

Toward the functional interpretation of somatic structural variations: bulk- and single-cell approaches

Dohun Yi, Jin-Wu Nam  and Hyobin Jeong 

Corresponding authors: Jin-Wu Nam, Department of Life Science, College of Natural Sciences, Hanyang University, Wangsimni-ro 222, Seongdong-gu, Seoul 04763, Republic of Korea. Tel.: +82-2-2220-2428; Fax: +82-2-2298-0319. E-mail: jwnam@hanyang.ac.kr; Hyobin Jeong, Department of Life Science, College of Natural Sciences, Hanyang University, Wangsimni-ro 222, Seongdong-gu, Seoul 04763, Republic of Korea. Tel.: +82-2-2220-2494; E-mail: hyobinjeong@hanyang.ac.kr

Abstract

Structural variants (SVs) are genomic rearrangements that can take many different forms such as copy number alterations, inversions and translocations. During cell development and aging, somatic SVs accumulate in the genome with potentially neutral, deleterious or pathological effects. Generation of somatic SVs is a key mutational process in cancer development and progression. Despite their importance, the detection of somatic SVs is challenging, making them less studied than somatic single-nucleotide variants. In this review, we summarize recent advances in whole-genome sequencing (WGS)-based approaches for detecting somatic SVs at the tissue and single-cell levels and discuss their advantages and limitations. First, we describe the state-of-the-art computational algorithms for somatic SV calling using bulk WGS data and compare the performance of somatic SV detectors in the presence or absence of a matched-normal control. We then discuss the unique features of cutting-edge single-cell-based techniques for analyzing somatic SVs. The advantages and disadvantages of bulk and single-cell approaches are highlighted, along with a discussion of their sensitivity to copy-neutral SVs, usefulness for functional inferences and experimental and computational costs. Finally, computational approaches for linking somatic SVs to their functional readouts, such as those obtained from single-cell transcriptome and epigenome analyses, are illustrated, with a discussion of the promise of these approaches in health and diseases.

Keywords: somatic structural variation, whole-genome sequencing, single-cell whole-genome sequencing

INTRODUCTION

In the last few decades, population-level genomic studies have uncovered genomic variants that contribute to pathological phenotypes as well as disease development and progression. Conventionally, these studies took advantage of single-nucleotide polymorphism arrays to profile known genomic variants, but whole-genome sequencing (WGS) was later used to detect both known and unknown genetic variants in an unbiased manner. Initially, considerable effort was devoted to identifying single-nucleotide variants (SNVs); however, recent studies have highlighted the importance of large (>50 bp) genomic rearrangements called structural variants (SVs), which involve the deletion, amplification or reordering of genomic segments. SV detection often involves searching for breakpoints, that is, the chromosomal position at which a DNA break occurs. SVs have often been overlooked in disease studies and clinical decision-making because of the difficulties in their detection, although they tend to have a greater effect on gene expression than SNVs or indels [1, 2]. Although SVs occur much less frequently than SNVs, they are often involved in the loss- or gain-of-function of cancer-related genes and therapeutic targets in cancer genomes [3].

The genetic pathology of complex diseases is often explained by germline variations inherited from parents and somatic mutations acquired over a lifespan (Figure 1). Germline and somatic variants can be distinguished by their allele frequencies because germline variants are shared by all cells in an individual, whereas somatic variants are shared by only a subset of cells, owing to normal admixture (known as tumor purity), tumor heterogeneity and somatic mosaicism. Therefore, somatic events are often escaped from the radar of detection algorithms.

Detecting somatic SVs is difficult for other reasons. Somatic alterations are sometimes complexed with other rearrangements via catastrophic genomic events such as chromothripsis and chromoplexy (Figure 1). Chromothripsis is a cluster of many SVs that occur in a single catastrophic event, with an oscillating copy number profile and localized clustering of tens to hundreds of breakpoints. Chromoplexy is a set of reciprocal SVs that results from several simultaneous double-stranded DNA breaks in several chromosomes that are incorrectly rejoined, leading to a balanced chain of rearrangements [4]. Hence, resolving such complex rearrangements resembles untying terrible entangled paths, which is not a trivial problem, because SV callers generally take

Dohun Yi is a PhD candidate at the Department of Life Science at Hanyang University and is working with computational methods to detect structural variations and RNA modifications.

Jin-Wu Nam is a professor at Hanyang University. He received a PhD in bioinformatics from Seoul National University and studied computational biology at the Whitehead Institute for Biomedical Research.

Hyobin Jeong is an associate professor at Hanyang University. She received a PhD in systems biology from the Pohang University of Science and Technology and studied single-cell multiomics at the European Molecular Biology Laboratory, Heidelberg.

Received: May 10, 2023. **Revised:** July 5, 2023. **Accepted:** July 23, 2023

© The Author(s) 2023. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

advantage of clipped and/or discordant reads as well as read depth to detect SVs [5]. Pan-Cancer Analysis of Whole Genomes (PCAWG) has shown that a considerable number of unclassified complex SVs are present across different tumor types [4]. In addition, different somatic rearrangements in a locus may result from the clonal heterogeneity of bulk tumors. In this situation, very few informative genomic reads support the breakpoints of each rearrangement, often resulting in unresolved SVs.

However, the precise prediction of somatic SVs usually requires both tumor and matched-normal samples to reduce false-positive calls of germline SVs. However, matched-normal samples are often unavailable in clinical situations because of requirements for Internal Review Board approval, need for patient consent and economic issues, among other reasons [6]. Even when false-positives are removed using cataloged germline variations, the prediction of personal somatic variations without matched-normal samples produces many additional false-positives [7].

For these reasons, the high-confidence detection of somatic SVs remains challenging. To overcome these difficulties, large computational efforts have recently been made to detect somatic SVs in tissues using conventional bulk-cell WGS technology. Additionally, new experimental and computational developments have allowed somatic SVs to be detected or inferred using single-cell genomic technologies.

In this review, we provide an overview of the recent advances in the detection of somatic SVs in both tissue and single-cell samples. We first discuss the computational methods used for

the analysis of bulk samples, highlighting challenges related to tumor heterogeneity and the lack of matched-normal samples, as well as efforts to reduce costs and overcome limitations. Next, we discuss the newly developed single-cell-based experimental and computational techniques and compare them with the bulk sample-based techniques. Finally, we provide new insights into single-cell multiomics studies and perspectives on their future contributions to the field.

ADVANCES IN SOMATIC SV DETECTION METHODS IN BULK GENOME SEQUENCING DATA

A typical method for identifying SVs is to search for abnormalities when sequencing reads are mapped to a reference genome. Such abnormalities in the mapped reads include discordant and clip-/split-reads. Discordant reads occur when a pair of reads is mapped in a non-canonical orientation or when the insert size between them is unusually long or short. This situation arises when the breakpoint of the SV exists between a pair of reads, making the sequence generated by the variation not necessarily visible in the read.

In contrast, clip-/split-reads are identified when there is at least one SV breakpoint within the read sequence. If only a portion of the read is mapped, whereas the rest is not, it is called a clip-read, and the unmapped part is called clipped. However, if the remaining portion of the read is mapped to different locations

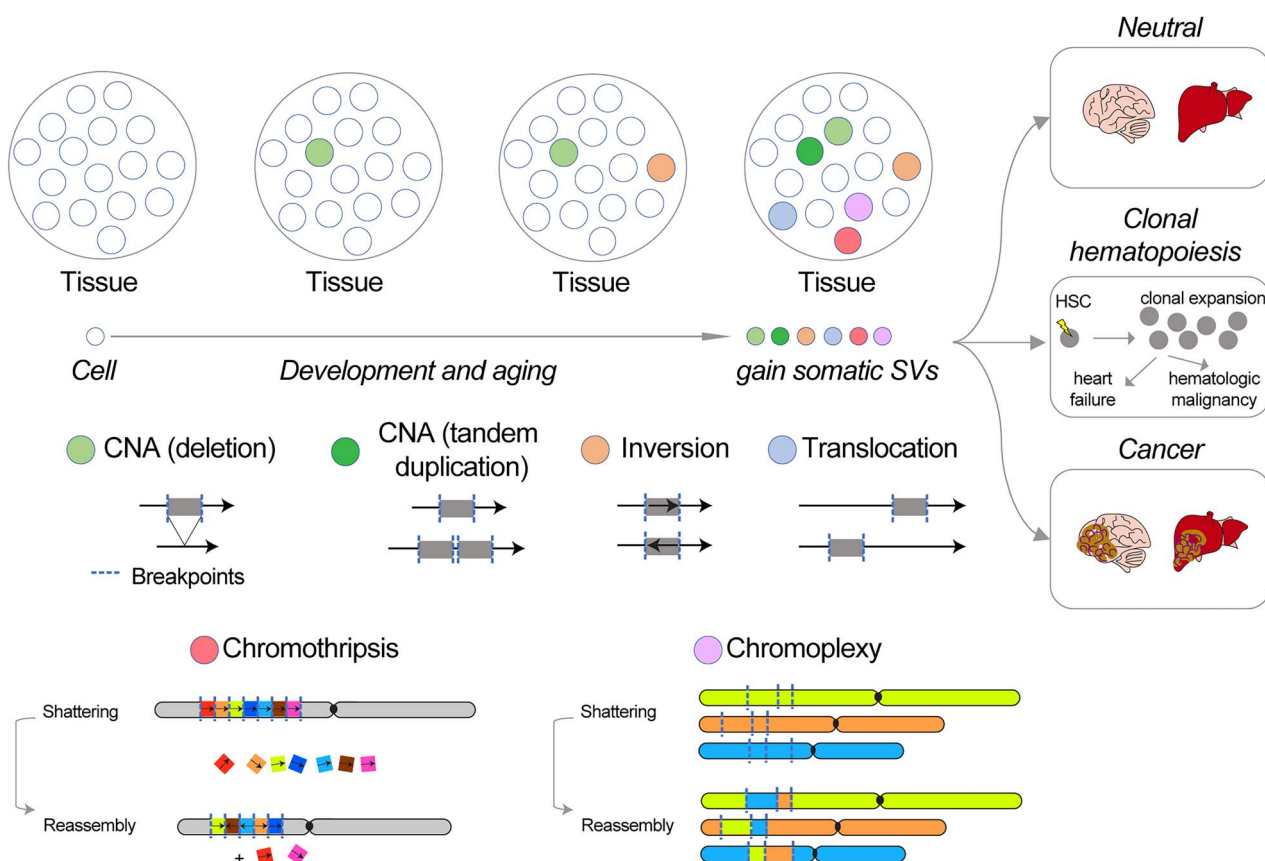


Figure 1. Accumulation of somatic SVs in cellular systems. SVs are genetic rearrangements that include CNAs, inversions, translocations, and complex events such as chromothripsis and chromoplexy. The dashed lines represent the breakpoints of SVs. Unlike germline SVs shared by all cells in an individual, somatic SVs are shared only by a subset of cells; such SVs are newly generated in individual somatic cells and then accumulate during divisions of those cells during the lifespan. Somatic SVs can be neutral or pathological; the latter can affect clonal hematopoiesis in the blood compartment and lead to cancer at various sites in the body. HSC, hematopoietic stem cell.

