewitt, J., Bowles, J., Quinn, A., Tan, J., … Sinclair, A. (2018). Human sex reversal is caused by duplication or deletion of core enhancers upstream of SOX9. *Nature Communications*, *9*, 5319. doi: 10.1038/s41467-018-07784-9.

Davis, C. F., Ricketts, C. J., Wang, M., Yang, L., Cherniack, A. D., Shen, H., … Fahey, C. C. (2014). The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell*, *26*, 319–330. doi: 10.1016/j.ccr.2014.07.014.

De Coster, W., De Rijk, P., De Roeck, A., De Pooter, T., D'Hert, S., Strazisar, M., … Van Broeckhoven, C. (2019). Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome Research*, *29*, 1178–1187. doi: 10.1101/gr.244939.118.

Dixon, J. R., Xu, J., Dileep, V., Zhan, Y., Song, F., Le, V. T., … Yue, F. (2018). Integrative detection and analysis of structural variation in cancer genomes. *Nature Genetics*, *50*, 1388–1398. doi: 10.1038/s41588-018-0195-8.

Fan, X., Chaisson, M., Nakhleh, L., & Chen, K. (2017). HySA: A hybrid structural variant assembly approach using next-generation and single-molecule sequencing technologies. *Genome Research*, *27*, 793–800. doi: 10.1101/gr.214767.116.

Fang, L., Kao, C., Gonzalez, M. V., Mafra, F. A., Pellegrino da Silva, R., Li, M., … Wang, K. (2019). LinkedSV for detection of mosaic structural variants from linked-read exome and genome sequencing data. *Nature Communications*, *10*, 5585. doi: 10.1038/s41467-019-13397-7.

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews Genetics*, *7*, 85–97. doi: 10.1038/nrg1767.

Franke, M., Ibrahim, D. M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., … Mundlos, S. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, *538*, 265–269. doi: 10.1038/nature19800.

Georgieva, L., Rees, E., Moran, J. L., Chambert, K. D., Milanova, V., Craddock, N., … Kirov, G. (2014). De novo CNVs in bipolar affective disorder and schizophrenia. *Human Molecular Genetics*, *23*, 6677–6683. doi: 10.1093/hmg/ddu379.

Glessner, J. T., Wang, K., Cai, G., Korvatska, O., Kim, C. E., Wood, S., … Hakonarson, H. (2009). Autism genome-wide copy number

variation reveals ubiquitin and neuronal genes. *Nature*, *459*, 569–573. doi: 10.1038/nature07953.

Gong, L., Wong, C. H., Cheng, W. C., Tjong, H., Menghi, F., Ngan, C. Y., … Wei, C. L. (2018). Picky comprehensively detects high-resolution structural variants in nanopore long reads. *Nature Methods*, *15*, 455–460. doi: 10.1038/s41592-018-0002-6.

Greenman, C. D., Pleasance, E. D., Newman, S., Yang, F., Fu, B., Nik-Zainal, S., … Campbell, P. J. (2011). Estimation of rearrangement phylogeny for cancer genomes. *Genome Research*, *22*, 346–361. doi: 10.1101/gr.118414.110.

Groffen, J., Stephenson, J. R., Heisterkamp, N., de Klein, A., Bartram, C. R., & Grosveld, G. (1984). Philadelphia chromosomal breakpoints are clustered within a limited region, bcr, on chromosome 22. *Cell*, *36*, 93–99. doi: 10.1016/0092-8674(84)90077-1.

Gröschel, S., Sanders, M. A., Hoogenboezem, R., de Wit, E., Bouwman, B. A. M., Erpelinck, C., … Van Lom, K. (2014). A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell*, *157*, 369–381. doi: 10.1016/j.cell.2014.02.019.

Handsaker, R. E., Korn, J. M., Nemesh, J., & McCarroll, S. A. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature Genetics*, *43*, 269–276. doi: 10.1038/ng.768.

Hastings, P. J., Lupski, J. R., Rosenberg, S. M., & Ira, G. (2009). Mechanisms of change in gene copy number. *Nature Reviews Genetics*, *10*, 551–564. doi: 10.1038/nrg2593.

