Liu *et al. Genome Biology* (2022) 23:68
https://doi.org/10.1186/s13059-022-02636-8

Genome Biology

REVIEW

Open Access

Towards accurate and reliable resolution of structural variants for clinical diagnosis

Zhichao Liu[1], Ruth Roberts[2,3], Timothy R. Mercer[4,5,6], Joshua Xu[1], Fritz J. Sedlazeck[7*] and Weida Tong[1*]

*Correspondence: fritz.sedlazeck@bcm.edu; Weida.Tong@fda.hhs.gov
[1] National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR 72079, USA[7] Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA
Full list of author information is available at the end of the article

Abstract

Structural variants (SVs) are a major source of human genetic diversity and have been associated with different diseases and phenotypes. The detection of SVs is difficult, and a diverse range of detection methods and data analysis protocols has been developed. This difficulty and diversity make the detection of SVs for clinical applications challenging and requires a framework to ensure accuracy and reproducibility. Here, we discuss current developments in the diagnosis of SVs and propose a roadmap for the accurate and reproducible detection of SVs that includes case studies provided from the FDA-led SEquencing Quality Control Phase II (SEQC-II) and other consortium efforts.

Introduction

Structural variants (SVs) are typically defined as genetic variants of greater than 50 base pairs (bp) in length and include insertions, deletions, duplications, and chromosome rearrangements [1]. Advances in sequencing technologies are constantly driving the improvements in SV identification, enhancing our understanding of the complex interplay between genetic makeup and associated phenotype [2]. Significant progress has been made in unraveling the importance of SVs in disease etiology, population genetic evolution, ethnic diversity, and gene expression regulation [3]. For example, the role of SVs in the key mutational process of various cancer types has been discovered including that rearrangements delete, amplify, or re-order large genomic segments [4–6]. Furthermore, SVs contribute to phenotypic diversity in neurological and rare diseases such as developmental disorders [7], autism spectrum disorders (ASDs) [8, 9], and schizophrenia [10]. Therefore, SVs have great potential in precision medicine via an understanding of SV patterns at the population level [11–15] and/or via implementation of pharmacogenomics (PGx) biomarkers [16].

Although great progress is being made in detecting SVs in the human genome and correlating this with phenotypic impact, accurately and precisely identifying SVs in specific samples and/or across samples is challenging [3]. The pressing challenges in SV calling are multifactorial, comprising SV types (insertion/deletion/re-order), size (where

the allele often exceeds standard NGS technologies read length), genomics technologies (with differing abilities to identify variations in repetitive regions), and SV calling algorithms (which might use different thresholds and heuristics) [1, 3, 17].

SVs are responsible for more nucleotide changes than any other genetic variants in humans and other species. Many technologies have been developed to identify different types of SVs in the past 15 years, from cytogenetic-based detection (e.g., Karyotyping), array-based technologies (e.g., SNParray and FISH), short-read high throughput sequencing (e.g., NovaSeq), and linked-read sequencing (e.g., 10X Genomics Chromium Technology) to long-read sequencing (e.g., PacBio and Nanopore) [1, 18, 19]. Accordingly, numerous SV-calling algorithms have been developed specific to each technology [3]. For example, approximately 80 SV calling tools are available for short-read whole-genome sequencing (WGS) alone. Given these developments, how can we improve the detection of SVs with sufficient accuracy and precision for use in clinical diagnosis?

The accurate and reproducible detection of SVs underpins reliable clinical implementation and reduces the chance of misdiagnosis. Similar to the quality control steps made for SNVs and small indels [20, 21], initial steps have been made for SV detection [3, 22–26] with an effort to standardize (i) procedures to reduce false positives and false negatives through benchmark call set development (e.g., Genome in a Bottle (GIAB), Human Genome Structural Variation Consortium (HGSV)); (ii) high-confidence calling region establishment (e.g., GIAB), (iii) reproducible SV calling assessment, and (iv) reporting standards (e.g., GA4GH) and best practice guideline recommendations that are ongoing, led by large consortiums and government agencies [22, 23, 27].