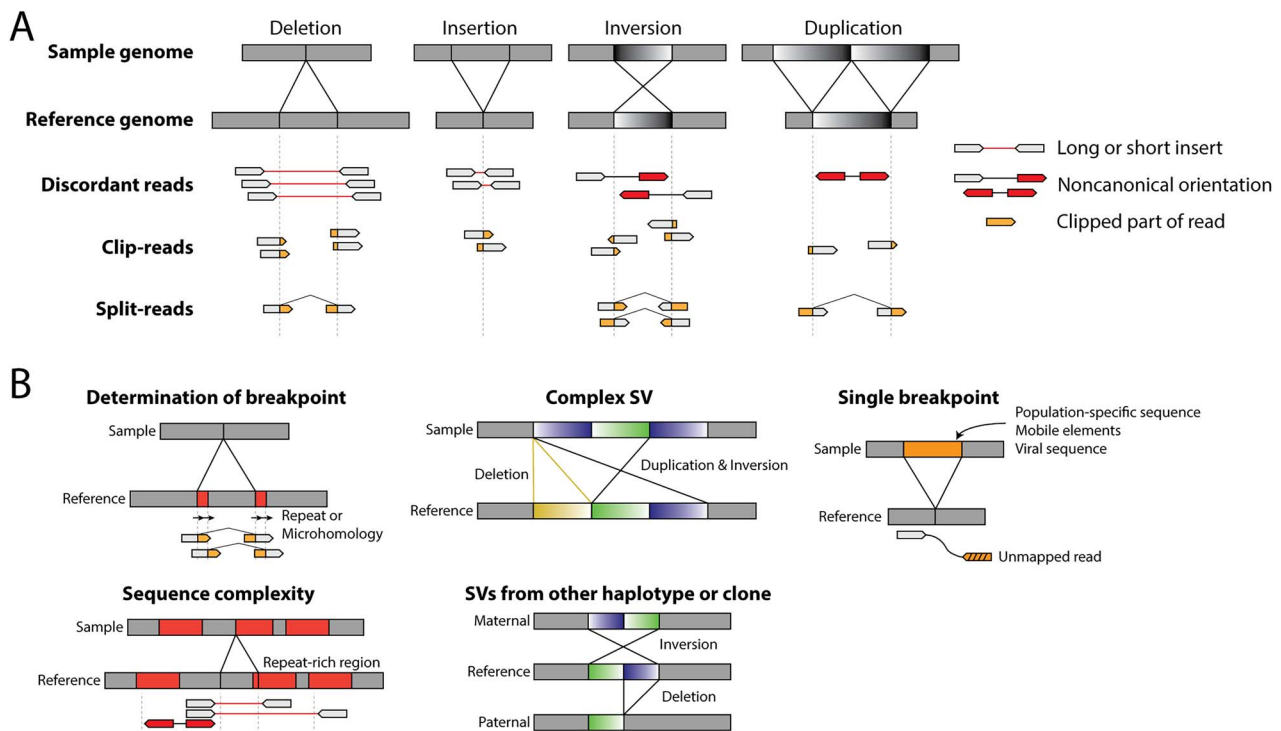


Figure 2. Abnormalities in mapped reads and complications affecting SV detection. **(A)** Abnormalities in mapped reads for different types of SVs. Sequencing reads are represented as arrows, and paired reads are connected by lines. In discordant reads, a short or long insert is denoted by a red line, and a noncanonical orientation of reads is indicated by red arrows. For both clip-reads and split-reads, the clipped portion of the read is marked in orange. In split-reads, a single read is mapped to two distinct sites, and corresponding clipped sequences are also marked in orange. For simplicity, only forward reads are displayed, and one of the read pairs is displayed on clip-reads and split-reads sections. **(B)** Challenges for the detection of SVs. Repetitive sequences are denoted as red boxes, and inserted sequences absent from the reference genome are marked as orange boxes. These could originate from population-specific sequences, mobile elements, or viral sequences.

in the reference genome, indicating that two or more segments of the read originated from different genomic locations, the read is called split-read (Figure 2A). If SV is accompanied by a change in the copy number, the sequencing depth of the genomic region should also be considered. Some previous tools have also used local assembly around breakpoints to resolve the structure and sequence of SVs [8–11]. However, multiple factors can complicate the detection of SVs (Figure 2B). These include the presence of repeats or microhomology at breakpoints, complex SVs, single breakpoints, SVs in repetitive regions or the presence of different SVs from two alleles or clones at the same locus.

Previous studies based on long-read sequencing have shown that healthy individuals can have as many as 20 000 SVs on average compared with the reference genome [12–15], which is double the number detected by short-read sequencing [16]. The high number of SVs in healthy individuals indicates the prevalence of both somatic and germline SVs, leading to a focus on somatic SVs in cancer genomic studies. The most common and effective way to filter germline variants from somatic variants is to compare a tumor sample to a matched-normal sample [17] because germline but not somatic variants would be found in both samples. Although the use of matched-normal samples is the *de facto* standard for somatic variant detection because of the reliability of this approach [7, 17], the cost of sequencing and interpreting both tumor and normal genomes, as well as issues related to patient consent and the availability of pre-existing data, mean that matched-normal samples are not always used.

Reducing sequencing and computing costs for somatic variant analysis is important, particularly in cohort-level genome studies

and when there is an urgent clinical demand. Reductions can be achieved by optimizing bioinformatics analysis pipelines, developing new algorithms [18], designing dedicated hardware [19] or developing more reliable methods [18, 20], particularly in cases where matched-normal samples are not available.

Somatic SV detection with matched-normal samples

Somatic SV detection with matched-normal samples typically involves a two-step process: detection of SVs in both cancerous and matched-normal samples, followed by filtration of common variants, which are likely to be germline variants. Different somatic SV callers utilized different filtration approaches (Table 1). For example, Lumpy [21], DELLY [22] and GRIDSS [8] filter SVs that are also present in normal samples, whereas Manta [10] uses a probabilistic filter rather than a deterministic filter by comparing normal and cancer-oriented reads. To detect somatic SVs with high sensitivity and specificity, SvABA [9] uses local assemblies made from data of both cancerous and normal samples. However, local assemblies often generate misassemblies, leading to false identification of somatic SVs and their breakpoints, as well as high computational costs and time. Recently, ETCHING [18] was developed; it utilizes a *k*-mer filtration strategy to remove non-informative normal reads from cancer reads. These reads are unlikely to support tumor-specific somatic variants and are removed before SV calling. The advantage of this approach is that it reduces computing cost and time and increases specificity by pre-filtration of non-informative reads.

Table 1. Computational methods that call somatic SVs using bulk and single-cell WGS

Tool	Applicable scenarios	Input data	Principle	Other features			References
				Local assembly	Segmentation	Machine learning	
ETCHING	Detecting SVs from bulk-cell, with and without matched-normal.	Bulk WGS	Utilization of pan-genome k-mer sets to filter population shared SVs. Random Forest model was applied for the removal of false positives.	-	-	O	[18]
DELIY	Detecting SVs from bulk-cell, with matched-normal.	Bulk WGS	Finding SVs via discordant read-pairs and clip-/split-reads.	-	-	-	[22]
Manta	Detecting SVs from bulk-cell, with matched-normal.	Bulk WGS	Scoring an association graph using read-pair and split read.	-	O	-	[10]
Lumpy	Detecting SVs from bulk-cell, with matched-normal.	Bulk WGS	Integration of SV evidences from read-pair, split-read, and read-depth to calculate probabilistic distribution of breakpoints.	-	-	-	[21]
novoBreak	Detecting SVs from bulk-cell, with matched-normal.	Bulk WGS	Tumor reads are collected via k-mer filtration, cluster & local assembly, then aligned to genome.	O	-	-	[11]
GRIDSS	Detecting SVs from bulk-cell or Colony-based WGS	Bulk WGS	Assembly using de Bruijn graph from mapped abnormal reads and the assembled contigs realigned to genome for variant call.	O	-	-	[8]
BRASS	Detecting SVs from bulk-cell or Colony-based WGS	Bulk WGS	Identification of breakpoint by assembling clustered abnormal read pairs.	O	-	-	Sanger Institute
Cue	Detecting SVs from bulk-cell, with matched-normal.	Bulk WGS	Converts alignments into images encoding SV signals, then uses a stacked hourglass convolutional neural network to predict SV type, genotype, and genomic location.	-	-	O	[80]
SvABA	Detecting SVs from bulk and scWGS	Bulk or Single-cell WGS	Assembly of aberrantly aligned reads using SGA	O	-	-	[9]
HMMCopy	Detecting CNAs from scWGS	Bulk or Single-cell WGS	Divides genome into 1 kb windows, extracts read depth, applies filters and GC content correction, computes log ratio for each window.	-	O	-	[45]
CopyNumber	Detecting CNAs from bulk and scWGS	Bulk or Single-cell WGS	Using copy number segmentation and penalization for optimal least squares solution.	-	O	-	[49]
Ginkgo	Detecting single-cell CNAs from scWGS	Single-cell WGS	Division of genome into bins. Performing OC and segmentation to determine copy-number states.	-	O	-	[46]
AneuFinder	Detecting single-cell CNAs from scWGS	Single-cell WGS	Applying a Hidden Markov Model to bin read counts to predict copy number states.	-	O	-	[47]
SCNV	Detecting single-cell CNAs from scWGS	Single-cell WGS	Use of a bin-free segmentation approach for highest resolution breakpoint detection and copy number calling.	-	O	-	[48]
SCOPE	Detecting single-cell CNAs from scWGS	Single-cell WGS	Utilization of a Poisson latent factor model for normalization, cross-sample segmentation, and breakpoint identification.	-	O	-	[50]
CHISEL	Detecting single-cell CNAs and copy-neutral LOH from scWGS	Single-cell WGS	Inference of allele-specific and haplotype-specific copy numbers in single cells from low coverage DNA sequencing data by integrating the weak allelic signals across thousands of individual cells.	-	O	-	[51]
Mosaicatcher	Detecting single-cell SVs from Strand-seq	Single-cell WGS (Strand-seq)	Detection of various SV classes in single-cells based on single-cell tri-channel processing including depth, strand, and phase.	-	O	-	[55]
TranslocatoR	Detecting single-cell translocations from Strand-seq	Single-cell WGS (Strand-seq)	Inference of translocation partners via single-cell tri-channel processing based on template strand co-segregation analysis.	-	O	-	[55]

^aINV: inversion; TRL: translocation; INS: insertion; LOH: loss of heterozygosity.