Hehir-Kwa, J. Y., Pfundt, R., & Veltman, J. A. (2015). Exome sequencing and whole genome sequencing for the detection of copy number variation. *Expert Review of Molecular Diagnostics*, *15*, 1023–1032. doi: 10.1586/14737159.2015.1053467.

Heller, M. J. (2002). DNA microarray technology: Devices, systems, and applications. *Annual Review of Biomedical Engineering*, *4*, 129–153. doi: 10.1146/annurev.bioeng.4.020702.153438.

Holland, A. J., & Cleveland, D. W. (2012). Chromoanagenesis and cancer: Mechanisms and consequences of localized, complex chromosomal rearrangements. *Nature Medicine*, *18*, 1630–1638. doi: 10.1038/nm.2988.

Kadalayil, L., Rafiq, S., Rose-Zerilli, M. J. J., Pengelly, R. J., Parker, H., Oscier, D., … Collins, A. (2014). Exome sequence read depth methods for identifying copy number changes. *Briefings in Bioinformatics*, *16*, 380–392. doi: 10.1093/bib/bbu027.

Kidd, J. M., Graves, T., Newman, T. L., Fulton, R., Hayden, H. S., Malig, M., … Eichler, E. E. (2010). A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell*, *143*, 837–847. doi: 10.1016/j.cell.2010.10.027.

Kishikawa, T., Momozawa, Y., Ozeki, T., Mushiroda, T., Inohara, H., Kamatani, Y., … Okada, Y. (2019). Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data. *Scientific Reports*, *9*, 1784. doi: 10.1038/s41598-018-38346-0.

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., & Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology*, *20*, 117. doi: 10.1186/s13059-019-1720-5.

Lee, A. Y., Ewing, A. D., Ellrott, K., Hu, Y., Houlahan, K. E., Bare, J. C., … Boutros, P. C. (2018). Combining accurate tumor genome simulation with crowdsourcing to benchmark somatic structural variant detection. *Genome Biology*, *19*, 188. doi: 10.1186/s13059-018-1539-5.

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L. J., … Park, P. J. (2012). Landscape of somatic retrotransposition in human cancers. *Science*, *337*, 967–971. doi: 10.1126/science.1222077.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. [preprint]. Retrieved from http://arxiv.org/abs/1303.3997.

Li, Y., Roberts, N. D., Wala, J. A., Shapira, O., Schumacher, S. E., Kumar, K., … Campbell, P. J. (2020). Patterns of somatic structural variation in human cancer genomes. *Nature*, *578*, 112–121. doi: 10.1038/s41586-019-1913-9.

Li, Y., Schwab, C., Ryan, S. L., Papaemmanuil, E., Robinson, H. M., Jacobs, P., … Harrison, C. J. (2014). Constitutional and somatic rearrangement of chromosome 21 in acute lymphoblastic leukaemia. *Nature*, *508*, 98–102. doi: 10.1038/nature13115.

Lin, C. F., Naj, A. C., & Wang, L. S. (2013). Analyzing copy number variation using SNP array data: Protocols for calling CNV and association tests. *Current Protocols in Human Genetics*, *79*, 1.27.1–1.27.15. doi: 10.1002/0471142905.hg0127s79.

Liu, P., Erez, A., Nagamani, S. C. S., Dhar, S. U., Kołodziejska, K. E., Dharmadhikari, A. V., … Bi, W. (2011). Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell*, *146*, 889–903. doi: 10.1016/j.cell.2011.07.042.

Lockwood, W. W., Chari, R., Chi, B., & Lam, W. L. (2006). Recent advances in array comparative genomic hybridization technologies and their applications in human genetics. *European Journal of Human Genetics*, *14*, 139–148. doi: 10.1038/sj.ejhg.5201531.

Lodato, M. A., Rodin, R. E., Bohrson, C. L., Coulter, M. E., Barton, A. R., Kwon, M., … Walsh, C. A. (2018). Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science*, *359*, 555–559. doi: 10.1126/science.aao4426.