The Sequencing Quality Control Phase II (SEQC-II), led by the U.S. FDA, is the most current initiative to develop actionable best practice for sequencing data analysis. The aim is to define reproducible and accurate genetic variant calling to facilitate the development and regulation of precision medicine in clinical practice [28–32]. In this initiative, multi-platform and multi-lab sequencing data of the standard reference materials was carried out, encompassing the broad spectrum of wet lab factors (e.g., library preparation and gene capture strategy). The data generated is allowing the identification of key factors that impact SV detection, such as calling algorithms, SV size and type, and genomics technologies resulting in high-quality calling sets for further application.

In this perspective, we discuss findings from these consortiums and identify gaps where current approaches for SV calling may be suboptimal, as well as propose potential solutions. Our analysis highlights three key components essential for an accurate and reproducible SV detection (Fig. 1): (1) characterization of standard reference materials, (2) determination of sequencing technology to improve SV detection, and (3) establishment of high-quality calling sets. We discuss the critical aspects of each component and propose potential solutions based on examples drawn from SEQC-II and other consortia. We also highlight the opportunity provided by artificial intelligence (AI) to potentially improve the accuracy of SV calling and propose specific deep learning solutions for the future.

Characterization of standard reference materials

Reference standards can be used to measure the false-positive and false-negative rate of SV calling [33]. Specifically, a reference sample can be used to evaluate the