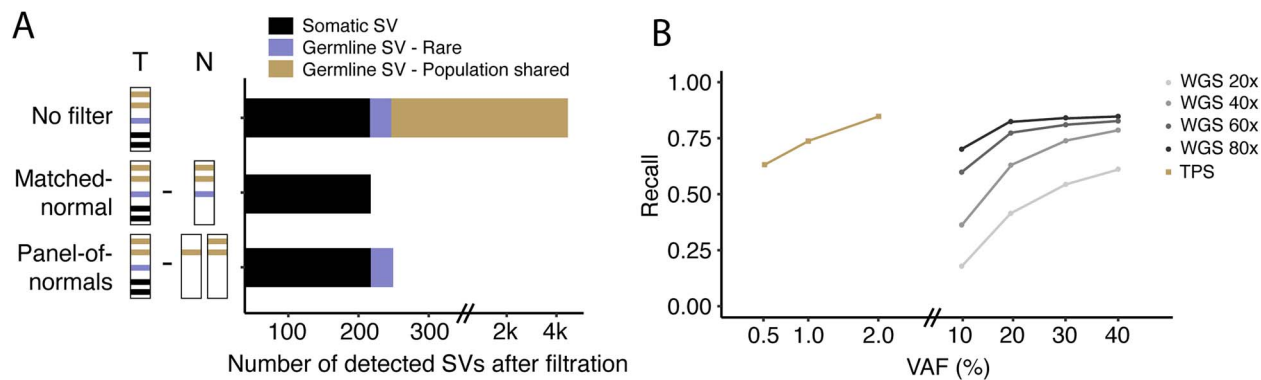


Figure 3. Benchmarking of technologies/algorithms to explore somatic SVs using bulk WGS. **(A)** (left) Schematic diagram of different filtration strategies. 'T' and 'N' indicate tumor and normal, respectively. The normal sample can either be absent (no filter), derived from the same patient (matched-normal) or sourced from a catalog of SV set (panel-of-normals). (right) The number of SVs expected to be detected by three different filtration strategies. **(B)** Mean detection power for somatic SVs across varying techniques and sequencing depths.

Somatic SV detection without matched-normal samples

Several approaches have been proposed for removing germline variants in the absence of matched-normal samples. One approach is to consider variant allele frequency (VAF). This method assumes that heterozygous germline variants have an expected VAF of 0.5, whereas heterozygous somatic variants have a VAF much lower than 0.5 [7]. However, VAF is strongly affected by copy number alterations (CNAs) as well as tumor purity and heterogeneity, which could result in false-positive and false-negative results. VAF has mostly been used to account for SNVs and indels; estimating the VAF of SVs is not straightforward because of the complexity of SVs and the effects of altered copy numbers [23].

Another approach is to utilize a 'panel-of-normals' (PON), which leverages the catalog of germline SVs compiled from the Database of Genomic Variants [24], 1000 Genome Project [25] and gnomAD [26, 27], to filter out recurrent germline variants or recurrent technical artifacts. Although the PON approach has been widely adopted for the analysis of small variants, its application to the analysis of SVs is relatively recent and is still under development [18, 20]. Although the PON of germline SVs is required for successful somatic SV detection, the catalog of germline SVs is far from complete, partly due to the low accuracy of current germline SV callers. Several tools that estimate population allele frequencies of SVs can be adopted as alternative approaches, although they are not designated for the filtration of germline variants [28–31]. This approach can also be logically extended to predict somatic variants by searching for unique variants that are not present in germline SV databases.

The most recently developed SV caller, ETCHING [18], takes advantage of a PON-like approach for germline filtration in tumor-only cases. Instead of using a cataloged set of SVs, ETCHING utilizes a unique set of *k*-mers constructed from pangenomes, which implicitly include recurrent germline structural variations. Alternatively, STIX [20] utilizes a pre-indexed alignment of population-level sequencing data to identify reads that support SVs. An SV supported by reads from multiple samples can be considered a shared germline variant and is less likely to be a somatic variant. Both ETCHING and STIX display filtering performances comparable to those obtained by methods using matched-normal samples.

Despite the success of somatic SV prediction methods for tumor-only samples, these methods still produce false-positive calls due to rare or private germline variants (Figure 3A). Additionally, the use of a PON can mistakenly remove true somatic

variants that match germline variants by chance [7, 17]. Moreover, compared to germline SNVs and indels [32], germline SVs are less recurrent, with 92% having a VAF < 1% in gnomAD data [26]. Therefore, the use of matched-normal samples is often considered an ideal approach for somatic variant detection.

Inference of SV allele frequencies in bulk genome sequencing data

Clinical interpretation of variants is influenced by tumor purity and heterogeneity [33, 34]. The general procedure to account for these factors is to precisely infer the VAF of genetic variants. In the case of SNVs, VAF is calculated by dividing the number of reads supporting a certain variant allele by the total read depth at a certain position. Theoretically, the VAF of an SV can be calculated similarly by counting the number of reads supporting a breakpoint divided by the total read depth at the breakpoint (Figure 4A). However, SVs are generally associated with two breakpoints with unequal VAFs and are often involved in CNA, hindering their VAF estimations. In addition, the criterion for selecting breakpoint-supporting reads is not straightforward and requires proper optimization. For instance, SVclone [23] optimized the criterion for supporting reads by simulating SVs with known VAFs and tumor purity, whereas Meltos [35] and Breakdown [36] used the maximum likelihood estimation of the probability distribution of the VAFs.

A standard sequencing depth (30× for normal samples and 60× for tumor samples) for detecting somatic variations is often insufficient to infer the low VAFs of SVs. For example, the detection power rapidly decreased for SV VAFs < 20% [37] (Figure 3B). This phenomenon is problematic because some tumors, such as lung adenocarcinoma, lung squamous cell carcinoma, and kidney renal clear cell carcinoma, often have a tumor purity < 50% [38], making it difficult to detect SVs present in < 40% of the cells in the tumor population via standard WGS. More sequencing reads can be generated to detect SVs with low VAFs; however, the cost remains an issue. Targeted panel sequencing (TPS) can be an alternative approach for detecting SVs with low VAFs; TPS can detect SVs with a VAF of merely 0.5% with thousands of reads in the targeted region (Figure 3B). However, the TPS approach is limited in its ability to detect long-range rearrangements or interchromosomal translocations without defining target regions.

The inferred VAF of SVs can be used to decipher the cell fraction (CF) with SVs, which refers to the clonal fraction of cancer cells carrying SVs. In copy-neutral genomes, the CF would be doubled to the VAF in the case of heterozygous variants (Figure 4A and B).