Luan, M.-W., Zhang, X.-M., Zhu, Z.-B., Chen, Y., & Xie, S.-Q. (2020). Evaluating structural variation detection tools for long-read sequencing datasets in saccharomyces cerevisiae. *Frontiers*

in Genetics, *11*, 159. doi: 10.3389/fgene.2020.00159.

Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., … Mundlos, S. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, *161*, 1012–1025. doi: 10.1016/j.cell.2015.04.004.

Maciejowski, J., Li, Y., Bosco, N., Campbell, P. J., & de Lange, T. (2015). Chromothripsis and kataegis induced by telomere crisis. *Cell*, *163*, 1641–1654. doi: 10.1016/j.cell.2015.11.054.

Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., … Campbell, P. J. (2015). High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, *348*, 880–886. doi: 10.1126/science.aaa6806.

McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., … Altshuler, D. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics*, *40*, 1166–1174. doi: 10.1038/ng.238.

Mertens, F., Johansson, B., Fioretos, T., & Mitelman, F. (2015). The emerging complexity of gene fusions in cancer. *Nature Reviews Cancer*, *15*, 371–381. doi: 10.1038/nrc3947.

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., … Korbel, J. O. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature*, *470*, 59–65. doi: 10.1038/nature09708.

Molenaar, J. J., Koster, J., Zwijnenburg, D. A., Van Sluis, P., Valentijn, L. J., Van der Ploeg, I., … Versteeg, R. (2012). Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature*, *483*, 589–593. doi: 10.1038/nature10910.

Mullaney, J. M., Mills, R. E., Pittard, W. S., & Devine, S. E. (2010). Small insertions and deletions (INDELs) in human genomes. *Human Molecular Genetics*, *19*, R131–R136. doi: 10.1093/hmg/ddq400.

Muzny, D. M., Bainbridge, M. N., Chang, K., Dinh, H. H., Drummond, J. A., Fowler, G., … Gibbs, R. A. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, *487*, 330–337. doi: 10.1038/nature11252.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., … Stratton, M. R. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, *534*, 47–54. doi: 10.1038/nature17676.

Northcott, P. A., Lee, C., Zichner, T., Stütz, A. M., Erkek, S., Kawauchi, D., … Sturm, D. (2014). Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature*, *511*, 428–434. doi: 10.1038/nature13379.

Pinkel, D., & Albertson, D. G. (2005). Comparative genomic hybridization. *Annual Review of Genomics and Human Genetics*, *6*, 331–354. doi: 10.1146/annurev.genom.6.080604.162140.

Pirooznia, M., Goes, F., & Zandi, P. P. (2015). Whole-genome CNV analysis: Advances in computational approaches. *Frontiers in Genetics*, *6*, 138. doi: 10.3389/fgene.2015.00138.

Rausch, T., Jones, D. T. W., Zapatka, M., Stütz, A. M., Zichner, T., Weischenfeldt, J., … Korbel, J. O. (2012a). Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell*, *148*, 59–71. doi: 10.1016/j.cell.2011.12.013.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012b). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*, i333–i339. doi: 10.1093/bioinformatics/bts378.

Redin, C., Brand, H., Collins, R. L., Kammin, T., Mitchell, E., Hodge, J. C., … Talkowski, M. E. (2017). The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nature Genetics*, *49*, 36–45. doi: 10.1038/ng.3720.

Ritz, A., Bashir, A., Sindi, S., Hsu, D., Hajirasouliha, I., & Raphael, B. J. (2014). Characterization of structural variants with single molecule and hybrid sequencing approaches. *Bioinformatics*, *30*, 3458–3466. doi: 10.1093/bioinformatics/btu714.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*, 24–26. doi: 10.1038/nbt.1754.

Rosales, R. A., Drummond, R. D., Valieris, R., Dias-Neto, E., & da Silva, I. T. (2016). signeR: An empirical Bayesian approach to mutational signature discovery. *Bioinformatics*, *33*, 8–16. doi: 10.1093/bioinformatics/btw572.

Ruderfer, D. M., Hamamsy, T., Lek, M., Karczewski, K. J., Kavanagh, D., Samocha, K. E., … Purcell, S. M. (2016). Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nature Genetics*, *48*, 1107–1111. doi: 10.1038/ng.3638.