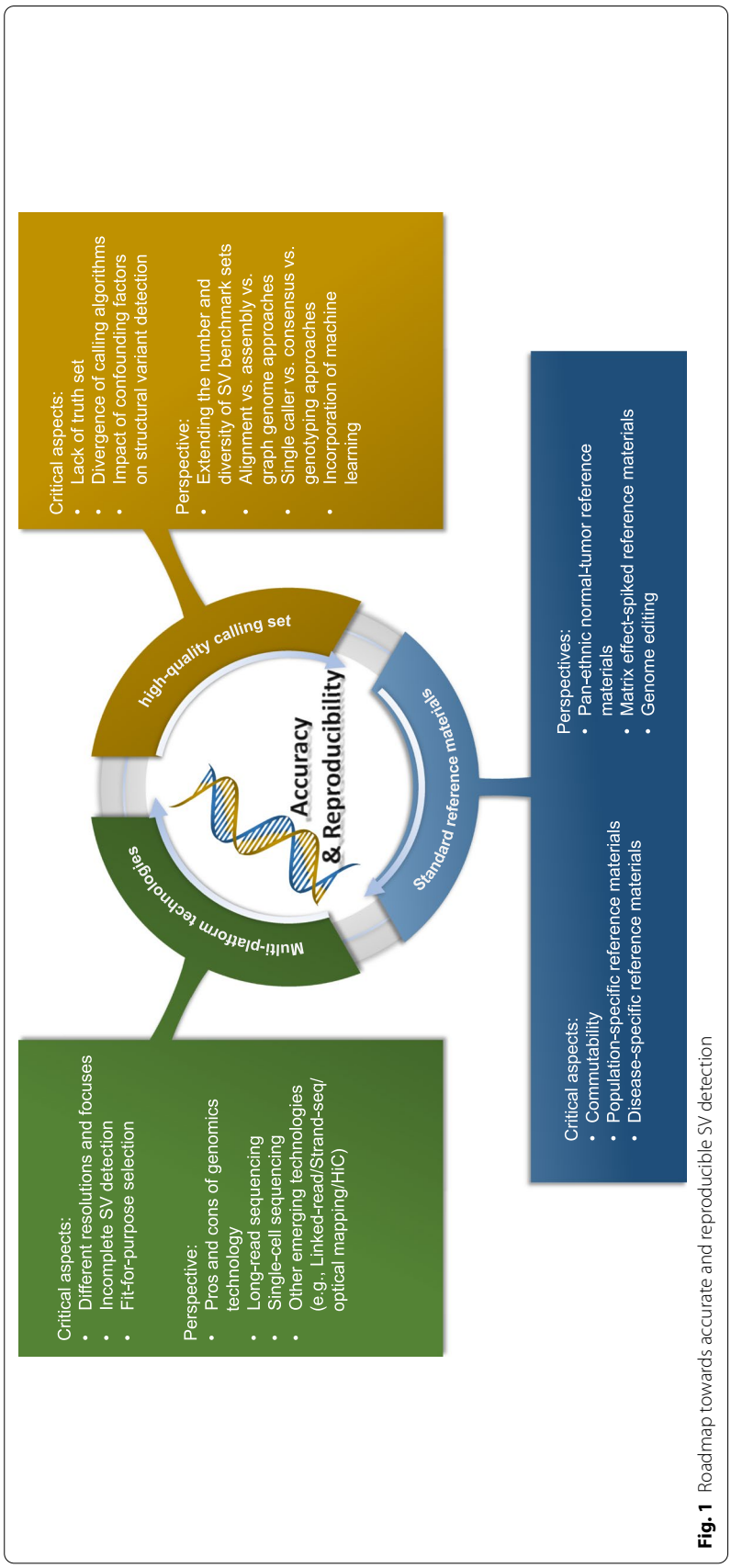

Fig. 1 Roadmap towards accurate and reproducible SV detection

accuracy and reproducibility of sequencing technologies, prioritize confounding factors that contribute to systematic errors, and establish best practice for SV detection. The use of a reference standard that includes well-characterized genetic material or synthetic spike-in control to calibrate sequencing platforms and measure the capability of SV detection is broadly supported [33, 34]. Well-characterized and broadly available reference samples are the foundation for assessing SV calling accuracy and reproducibility, as well as for understanding potential biases in data preparation or analysis. For quality control in the detection of SVs, reference materials need to be relevant for a wide range of SVs and variant allele frequencies (VAFs) to enable a comprehensive assessment [18]. The SEQC-II consortium established several working groups to characterize reference samples and then perform high-depth multi-platform sequencing across different laboratories to make progress in these areas, aiming to facilitate the qualification and validation of genetic variant detection and to promote reproducible science (Fig. 2).

Germline benchmarking

When reference materials are designed to improve assay development for inherited disease diagnosis, patient-parent trio samples offer more sensitive identification of the de novo SVs and any underlying predisposition mechanism. Examples of such trio-based reference materials (e.g., Reference Material 8392) were employed by the National Institute of Standards and Technology (NIST) [22] and the Human Genome Structural Variation Consortium [14] for the evaluation of SV detection.

Unlike these two working groups focusing on somatic mutations, the SEQC-II germline mutation working group focuses on a comprehensive and reproducible assessment of the detection of germline variants based on a trio sample set. For this, sample materials are the HapMap trio from the 1000 Genome project [37] and HG001 for the Genome in a Bottle (GIAB) led by the NIST [21]. Notably, Chinese Quartet reference DNA materials derived from normal B-lymphocyte cell line and blood samples are included, enabling the comparison of SV and SNV detection in different biological matrices and ethnicities [36, 38, 39]. Furthermore, the ABRF NGS phase 2 DNA-seq study leverages RMs (National Institute of Standards and Technology (NIST) RM 8392, known as the Ashkenazi trio; mother (HG004), father (HG003), and son (HG002), a family trio consented through the Personal Genome Project (PGP)) [40] to provide insight into currently popular sequencing instruments. Inter-laboratory and intra-laboratory DNA-seq replicates of the Ashkenazi trio are analyzed, as well as three individual bacterial strains and a metagenomic mixture of ten bacterial species to study the effects of GC content and library complexity [41]. These are novel and important studies since they not only assess like other studies the ability to call SV or SNV across certain data sets, but really dive into the question what causes the variability between different sequencing events. This is especially important when sequencing multiple samples (e.g., from a large cohorts) across different sequencing centers. All three studies revealed detail insights in the contribution of sequencing centers themselves including library preparations but further investigate the impact of different filtering (deduplication, QV recalibrations) and analytical approaches [36, 38, 39].