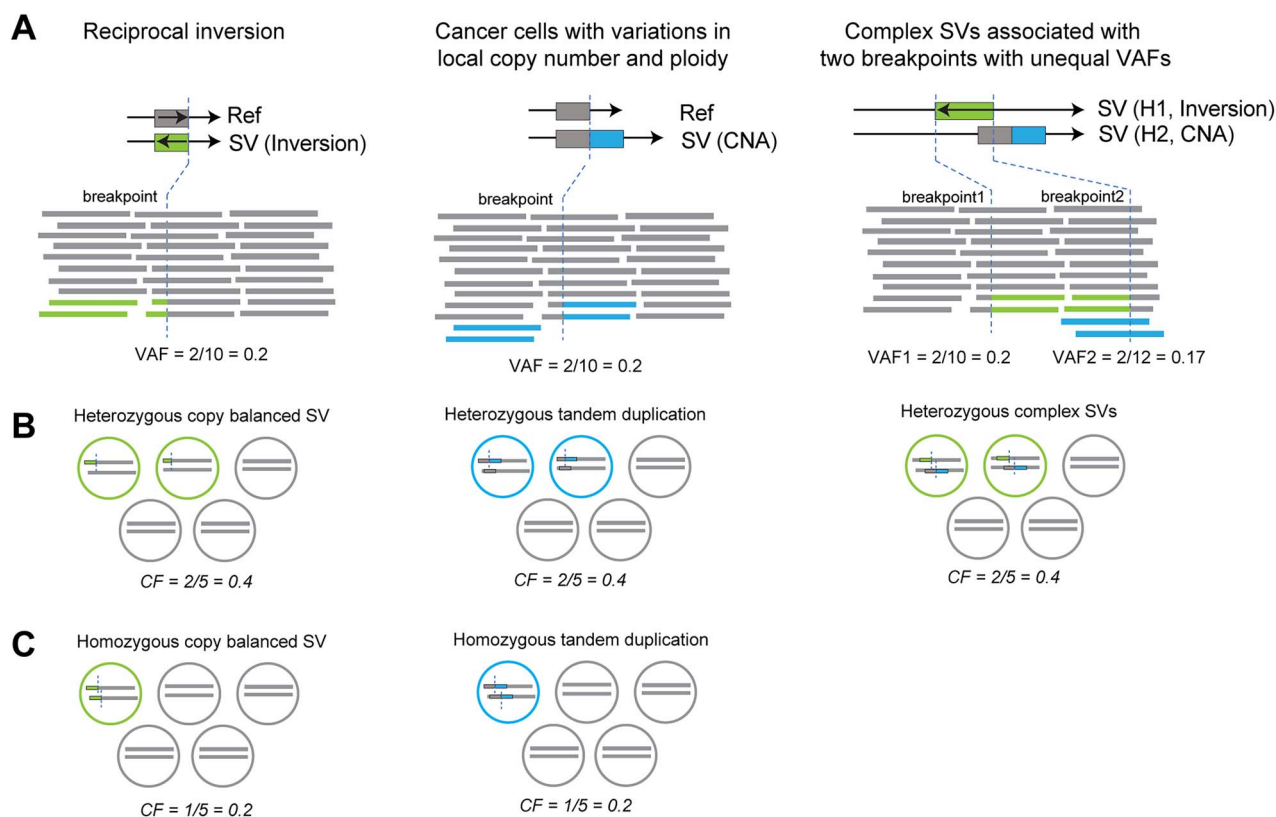


Figure 4. Conceptual comparison of the VAF and CF used in bulk and single-cell WGS, respectively. **(A)** VAF indicates the number of variant reads divided by the total number of reads observed at the SV breakpoint and is often used for bulk WGS analysis. Three example SV cases are shown in this panel. The upper panel shows the genomic configuration of SVs, and the lower panel demonstrates the possible outcome of sequencing read alignment to the reference genome for each SV case. Horizontal bars in the lower panel represent the sequencing read from WGS. Green and blue bars indicate the segments originated from rearranged genomic regions (green and blue for inversion and tandem duplicate, respectively), and the gray bars show the genomic segments originated from normal genomic regions. **(B–C)** CF is the fraction of cancer cells harboring a certain SV and is often used in single-cell analysis. The VAF and CF values can be interconvertible in specific situations by considering the heterozygosity **(B)** or homozygosity **(C)** of the variant. Each circle represents a single cell, and green and blue borders of circles indicate cells that contain DNA with rearranged genomic segments (SVs).

However, variations in the region with local CNA or aneuploidy of cancer cells complicate CF estimation (Figure 4B and C) [39]. The need to infer purity, copy number, and ploidy could introduce unavoidable errors and add noise to the CF estimations. Furthermore, complex SVs can hinder the inference of the true CF from the SVs because of imperfect estimations of the VAF and the admixture of clonal SVs in bulk WGS data [23].

Genotypic analysis at the single-cell level using single-cell sequencing offers a promising means of addressing the limitations described above. Sequencing DNA from individual cells allows tumor CF with SVs to be determined directly, even if the variant is present in only a small fraction of the cells. Single-cell sequencing also allows the quantification of cell-to-cell heterogeneity and identification of subclonal variants, which can be easily missed by bulk sequencing approaches.

ADVANCES IN METHODS FOR SOMATIC SV DETECTION IN SINGLE-CELL GENOME SEQUENCING DATA

Single-cell WGS (scWGS) approaches were developed to detect somatic SVs in single cells [40, 41] (Figure 5). ScWGS requires DNA amplification to sensitively detect somatic variants at the single-cell level. During DNA amplification, whole-genome amplification (WGA) or multiple displacement amplification (MDA) approaches are used. WGA relies on polymerase chain reaction (PCR)

amplification of a small portion of the genome, which is beneficial for CNA detection but not for SNV detection. In contrast, MDA can detect SNVs but not CNAs owing to its uneven coverage [42]. To overcome this problem, the primary template-directed amplification (PTA) approach [42] was recently introduced and is now commercially available. PTA has been applied to detect somatic SVs in primary human leukemia cells as well as SVs induced by CRISPR off-target editing in several cell lines and human cord blood cells [42]. In this study, SV calling was performed using SvABA, which was previously developed for bulk WGS analysis [9].

To increase the cell number profiled in an experiment and reduce biases from the DNA amplification step, an upgraded direct library preparation (DLP+) method for amplification-free scWGS following DNA tagmentation was developed using high-density nanowell arrays and picoliter volume piezo dispensing technology [43]. This method enables the sequencing of >50 k single cells from diverse cell types using microfluidic devices without a pre-amplification step, allowing the detection of CNAs at the single-cell level. The DLP+ method was recently applied to study the mutational processes occurring at the single-cell level. In this study, tens of thousands of genomes from wild-type and TP53 deficient mammary epithelial cells, as well as from primary triple-negative breast cancer and high-grade serous ovarian cancer cells, were sequenced [44]. To detect CNAs at the single-cell level, reads were first counted in non-overlapping 500 kb genomic bins. After

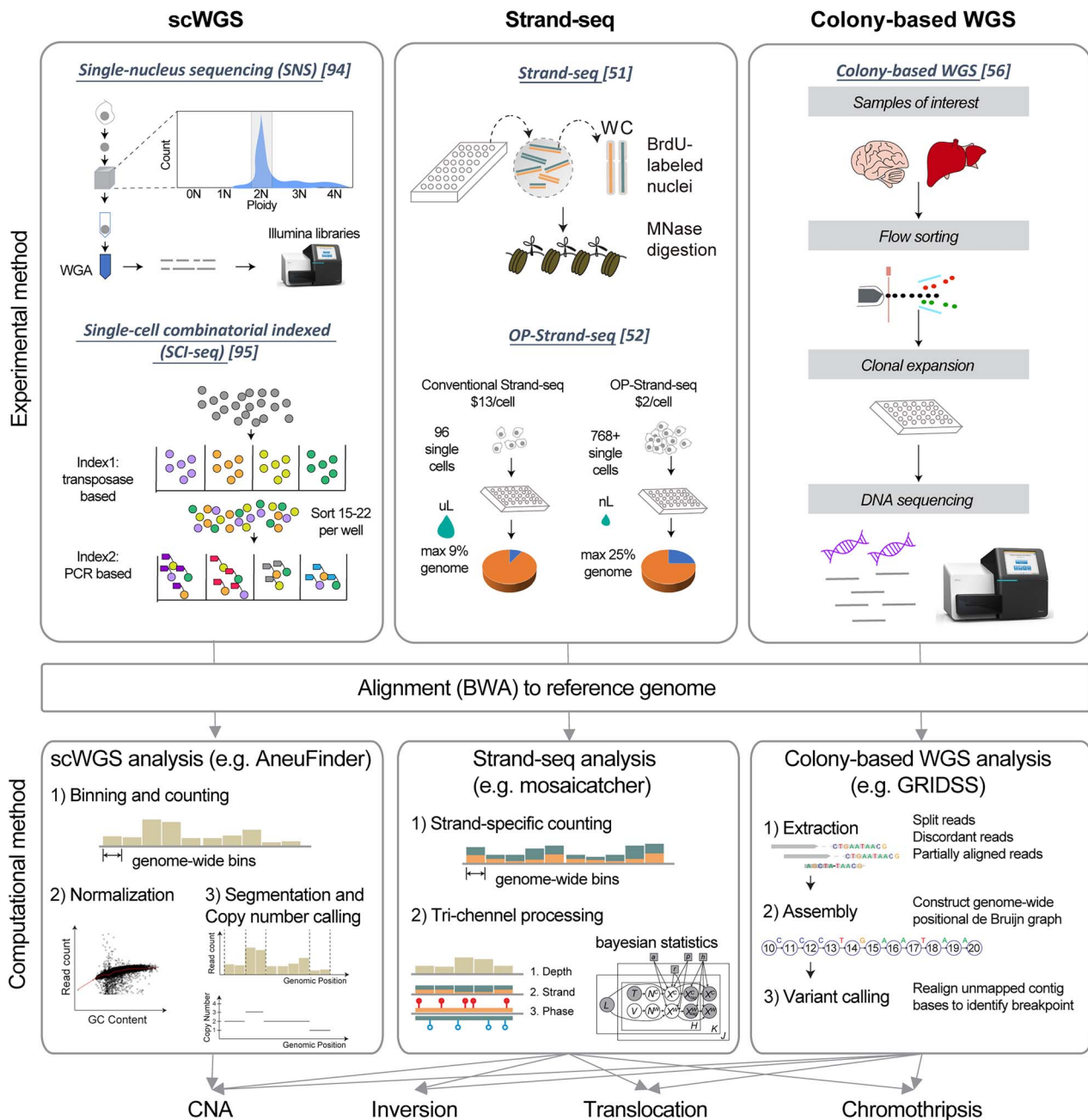


Figure 5. Technologies to explore somatic SVs using single-cell-derived genomes. Genomes in individual cells can be profiled either from DNA materials extracted from single cells or from colonies generated by clonal expansion of single cells. Both approaches include similar pre-processing and alignment steps (alignment, normalization and segmentation). Afterwards, scWGS [40, 41] and Strand-seq use single cell-specific algorithms to detect somatic SVs. In the figure, W and C indicate the negative template strand (Watson, 'W') and the positive template strand (Crick, 'C') that is sequenced in Strand-seq. Colony-based WGS uses conventional algorithms originally developed for bulk WGS data analysis.

normalizing the GC bias, the copy number was predicted using the HMMcopy algorithm [45], which was originally developed for bulk WGS data.

Owing to the unique issues and challenges involved in scWGS data analysis, several algorithms have been designed (Figure 5 and Table 1). Ginkgo [46], AneuFinder [47], SCNv [48], CopyNumber [49], SCOPE [50] and CHISEL [51] are currently available tools for calling CNAs from scWGS data. These tools comprise the following common steps: binning, GC bias correction, mappability correction, removal of outlier bins and cells, segmentation, and calling of absolute copy numbers [52]. These methods are specialized for the detection of CNAs but not for copy-neutral SV types or complex events such as chromothripsis.