Schaaf, C. P., Wiszniewska, J., & Beaudet, A. L. (2011). Copy number and SNP arrays in clinical diagnostics. *Annual Review of Genomics and Human Genetics*, *12*, 25–51. doi: 10.1146/annurev-genom-092010-110715.

Schmitt, M. W., Loeb, L. A., & Salk, J. J. (2016). The influence of subclonal resistance mutations on targeted cancer therapy. *Nature Reviews Clinical Oncology*, *13*, 335–347. doi: 10.1038/nrclinonc.2015.175.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., … Wigler, M. (2007). Strong association of de novo copy number mutations with autism. *Science*, *316*, 445–449. doi: 10.1126/science.1138659.

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, *15*, 461–468. doi: 10.1038/s41592-018-0001-7.

Spies, N., Weng, Z., Bishara, A., McDaniel, J., Catoe, D., Zook, J. M., … Sidow, A. (2017). Genome-wide reconstruction of complex structural variants using read clouds. *Nature Methods*, *14*, 915–920. doi: 10.1038/nmeth.4366.

Stankiewicz, P., & Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, *61*, 437–455. doi: 10.1146/annurev-med-100708-204735.

Stephens, P. J., Greenman, C. D., Fu, B., Yang, F., Bignell, G. R., Mudie, L. J., … Campbell, P. J. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, *144*, 27–40. doi: 10.1016/j.cell.2010.11.055.

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazlsy, C., Thorne, N., … Dermitzakis, E. T. (2007). Relative impact of nucleotide and copy number variation on gene phenotypes. *Science*, *315*, 848–853. doi: 10.1126/science.1136678.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*, 75–81. doi: 10.1038/nature15394.

Takeda, D. Y., Spisák, S., Seo, J. H., Bell, C., O'Connor, E., Korthauer, K., … Freedman, M. L. (2018). A somatically acquired enhancer of the androgen receptor is a noncoding driver in advanced prostate cancer. *Cell*, *174*, 422–432.e13. doi: 10.1016/j.cell.2018.05.037.

Tang, Y. C., & Amon, A. (2013). Gene copy-number alterations: A cost-benefit analysis. *Cell*, *152*, 394–405. doi: 10.1016/j.cell.2012.11.043.

Trost, B., Walker, S., Wang, Z., Thiruvahindrapuram, B., MacDonald, J. R., Sung, W. W. L., … Scherer, S. W. (2018). A comprehensive workflow for read depth-based identification of copy-number variation from whole-genome sequence data. *American Journal of Human Genetics*, *102*, 142–155. doi: 10.1016/j.ajhg.2017.12.007.

Valentijn, L. J., Koster, J., Zwijnenburg, D. A., Hasselt, N. E., Van Sluis, P., Volckmann, R., … Molenaar, J. J. (2015). TERT rearrangements are frequent in neuroblastoma and identify aggressive tumors. *Nature Genetics*, *47*, 1411–1414. doi: 10.1038/ng.3438.

Viswanathan, S. R., Ha, G., Hoff, A. M., Wala, J. A., Carrot-Zhang, J., Whelan, C. W., … Meyerson, M. (2018). Structural alterations driving castration-resistant prostate cancer revealed by linked-read genome sequencing. *Cell*, *174*, 433–447.e19. doi: 10.1016/j.cell.2018.05.036.

Wan, T. S. K. (2014). Cancer cytogenetics: Methodology revisited. *Annals of Laboratory Medicine*, *34*, 413–425. doi: 10.3343/alm.2014.34.6.413.

Weischenfeldt, J., Symmons, O., Spitz, F., & Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: Insights from and for human disease. *Nature Reviews Genetics*, *14*, 125–138. doi: 10.1038/nrg3373.

Wu, S., Turner, K. M., Nguyen, N., Raviram, R., Erb, M., Santini, J., … Mischel, P. S. (2019). Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature*, *575*, 699–703. doi: 10.1038/s41586-019-1763-5.

Xi, R., Hadjipanayis, A. G., Luquette, L. J., Kim, T.-M., Lee, E., Zhang, J., … Park, P. J. (2011). Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, E1128–E1136. doi: 10.1073/pnas.1110574108.