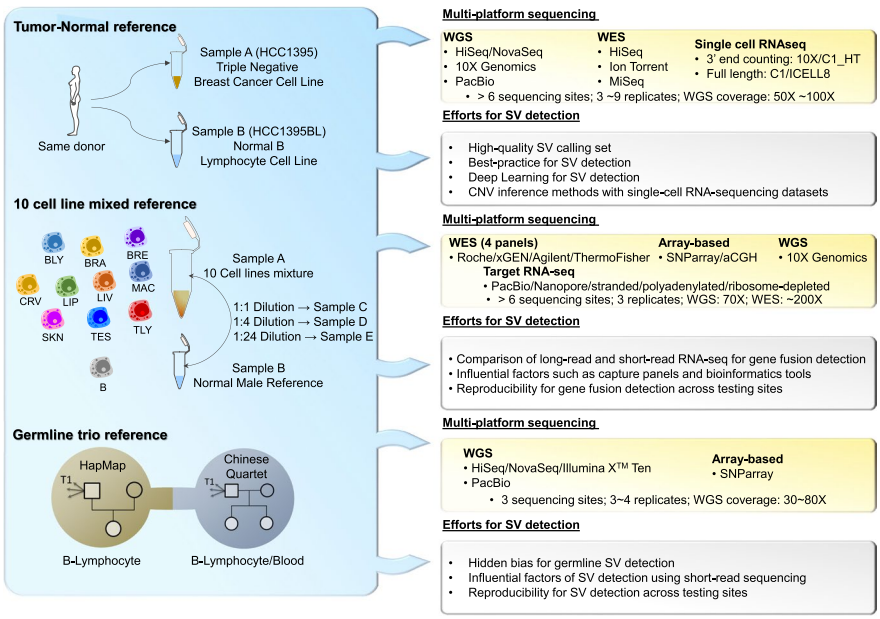

Fig. 2 An insight into the reference samples and efforts for SV detection by the SEQC-II consortium. Great strides have been made in advancing SV detection by the SEQC-II consortium (Fig. 2). First, the SEQC-II consortium established high-quality SV calling sets based on multi-platform sequencing of tumor- normal reference samples and partially verified this using orthogonal methods, including PCR-based validation, cytogenetic array BioNano optical mapping, as well as fusion gene detected from RNA-seq [35]. Meanwhile, SEQC-II systematically evaluated the reproducibility of somatic SV detections across platforms and benchmarked the performance of various software tools. Leveraging the developed high-quality SV calling sets, they developed a deep learning-based calling algorithm for SV detection using the convolutional neural network (CNN). The proposed deep learning models achieved high robustness across multiple sequencing technologies for fresh and FFPE DNA input, varying tumor/normal purities, and different coverages, with significant superiority over conventional detection approaches in general, as well as in challenging situations such as low coverage, low variant allele frequency, DNA damage, and complex genomic regions. Furthermore, the CNV inference method was developed based on the generated single-cell RNAseq data. Second, the SEQC-II consortium comprehensively investigated the performance and confounding factors (i.e., long or short-read sequencing, capture panels, and bioinformatics pipelines) of gene fusion detection [28]. It was found that long-read sequencing achieved higher precision and discovered more novel fusion genes. Short-read sequencing achieved greater sensitivity for detecting known fusion genes correlated with the endogenous expression of targeted genes. Third, the SEQC-II consortium prioritized SV detection divergent sources by using multiple illumina-based short-read sequencing of the Chinese quartet reference samples. Interestingly, mapping methods are significant resources of calling variability, followed by sequencing centers and replicates. Surprisingly, SV supported by only one site or technical replicate often represented true positives defined by long-read PacBio sequencing, consistent with an overall higher false-negative rate for SV calling [36]

Somatic benchmarking

It is apparent that this is yet to be solved as many samples are "only" healthy cells derived from limited ethnic backgrounds. Especially in cancer, scientists are often interested in detangling complex SV that are chained, including virus insertions, somatic vs. germline mutations, low-frequency mutations escaping traditional diploid genomic patterns or even circularized DNA are not present in any of the reference materials yet. Several attempts have been made to develop reference material for somatic variants identification and NGS technology calibration in clinical laboratories. Specifically, the genome editing technologies such as CRISPR/Cas9 were adopted to biologically engineer mutation clones in mammalian cells [42, 43]. However, these

reference materials could not be used to benchmark whole genome-based technologies and cover a broad spectrum of structural variant types. Additionally, Approaches to simulated sequence alignment data were also established to evaluate the variant calling algorithms for SNV and SVs [44]. Also, the two normal cell lines were titrated at various concentrations to mimic tumor heterogeneity. A normal cell line with engineered mutations or synthetic DNA spike-in is difficult, if not impossible, to engineer somatic mutations in a whole genome scale to mimic the complexity and heterogeneity of a cancer genome. Therefore, whether these reference materials could capture and represent patient tumor samples regarding VAF, inter- and intratumor heterogeneity, prevalent copy number alterations (CNAs), and complex chromosomal rearrangements are largely questionable about biased inferences and highlight the need for whole genome-based reference materials.