Because there is only a limited amount of DNA in a single cell, it is challenging to acquire high sequencing coverage per cell (values typically range from 0.02× to 10×, depending on the scWGS technique). Such low coverage makes the detection of copy-neutral SVs challenging, because these events can only be detected by reads spanning breakpoints. Thus, single-cell template strand sequencing (Strand-seq), which detects strand-state changes, has the advantage of identifying copy-neutral SVs [53]. This is possible because the cells are treated with BrdU during library preparation. After one cell division, the non-template strand containing BrdU substitutions is removed, and only the template strand is sequenced. The original Strand-seq protocol is based on single-cell sorting in 96-well plates. To increase throughput and

enable sequencing of more cells in parallel, the one-pot (OP)-Strand-seq method, which involves the construction of ~768 single-cell libraries in one batch, was recently devised [54]. To analyze these data, the scTRIP pipeline was developed by implementing the Mosaicatcher algorithm, which uses depth, strand state, and haplotype information to efficiently detect CNAs, inversions, and complex rearrangements at the single-cell level [55]. Additionally, various computational algorithms, such as BreakpointR [56], StrandPhaseR [57] and TranslocatoR [55], are available for the in-depth analysis of Strand-seq data (Figure 5 and Table 1).

Although Strand-seq is advantageous for efficiently detecting copy-neutral rearrangements, the mean coverage is only $0.05\times$ per cell, meaning that SNVs or SVs smaller than 200 kb are difficult to detect at the single-cell level. By increasing genomic coverage to up to $\sim 0.1\text{--}0.25\times$ per cell, OP-Strand-seq enables breakpoint detection with increased confidence; however, the coverage is still too sparse for the analysis of SNV profiles or small SVs (<200 kb) [54]. As an alternative approach, colony-based WGS methods, in which individual cells are physically separated and cultured to generate single-cell-derived colonies for bulk WGS, are available. This approach was used to study clonal dynamics during hematopoiesis and led to the detection of 77 somatic SVs from 3579 genomes of single-cell-derived colonies across 10 human subjects [58]. In this study, somatic SVs were called using GRIDSS [8]. This approach was also applied for the detection of somatic SVs from single bronchial epithelial cells across 16 samples to examine the relationship between smoking and somatic mosaicism using the variant caller BRASS [59].

CF denotes the proportion of cells harboring the somatic mutation of interest [23]. While the VAF is commonly calculated in bulk WGS analysis, the CF is often calculated in single-cell analysis, even though the two metrics are interconvertible (Figure 4). For a balanced heterozygous SV, the CF is twice that of the VAF, whereas for a balanced homozygous SV, the CF is the same as the VAF. Computational algorithms that automatically convert VAF into CF are available [23]. In this process, tumor purity is used to calibrate the supporting read counts for the SVs. Although the CF of a given somatic SV can be inferred from bulk WGS data [23], calculating the CF from scWGS is more straightforward, especially for CNAs. An advantage of the single-cell-based approach is that not only subclonal SVs shared by multiple cells, but also *de novo* SVs that appear in a single cell or in a very low CF can be detected. Additionally, the single-cell approach enables us to easily understand the subclonal structure by revealing the co-occurrence of multiple somatic SVs in certain cell subpopulations in one sample [60].

INFERENCE OF SVS THROUGH SINGLE-CELL TRANSCRIPTOME AND EPIGENOME ANALYSIS

An orthogonal computational method was used to infer somatic SVs from single-cell transcriptomic data (Figure 6). The original purpose of this method was to link genetic variations (somatic SVs) to phenotypic variations (transcriptomic alterations). This method assumes that the average gene expression level in a genomic bin of a certain size (e.g. ≥ 100 expressed genes) is proportional to the DNA copy number in the genomic bin. More specifically, genomic bins that contain somatic CNA (SCNA) show abnormally larger (gain of copy number) or smaller (loss

of copy number) average gene expression levels than other bins in karyotypically normal positions. Based on this assumption, Patel et al. [61] attempted to infer single-cell SCNAs (scSCNAs) using single-cell RNA-seq (scRNA-seq) data from patients with glioblastoma. The computational algorithm was implemented using the InferCNV package [61], which is now widely used to infer scSCNAs from scRNA-seq data, regardless of the single-cell library preparation platform. Another method, CONICS [62], uses a Gaussian mixture model to infer the presence of SCNAs at the clonal level. In contrast, HoneyBadger [63] and Numbat [64] attempted to improve the accuracy of scSCNA prediction by phasing haplotypes and performing B-allele frequency estimation. The inferred scSCNAs are integrated with a functional readout (gene expression) using algorithms such as LIGER [65] and PAGODA [66] to investigate the functional impact of a CNA at the single-cell level. STARCH [67] was recently developed to infer copy numbers and clonal structures based on spatially resolved transcriptomes. In addition to scRNA-seq, single-cell ATAC-seq (scATAC-seq) can be used to infer scSCNAs using Alleloscope [68] and Copy-scAT [69]. The resulting scSCNAs can be integrated with functional readouts of chromatin accessibility using ChromVAR [70] and STREAM [71] to gain biological insights into the effects of scSCNAs on cellular phenotypes. scSCNA inferences provide a great opportunity to link genetic variations and cellular and molecular phenotypic alterations in thousands to millions of cells, which are regarded as experimental replicates. However, these approaches have technical limitations, as: (1) the SV size should be large enough (≥ 10 Mb); (2) the SV should be copy-imbalanced and (3) the SV should have a high allele frequency.

Recent efforts to improve the confidence with which scSCNAs can be inferred from scRNA and scATAC-seq profiles have exploited the potential to integrate sample-matched bulk WGS or whole exome sequencing (WES) data. These approaches provide high-confidence and high-resolution predictions of scSCNAs by including precise breakpoint information, allowing for the detection of even small scSCNAs (<10 Mb). The improved results led to the clustering of cells into subclones harboring a certain scSCNA and better inferences of molecular phenotypes for distinct subclones at a large scale ($\sim 1\,000\,000$ cells). Using these advantages, the Cardelino method integrates somatic clonal architecture from bulk DNA sequencing and scRNA-seq data to investigate differentially expressed genes between subclones [72]. Alleloscope combines scATAC-seq and sample-matched bulk WGS/WES data to infer subclonal epigenome profiles associated with DNA rearrangements at the single-cell level [68]. An unmet need for this approach is the lack of relevant algorithms to infer copy-balanced SVs from scRNA-seq or scATAC-seq data. Copy-balanced SVs can only be detected if reads from scRNA-seq and scATAC-seq span breakpoints.

As an alternative approach, single-cell nucleosome occupancy and genetic variation analysis (scNOVA), which analyzes Strand-seq profiles to extract somatic SVs and nucleosome occupancy (NO) signals from the same single cells, was recently introduced [73]. In addition to detecting various SV types, this method takes advantage of the NO signal from MNase treatment during library preparation to infer transcriptional activity across promoters and gene bodies. Although this approach produces a low number of cells per experiment ($\sim 100\text{--}1000$ cells) and displays a narrow dynamic range of functional readouts compared to scRNA-seq, it provides complementary information for examining the functional impact of somatic copy-balanced and complex rearrangements in addition to SCNAs.

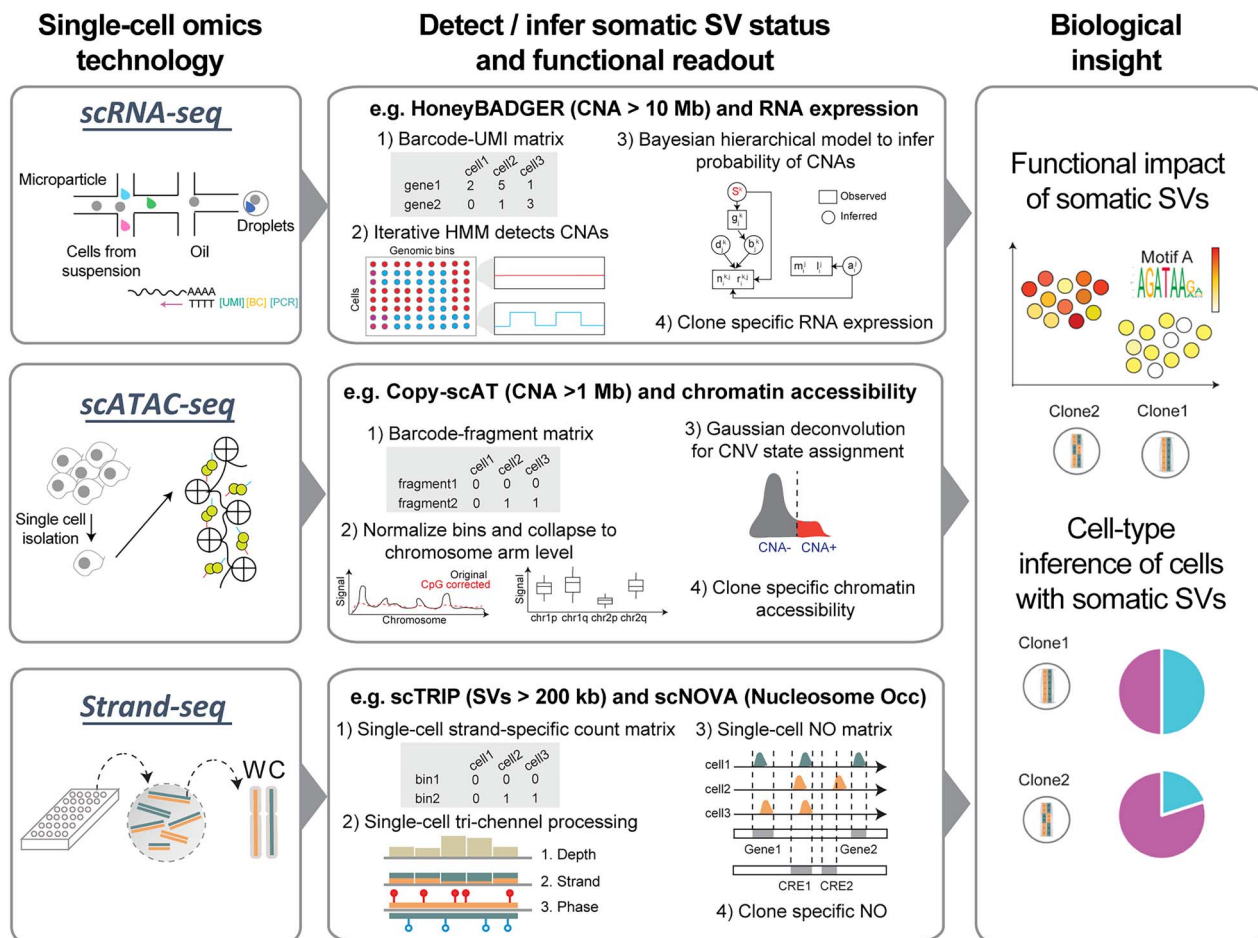


Figure 6. Functional analysis of somatic SVs using single-cell transcriptomes and epigenomes. To investigate the functional impact of somatic SVs, datasets obtained from single-cell omics technologies, such as scRNA-seq, scATAC-seq, and Strand-seq, can be used. From these datasets, somatic SVs are first detected or inferred to assign cells into two groups: (1) subclones that harbor certain somatic SVs of interest or (2) subclones without those SVs of interest. Then, the functional readouts for the two groups are compared and may provide biological insights such as the functional impacts of somatic SVs and cell-type inferences with and without certain somatic SVs of interest. In the bottom left panel, W and C indicate the negative template strand (Watson, 'W') and the positive template strand (Crick, 'C') that is sequenced in Strand-seq, as described in Figure 5.