Xu, C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal*, *16*, 15–24. doi: 10.1016/j.csbj.2018.01.003.

Yang, L., Lee, M.-S., Lu, H., Oh, D.-Y., Kim, Y. J., Park, D., … Haseley, P. S. (2016). Analyzing somatic genome rearrangements in human cancers by using whole-exome sequencing. *American Journal of Human Genetics*, *98*, 843–856. doi: 10.1016/j.ajhg.2016.03.017.

Yang, L., Luquette, L. J., Gehlenborg, N., Xi, R., Haseley, P. S., Hsieh, C.-H., … Park, P. J. (2013). Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell*, *153*, 919–929. doi: 10.1016/j.cell.2013.04.010.

Yang, L., Wang, S., Lee, J. J.-K., Lee, S., Lee, E., Shinbrot, E., … Park, P. J. (2019). An enhanced genetic model of colorectal cancer progression history. *Genome Biology*, *20*, 168. doi: 10.1186/s13059-019-1782-4.

Yao, R., Yu, T., Qing, Y., Wang, J., & Shen, Y. (2019). Evaluation of copy number variant detection from panel-based next-generation sequencing data. *Molecular Genetics and Genomic Medicine*, *7*, e00513. doi: 10.1002/mgg3.513.

Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., … Mermel, C. H. (2013). Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, *45*, 1134–1140. doi: 10.1038/ng.2760.

Zarate, S., Carroll, A., Krashenina, O., Sedlazeck, F. J., Jun, G., Salerno, W., … Gibbs, R. (2018). Parliament2: Fast structural variant calling using optimized combinations of callers. *bioRxiv*, 424267, [preprint]. Retrieved from https://www.biorxiv.org/content/10.1101/424267v1.

Zare, F., Dow, M., Monteleone, N., Hosny, A., & Nabavi, S. (2017). An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinformatics*, *18*, 286. doi: 10.1186/s12859-017-1705-x.

Zhang, C.-Z., Spektor, A., Cornils, H., Francis, J. M., Jackson, E. K., Liu, S., … Pellman, D. (2015). Chromothripsis from DNA damage in micronuclei. *Nature*, *522*, 179–184. doi: 10.1038/nature14493.

Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009). Copy number variation in human

health, disease, and evolution. *Annual Review of Genomics and Human Genetics*, *10*, 451–481. doi: 10.1146/annurev.genom.9.081307.164217.

Zhang, L., Bai, W., Yuan, N., & Du, Z. (2019). Comprehensively benchmarking applications for detecting copy number variation. *PLoS Computational Biology*, *15*, e1007069. doi: 10.1371/journal.pcbi.1007069.

Zhang, X., Choi, P. S., Francis, J. M., Gao, G. F., Campbell, J. D., Ramachandran, A., … Meyerson, M. (2018). Somatic superenhancer duplications and hotspot mutations lead to oncogenic activation of the KLF5 transcription factor. *Cancer Discovery*, *8*, 108–125. doi: 10.1158/2159-8290.CD-17-0532.

Zhang, X., Choi, P. S., Francis, J. M., Imielinski, M., Watanabe, H., Cherniack, A. D., & Meyerson, M. (2015). Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. *Nature Genetics*, *48*, 176–182. doi: 10.1038/ng.3470.

Zhang, Y., Yang, L., Kucherlapati, M., Chen, F., Hadjipanayis, A., Pantazi, A., … Creighton, C. J. (2018). A pan-cancer compendium of genes deregulated by somatic genomic rearrangement across more than 1,400 cases. *Cell Reports*, *24*, 515–527. doi: 10.1016/j.celrep.2018.06.025.

Zhao, M., Wang, Q., Wang, Q., Jia, P., & Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: Features and perspectives. *BMC Bioinformatics*, *14*, S1. doi: 10.1186/1471-2105-14-S11-S1.

Zook, J. M., Hansen, N. F., Olson, N. D., Chapman, L. M., Mullikin, J. C., Xiao, C., … Salit, M. (2020). A robust benchmark for detection of germline large deletions and insertions. *Nature Biotechnology*, [Epub ahead of print]. doi: 10.1038/s41587-020-0538-8.