One approach could be to use tumor-normal or mixed cancer cell lines samples with different tumor purity and heterogenicity that may be closer to actual cancer patients. Cancer genomics aims to identify somatic (tumor-specific) variants with potential diagnostic, prognostic, and therapeutic implications and to detect germline variants with inherent information for both patients and their families. Although substantial clinical tumor testing does not currently involve analysis of a matched germline sample due to cost and time to delivery, concerns have been raised that the detected genetic variants are truly somatic variants [45]. Moreover, consortiums such as NHGRI/NCI Clinical Sequencing Exploratory Research Consortium Tumor Working Group and the American College of Medical Genetics and Genomics (ACMG) released a set of guidelines recommending that laboratories performing cancer sequencing tests should include germline variants [46, 47]. The International Cancer Genome Consortium used tumor-normal sample pairs from two different types of cancer, chronic lymphocytic leukemia (CLL), and medulloblastoma (MB), for a comprehensive assessment of somatic variants identification, emphasizing on the imperativeness of real, not simulated, mutations are more helpful in dissecting performance of mutation callers on establishing whole-genome somatic variant signatures incredibly complex SVs [48]. However, the ratio of DNA amount between tumor and matched control in the reference materials developed by ICGC is predefined with low mutation burden and minimal structural changes, limiting its utility to assess the detection limit of various genomics technologies.

Specifically, the SEQC-II somatic mutation working group developed a tumor-normal reference sample using a normal B-lymphocyte cell line (i.e., HCC1395BL) and a triple-negative breast cancer cell line (i.e., HCC1395) from the same donor purchasable through the American Type Culture Collection (ATCC) [35]. The HCC1395 cell line has been characterized with conventional genomics approaches such as cytogenetic analysis [49] and array-based comparative genomic hybridization [50], consisting of rich genetic variant types including ~40,000 SNVs, ~2000 small indels, CNAs covering over 50% of the genome, more than 250 complex genomic rearrangements [51], and an aneuploid genome and BRCAness [52]. Moreover, the HCC1395 DNA was pooled with HCC1395BL DNA at different ratios to create a range of admixtures that mimicked tumor purity levels of 100%, 75%, 50%, 20%, 10%, 5%, and 0%. Furthermore, this tumor-normal reference sample set may also mimic different biospecimen types (i.e., fresh vs.

formalin-fixed, paraffin-embedded) for investing the effect of fixatives and sample handling on variant detection.

While large whole-exome and whole-genome sequencing studies are capable of providing new insights of genetic variants into cancers, clinical adoption of these approaches is still lagged and is not routinely offered by clinical laboratories [53]. In contrast, large, low-cost, and short turnaround time targeted cancer panels have been widely utilized in clinical laboratories. Encouragingly, some panel-based NGS have been approved or cleared by regulatory agencies such as U.S. FDA to promote precision medicine (https://www.fda.gov/medical-devices/in-vitro-diagnostics/list-cleared-or-approved-companion-diagnostic-devices-in-vitro-and-imaging-tools). To enhance the clinical application of panel-based sequencing for cancer diagnosis, we suggest it is indispensable to establish reliable, robust, continuous, and generally available genomics reference samples for assessing and calibrating different NGS assays.