BULK AND SCWGS APPROACHES FOR SOMATIC SV DETECTION

Recent advances in computational and experimental techniques for detecting somatic SVs at the bulk and single-cell levels are discussed (Table 1). Bulk- and single-cell WGS approaches have their own strengths and weaknesses in the analysis of somatic SVs in terms of the type, size and VAF (or CF) of detectable SVs and the cost of analysis (Table 2). Although bulk WGS covers most SV types, including CNAs and copy-balanced rearrangements, scWGS approaches display different detection capabilities, depending on the associated library preparation strategies. For instance, scWGS using the PTA method produces a relatively high coverage of up to $\sim 10\times$, which allows for higher sensitivity for detecting all SV types. However, highly multiplexed single-nucleus sequencing methods produce $0.02\text{--}0.1\times$ coverage, enabling only the detection of CNAs [74]. The recently introduced Strand-seq provides an alternative method for detecting all types of SVs, including copy-balanced events [55].

Bulk WGS can detect SVs with a broader range of sizes than scWGS when the value of CF harboring a specific somatic SV is high. However, the accuracy of SV detection is dramatically reduced when the CF is low [18]. In contrast, although scWGS and Strand-seq have a much lower sequencing coverage than bulk

WGS, they display a higher sensitivity for SVs of $\geq 100\text{--}200$ kb in length [52] even when the CF is low [73]. PTA results in higher sequencing coverage ($\sim 10\times$), which makes it possible to detect interchromosomal SVs and CNAs < 50 kb in length [42, 75].

The number of samples/cells per run in bulk and single-cell WGS experiments can also be a technical limitation of SV calling. For instance, the number of cells that can be sequenced in one experiment is limited to 96 cells/plate for single-cell WGS, such as the WGA and PTA methods, which makes it difficult to detect somatic SVs whose CF is $< 1\%$. In addition, when SV is detected in only one of 96 cells, it is often difficult to determine whether it is a technical artifact or a true biological signal. As an alternative, DLP+ provides a higher throughput that can profile up to 1000 cells/run. For bulk WGS, there is no limitation to increasing the number of cells in the starting material; however, one experiment generally requires ~ 300 ng of gDNA, and a maximum of 24 samples/run can be performed in parallel.

Regarding the cost of analysis per sample, bulk WGS with a matched-normal sample costs approximately 2000 US dollars, while that without a matched-normal sample costs approximately two-thirds. Although scWGS generally costs much more than bulk WGS, the cost of scWGS largely depends on the number of cells in the sequencing library, which ranges from

Table 2. Comparison between bulk-cell- and single-cell-based somatic SV identification

Detectable SVs	Bulk WGS approach		Single-cell approach					
	With matched normal	Without matched normal	Single-cell WGS			Strand-seq		CNA inference from scRNA-seq
			WGA	PTA	DLP+	Original	OP	
CNAs	O	O	O	O	O	O	O	O
Inversion	O	O	–	O	–	O	O	–
Translocation	O	O	–	O	–	O	O	–
Chromothripsis	O	O	–	O	–	O	O	–
Size resolution of SVs	No limit	No limit	100 kb	No limit	500 kb	200 kb	200 kb	10 Mb
Minimum VAF or CF	0.2	0.2	no limit (1 cell)	no limit (1 cell)	no limit (1 cell)	no limit (1 cell)	no limit (1 cell)	no limit (1 cell)
Number of sequenced cells	No limit	No limit	96/plate	96/plate	1000	96/plate	768	10~50 k cells
Cost	Approx. \$2000/sample	About two-thirds of the cost with matched normal	Approx. thousand to several thousand dollars/sample					

approximately thousand to several thousand US dollars per sample.

FUTURE PERSPECTIVES
Clinical applications of somatic SV detection

The detection of cancer-driving and druggable somatic SVs has become clinically common for companion diagnostics for targeted cancer therapies. Such clinical procedures are mostly carried out using cancer panel sequencing (also known as targeted sequencing), often focusing on fusion genes and CNAs of oncogenes and tumor suppressor genes [76, 77]. Recently, a dramatic reduction in sequencing costs has led to the clinical use of genome-scale analysis (WGS) for detecting therapeutically targetable SVs. However, WGS-based somatic SV prediction as a diagnostic application requires low computing costs, high accuracy, and the availability of a matched-normal sample. A newly developed SV calling program, ETCHING [18], significantly reduced the computing time of WGS-based SV prediction to within a couple of hours and allowed the effective detection of somatic cancer-driving SVs even in clinical situations in the absence of a matched-normal sample. Another challenge associated with WGS-based SV prediction in the clinical setting is the need to determine which SV is an actual cancer driver and/or therapeutic target among the tens to hundreds of resulting somatic SVs. Similar to the SNVs, the majority of the resulting somatic SVs are passengers. Moreover, driver SVs can affect large genomic regions harboring multiple cancer-related genes, raising the question of which of them is the cancer driver. To address these questions, somatic SV calls can be integrated with single-cell transcriptome or epigenome analysis (Figure 7). The SV breakpoints detected by bulk WGS could guide the confident inference of scSCNAs from single-cell functional genomic (transcriptome or epigenome) profiles, which could uncover a subclonal structure. Subsequent comparisons of single-cell expression and chromatin accessibility among the subclones could reveal the phenotypic consequences of somatic SVs, which would help narrow down the potential driver SVs and genes. In fact, scNOVA can be an efficient option for investigating the functional consequences of somatic SVs (≥ 200 kb) and identifying cancer-driving genes using a single-cell approach.

Somatic SV mosaicism that occurs with aging
Somatic SVs are detected not only in cancer cells but also in non-malignant normal cells. Somatic SVs that accumulate during

clonal expansion in hematopoiesis have also been examined in healthy individuals [58, 78]. Similar to somatic SVs in cancer cell clones, those in normal tissues are often present in a mosaic pattern with a very low allele frequency. Although the sensitivity of somatic SV callers in bulk tissue samples has increased, SV events with low allele frequencies are often missed in bulk WGS data. To detect rare SV variants (with a VAF $< 1\%$) resulting from somatic mosaicism, a single-cell approach would be a good choice because it can detect a rearrangement that is present in a single cell. Recently, hundreds of thousands of individual’s WGS datasets from government-led cohort-level WGS projects have been produced and analyzed [79, 80]. Along with germline variants, somatic variants were detected in the cohort-level WGS datasets. The joint utilization of bulk and scWGS data can help increase our understanding of somatic SV mosaicism at the population level.

Deep learning on structural variation
Similar to other genomic studies, deep learning approaches have also been applied to SV detection. For instance, Cue [81] leveraged convolutional neural networks (CNNs) to detect various SV types accurately, whereas DeepSVFilter [82] used a CNN to filter unclear SVs from previously identified SVs. Despite their potential, current deep learning methods have limitations such as a lack of high-quality, well-balanced data for model training, lack of interpretability and limited generalizability across different datasets. Hence, transfer learning [83], data augmentation [84] and the integration of multimodal features are promising strategies for enhancing SV detection. Deep learning has also shown significant promise in predicting the effects of SV on phenotypes. DeepSEA [85] and Basset [86] predict the effects of small non-coding variants on DNA accessibility and gene expression. More recently developed methods such as Sei [87] and ExPecto [88] extend the scope of variant sizes and predict outcomes directly from DNA sequences. Nonetheless, the validation efforts of these methods have primarily targeted small variants. Such models for understanding how SVs affect gene expression, DNA accessibility, and even higher-level phenotypes, such as disease risk, could lead to the identification of therapeutic targets for genetic diseases.

Structural variation analysis in pangenomes
Pangenomes represent the entire genomic content of all individuals in a cohort, along with large-scale efforts to catalog germline and somatic variations—the Database of Genomic Variants [24],

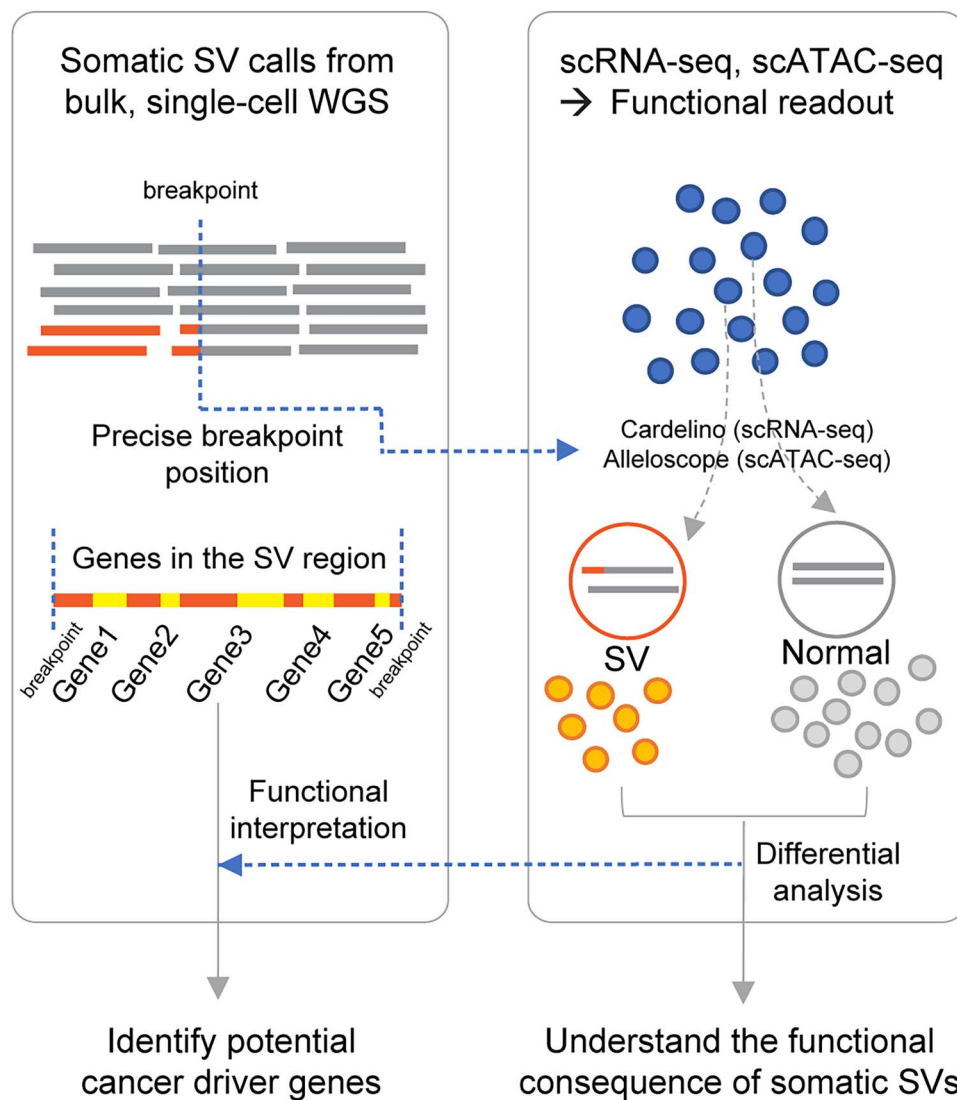


Figure 7. Complementary approaches of bulk WGS and single-cell genomics to understand the functional consequences of somatic SVs. To identify SV and genes that are cancer drivers and therapeutic targets, somatic SV calls can be integrated with single-cell transcriptome and epigenome analysis. Breakpoint positions obtained from bulk and single-cell WGS (scWGS) can improve the assignment of single-cell transcriptome and epigenome profiles in the corresponding subclones. Transcriptional and epigenomic comparisons between subclones can guide the functional interpretation of genes in the SV-affected region to identify potential cancer driver genes.

1000 Genomes Project [25], gnomAD [26, 27] and the Human Pangenome Reference Consortium (HPRC) [89, 90] for germline variants; The Cancer Genome Atlas (TCGA) [91] and the PCAWG [92] for somatic variants. Simultaneously, owing to recent computational advancements, the integration of pangenomes into everyday genomic analysis has become as practical as using a single reference genome [30, 89]. The field of SV particularly benefits from using pangenomes, employing them increasingly for identification and genotyping [29, 30, 89, 93]. Although these studies have largely focused on germline SVs, the detection of somatic SVs can also benefit from using pangenomes, as this approach helps reduce spurious alignments and better handle multi-allelic variations. This more accurate genomic representation within a tumor is especially beneficial when a matched-normal is unavailable [94]. Additionally, pangenomes better represent ethnicity-specific genomic variations, thereby increasing the accuracy and relevance of somatic SV detection in diverse populations. As efforts to construct pangenomes representing diverse populations continue (e.g. HPRC [90]), personalized medicine could benefit from the use of population-specific pangenome SVs.

Remaining computational challenges for SV detection at the bulk and single-cell level

The configuration of complex genomic regions remains to be resolved by SV callers based on bulk WGS data. Chromothripsis and chromoplexy involve many rearrangements occurring in parallel, which lead to the mixing of multiple chromosomes. Using existing SV callers, it is possible to detect breakpoints of these complex events based on split-reads; however, the reconstruction of the derivative chromosome from individual fragments remains challenging. Long-read WGS data, whose sequencing accuracy has recently improved, can provide useful complementary information regarding the order of scattered genomic pieces in this case. However, there is still a lack of SV callers that make use of both short- and long-read data modalities simultaneously. Therefore, new computational algorithms that can accurately reconstruct chromosomal configurations by systematically integrating short- and long-read bulk WGS datasets must be actively developed.

Furthermore, single-cell SV callers have largely focused on detecting CNAs because of the limited coverage of single-cell

sequencing data [52]. Recently, an attempt was made to call imbalanced SVs by pooling data from single cells, a process known as pseudo-bulk; the resulting SVs were subsequently subjected to single-cell analysis [95]. However, this approach has not been widely used. To obtain comprehensive SV profiles from individual cells, high-performance robust SV callers designed specifically for the analysis of data from single cells must be developed. Unlike for CNAs, there are no efficient methods for inferring copy-balanced SVs from scRNA-seq data. By investing more effort into the development of computational methods for detecting copy-balanced SVs from data derived from various single-cell technologies, we should achieve a broader understanding of the functional consequences of somatic SVs in the future.

Key Points

- This work reviews recent advances in the detection and interpretation of somatic SVs using WGS at both the tissue and single-cell levels.
- We describe and compare the performances of computational algorithms for somatic SV calling using bulk WGS data with or without a matched-normal control, and introduce recent works that reliably call SVs without matched-normal control.
- We highlight the unique features of cutting-edge, single-cell-based techniques for the analysis of somatic SVs and discuss the advantages and disadvantages of bulk and single-cell approaches, including sensitivity, usefulness for functional inferences and costs.
- We illustrate computational approaches for linking somatic SVs to functional readouts obtained from single-cell transcriptome and epigenome analyses, which hold promise for improving our understanding of the functional impact of somatic SVs on health and diseases.

ACKNOWLEDGEMENTS

We thank all members of the BIG lab and HY-IBB for their helpful discussions and comments.

FUNDING

National Research Foundation (NRF); Ministry of Science and ICT, Republic of Korea (2021R1A2C3005835, 2022M3E5F1018502, 2022M3A9I2082294 and RS-2023-00207840 to J.W.N.); Hanyang University (HY-202200000003414 to H.J.).

REFERENCES

- Chiang C, Scott AJ, Davis JR, et al. The impact of structural variation on human gene expression. *Nat Genet* 2017;**49**:692–9.
- Scott AJ, Chiang C, Hall IM. Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. *Genome Res* 2021;**31**:2249–57.
- Cosenza MR, Rodriguez-Martin B, Korbelt JO. Structural variation in cancer: role, prevalence, and mechanisms. *Annu Rev Genomics Hum Genet* 2022;**23**:123–52.
- Li Y, Roberts ND, Wala JA, et al. Patterns of somatic structural variation in human cancer genomes. *Nature* 2020;**578**:112–21.
- Kosugi S, Momozawa Y, Liu X, et al. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol* 2019;**20**:117.
- Zeng, Glicksberg BS, Li Y, Chen B. Selecting precise reference normal tissue samples for cancer research using a deep learning approach. *BMC Med Genomics* 2019;**12**:21.
- Jones S, Anagnostou V, Lytle K, et al. Personalized genomic analyses for cancer mutation discovery and interpretation. *Sci Transl Med* 2015;**7**:283ra53.
- Cameron DL, Schröder J, Penington JS, et al. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Res* 2017;**27**:2050–60.
- Wala JA, Bandopadhyay P, Greenwald NF, et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res* 2018;**28**:581–91.
- Chen X, Schulz-Trieglaff O, Shaw R, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 2016;**32**:1220–2.
- Chong Z, Ruan J, Gao M, et al. novoBreak: local assembly for breakpoint detection in cancer genomes. *Nat Methods* 2017;**14**:65–7.
- Pendleton M, Sebra R, Pang AWC, et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods* 2015;**12**:780–6.
- Sedlazeck FJ, Rescheneder P, Smolka M, et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* 2018;**15**:461–8.
- Chaisson MJP, Huddleston J, Dennis MY, et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* 2015;**517**:608–11.
- Wenger AM, Peluso P, Rowell WJ, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* 2019;**37**:1155–62.
- Mahmoud M, Gobet N, Cruz-Dávalos DI, et al. Structural variant calling: the long and the short of it. *Genome Biol* 2019;**20**:1–14.
- Mandelker D, Ceyhan-Birsoy O. Evolving significance of tumor-normal sequencing in cancer care. *Trends Cancer Res* 2020;**6**:31–9.
- Sohn J-I, Choi M-H, Yi D, et al. Ultrafast prediction of somatic structural variations by filtering out reads matched to pan-genome k-mer sets. *Nat Biomed Eng* 2023;**7**:853–66.
- Zhao S, Agafonov O, Azab A, et al. Accuracy and efficiency of germline variant calling pipelines for human genome data. *Sci Rep* 2020;**10**:1–12.
- Chowdhury M, Pedersen BS, Sedlazeck FJ, et al. Searching thousands of genomes to classify somatic and novel structural variants using STIX. *Nat Methods* 2022;**19**:445–8.
- Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* 2014;**15**:R84.
- Rausch T, Zichner T, Schlattl A, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 2012;**28**:i333–9.
- Cmero M, Yuan K, Ong CS, et al. Inferring structural variant cancer cell fraction. *Nat Commun* 2020;**11**:730.
- MacDonald JR, Ziman R, Yuen RKC, et al. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* 2014;**42**:D986–92.
- 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature* 2015;**526**:68–74.
- Collins RL, Brand H, Karczewski KJ, et al. A structural variation reference for medical and population genetics. *Nature* 2020;**581**:444–51.

27. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;**581**:434–43.
28. Garrison E, Sirén J, Novak AM, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol* 2018;**36**:875–9.
29. Ebert P, Audano PA, Zhu Q, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* 2021;**372**:eabf7117.
30. Sirén J, Monlong J, Chang X, et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* 2021;**374**:abg8871.
31. Nicholas TJ, Cormier MJ, Quinlan AR. Annotation of structural variants with reported allele frequencies and related metrics from multiple datasets using SVAFootnote. *BMC Bioinformatics* 2022;**23**:490.
32. Chen S, Francioli LC, Goodrich JK, et al. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. *bioRxiv* 2022. <https://doi.org/10.1101/2022.03.20.485034>.
33. Raphael BJ, Dobson JR, Oesper L, Vandin F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med* 2014;**6**:5.
34. McGranahan N, Swanton C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell* 2015;**27**:15–26.
35. Ricketts C, Seidman D, Popic V, et al. Meltos: multi-sample tumor phylogeny reconstruction for structural variants. *Bioinformatics* 2019;**36**:1082–90.
36. Fan X, Zhou W, Chong Z, et al. Towards accurate characterization of clonal heterogeneity based on structural variation. *BMC Bioinformatics* 2014;**15**:1–12.
37. Gong T, Hayes VM, Chan EKF. Detection of somatic structural variants from short-read next-generation sequencing data. *Brief Bioinform* 2020;**22**:bbaa056.
38. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. *Nat Commun* 2015;**6**:1–12.
39. Tarabichi M, Salcedo A, Deshwar AG, et al. A practical guide to cancer subclonal reconstruction from DNA sequencing. *Nat Methods* 2021;**18**:144–55.
40. Navin N, Kendall J, Troge J, et al. Tumour evolution inferred by single-cell sequencing. *Nature* 2011;**472**:90–4.
41. Vitak S, Torkenczy K, Rosenkrantz J, et al. Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat Methods* 2017;**14**:302–8.
42. Gonzalez-Pena V, Natarajan S, Xia Y, et al. Accurate genomic variant detection in single cells with primary template-directed amplification. *Proc Natl Acad Sci U S A* 2021;**118**:e2024176118.
43. Laks E, McPherson A, Zahn H, et al. Clonal decomposition and DNA replication states defined by scaled single-cell genome sequencing. *Cell* 2019;**179**:1207–1221.e22.
44. Funnell T, O'Flanagan CH, Williams MJ, et al. Single-cell genomic variation induced by mutational processes in cancer. *Nature* 2022;**612**:106–15.
45. Ha G, Roth A, Lai D, et al. Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Res* 2012;**22**:1995–2007.
46. Garvin T, Aboukhalil R, Kendall J, et al. Interactive analysis and assessment of single-cell copy-number variations. *Nat Methods* 2015;**12**:1058–60.
47. Bakker B, Taudt A, Belderbos ME, et al. Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. *Genome Biol* 2016;**17**:115.
48. Wang X, Chen H, Zhang NR. DNA copy number profiling using single-cell sequencing. *Brief Bioinform* 2018;**19**:731–6.
49. Nilsen G, Liestøl K, Van Loo P, et al. Copynumber: efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* 2012;**13**:591.
50. Wang R, Lin D-Y, Jiang Y. SCOPE: a normalization and copy-number estimation method for single-cell DNA sequencing. *Cell Syst* 2020;**10**:445–452.e6.
51. Zaccaria S, Raphael BJ. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. *Nat Biotechnol* 2021;**39**:207–14.
52. Mallory XF, Edrisi M, Navin N, Nakhleh L. Methods for copy number aberration detection from single-cell DNA-sequencing data. *Genome Biol* 2020;**21**:208.
53. Sanders AD, Falconer E, Hills M, et al. Single-cell template strand sequencing by strand-seq enables the characterization of individual homologs. *Nat Protoc* 2017;**12**:1151–76.
54. Hanlon VCT, Chan DD, Hamadeh Z, et al. Construction of strand-seq libraries in open nanoliter arrays. *Cell Rep Methods* 2022;**2**:100150.
55. Sanders AD, Meiers S, Ghareghani M, et al. Single-cell analysis of structural variations and complex rearrangements with tri-channel processing. *Nat Biotechnol* 2020;**38**:343–54.
56. Porubsky D, Sanders AD, Taudt A, et al. breakpointR: an R/Bioconductor package to localize strand state changes in strand-seq data. *Bioinformatics* 2020;**36**:1260–1.
57. Porubsky D, Garg S, Sanders AD, et al. Dense and accurate whole-chromosome haplotyping of individual genomes. *Nat Commun* 2017;**8**:1293.
58. Mitchell E, Spencer Chapman M, Williams N, et al. Clonal dynamics of haematopoiesis across the human lifespan. *Nature* 2022;**606**:343–50.
59. Campbell PJ, Stephens PJ, Pleasance ED, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* 2008;**40**:722–9.
60. Kuipers J, Jahn K, Beerenwinkel N. Advances in understanding tumour evolution through single-cell sequencing. *Biochim Biophys Acta Rev Cancer* 2017;**1867**:127–38.
61. Patel AP, Tirosh I, Trombetta JJ, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 2014;**344**:1396–401.
62. Müller S, Cho A, Liu SJ, et al. CONICS integrates scRNA-seq with DNA sequencing to map gene expression to tumor sub-clones. *Bioinformatics* 2018;**34**:3217–9.
63. Fan J, Lee H-O, Lee S, et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Res* 2018;**28**:1217–27.
64. Gao T, Soldatov R, Sarkar H, et al. Haplotype-aware analysis of somatic copy number variations from single-cell transcriptomes. *Nat Biotechnol* 2023;**41**:417–26.
65. Welch JD, Kozareva V, Ferreira A, et al. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* 2019;**177**:1873–1887.e17.
66. Fan J, Salathia N, Liu R, et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat Methods* 2016;**13**:241–4.
67. Elyanow R, Zeira R, Land M, Raphael BJ. STARCH: copy number and clone inference from spatial transcriptomics data. *Phys Biol* 2021;**18**:035001.

68. Wu C-Y, Lau BT, Kim HS, et al. Integrative single-cell analysis of allele-specific copy number alterations and chromatin accessibility in cancer. *Nat Biotechnol* 2021;**39**:1259–69.
69. Nikolic A, Singhal D, Ellestad K, et al. Copy-scAT: deconvoluting single-cell chromatin accessibility of genetic subclones in cancer. *Sci Adv* 2021;**7**:eabg6045.
70. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods* 2017;**14**:975–8.
71. Chen H, Albergante L, Hsu JY, et al. Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nat Commun* 2019;**10**:1903.
72. McCarthy DJ, Rostom R, Huang Y, et al. Cardelino: computational integration of somatic clonal substructure and single-cell transcriptomes. *Nat Methods* 2020;**17**:414–21.
73. Jeong H, Grimes K, Rauwolf KK, et al. Functional analysis of structural variants in single cells using strand-seq. *Nat Biotechnol* 2023;**41**:832–44.
74. Gao R, Davis A, McDonald TO, et al. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat Genet* 2016;**48**:1119–30.
75. Luquette LJ, Miller MB, Zhou Z, et al. Single-cell genome sequencing of human neurons identifies somatic point mutation and indel enrichment in regulatory elements. *Nat Genet* 2022;**54**:1564–71.
76. Maher CA, Kumar-Sinha C, Cao X, et al. Transcriptome sequencing to detect gene fusions in cancer. *Nature* 2009;**458**:97–101.
77. Heyer EE, Deveson IW, Wooi D, et al. Diagnosis of fusion genes using targeted RNA sequencing. *Nat Commun* 2019;**10**:1388.
78. Loh P-R, Genovese G, Handsaker RE, et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* 2018;**559**:350–5.
79. Halldorsson BV, Eggertsson HP, Moore KHS, et al. The sequences of 150,119 genomes in the UK biobank. *Nature* 2022;**607**:732–40.
80. Wheeler MM, Stilp AM, Rao S, et al. Whole genome sequencing identifies structural variants contributing to hematologic traits in the NHLBI TOPMed program. *Nat Commun* 2022;**13**:7592.
81. Popic V, Rohlicek C, Cunial F, et al. Cue: a deep-learning framework for structural variant discovery and genotyping. *Nat Methods* 2023;**20**:559–68.
82. Liu Y, Huang Y, Wang G, Wang Y. A deep learning approach for filtering structural variants in short read sequencing data. *Brief Bioinform* 2020;**22**:bbaa370.
83. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng* 2010;**22**:1345–59.
84. Wong SC, Gatt A, Stamatescu V, et al. Understanding data augmentation for classification: when to warp? *arXiv* 2016; [10.48550/arXiv.1609.08764](https://arxiv.org/abs/1609.08764).
85. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 2015;**12**:931–4.
86. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* 2016;**26**:990–9.
87. Chen KM, Wong AK, Troyanskaya OG, Zhou J. A sequence-based global map of regulatory activity for deciphering human genetics. *Nat Genet* 2022;**54**:940–9.
88. Zhou J, Theesfeld CL, Yao, et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet* 2018;**50**:1171–9.
89. Liao W-W, Asri M, Ebler J, et al. A draft human pangenome reference. *Nature* 2023;**617**:312–24.
90. Wang T, Antonacci-Fulton L, Howe K, et al. The Human Pangenome project: a global resource to map genomic diversity. *Nature* 2022;**604**:437–46.
91. Weinstein JN, Collisson EA, Mills GB, et al. The cancer genome atlas Pan-cancer analysis project. *Nat Genet* 2013;**45**:1113–20.
92. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* 2020;**578**:82–93.
93. Ebler J, Ebert P, Clarke WE, et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat Genet* 2022;**54**:518–25.
94. The Computational Pan-Genomics Consortium, Marschall T, Marz M, et al. Computational pan-genomics: status, promises and challenges. *Brief Bioinform* 2016;**19**:118–35.
95. Du C. *Identifying Structural Variants from Single-Cell DNA Sequencing Data*. Princeton University Senior Theses, 2020, <http://arks.princeton.edu/ark:/88435/dsp01pn89d965j>.