VAF of somatic mutation (e.g., VAF < 20%) is far less than the VAF in germline cells (e.g., ~50% and ~100%). It is technically challenging to utilize normal cell lines to establish a reference material covering a wide range of VAF magnitudes. Commonly available commercial reference samples typically consist of the limited number of genes and variants with distinct VAF ranges, hampering its utility for benchmark NGS assays in genetic variants with lower VAF (< 2~5%) and resulting in the overall variant detection performance is inversely correlated to the VAF of the analytes targeted (https://www.horizondiscovery.com/reference-standards/type/oncospan). To overcome the shortcoming, the SEQC-II Oncopanel working group synthesized sample A by equal mass pooling gDNA of ten cancer cell lines that were used to make the Universal Human Reference RNA (UHRR) (Catalog #740000, Agilent Technologies) to ensure the wide coverage of actionable cancer genes under different VAF magnitudes [28]. Sample A permitted the investigation of 40,000 variants down to 1% allele frequency with more than 25,000 variants having less than 20% allele frequency with 1653 variants in COSMIC-related genes, which is 5~100× more than existing commercially available samples. Also, a cell line derived from a normal male individual (Agilent OneSeq Human Reference DNA, PN 5190–8848) (termed "Sample B") was characterized and employed as a negative control for somatic variants but also generating different genomic backgrounds. By titrating sample A into a control sample B with varying ratios of mixing (i.e., 1:1/4/24/124), samples C, D, E, and F were created to mimic the somatic mutation noted in extremely low VAF (as low as 0.02%), a range suitable for assessing liquid biopsy panels for detecting mutations in circulating tumor DNA [54].

Commutability, defined as the ability of reference materials to perform comparably to actual patient samples, is one of the most critical parameters in qualifying reference materials [33]. To establish a more objective quality assessment of SV detection, we suggest developing additional biological certified reference materials as follows:

(1) Establish pan-ethnic normal-tumor reference materials. The ethnic diversity of SVs regarding size and types across different genome regions (e.g., rich GC contents or low complexity) in inherited diseases and cancers has been widely reported [23, 55]. Trio-based reference materials from different ethnic backgrounds have been developed, such as trio samples of Han Chinese, Puerto Rican, and Yoruban Nige-

rian ancestries developed by the Human Genome Structural Variation Consortium [14] and the Ashkenazi family trio (HG002, HG003, and HG004) in the GIAB [22]. However, few pan-ethnic normal-tumor reference materials are publicly available and/or purchasable. Therefore, an effort for coordinated, local implementation of the development of normal-tumor reference materials from diverse ancestries or admixtures is highly recommended to improve our understating of somatic SV differences in local populations.

(2) Matrix effect-spiked reference materials. The SV detection limit within formalin-fixed, paraffin-embedded, or liquid biopsy-based samples from patients is much lower than that in normal tissues. Furthermore, some complex effect of fixatives such as in vitro fertilization (IVF) and preimplantation genetic diagnosis (PGD) screening help patients to select embryos free of rare diseases [34]. As the part of efforts from the SEQC-II consortium, we proposed synthetic internal standards (IS) and methods for better control for technical error in NGS in assessment of circulating tumor DNA specimens, enabling measurement of low AF mutation not detected by current practices. The proposed Synthetic spike-in IS could be effective way to mimic the complex effect of fixatives and evaluate the NGS testing [56].

(3) Use genome editing for spiked specific SV types in the reference materials. Genome engineering technology with programmable nucleases (e.g., ZFNs, TALENs, and CRISPR/Cas9) has been well-established, enabling precise and efficient genome-editing spiking of specific SV types into cells, and offering the opportunity to develop synthetic reference materials [57]. For example, CRISPR/Cas9-based genome-editing protocols have been developed for the direct generation of deletions, duplications, and inversions of up to one million base pairs in zygotes [58]. However, a close examination of the potential unexpected off-target variants in the engineered cell line should be considered [59].

Determination of sequencing technologies to improve SV detection

Advances in genomics technologies continually improve the resolution of SV detections [1, 18]. From the use of microscopy to visualize karyotypes of short, condensed chromosomes to long-read sequencing to identify complex rearrangements that consist of multiple combinations of SV events, Table 1 summarizes some representative NGS technologies used for SV detection. The ability to more sensitively detect SVs and resolve more complex rearrangements has resulted in an exponential increase in the number and research on SVs identified over the past 15 years. However, the genome's intrinsic complexity, the technical errors introduced during sample preparation, and the limitation of the existing sequencing technologies have left a substantial fraction of SV undetectable and much of their complexity remains hidden.

An increasing body of literature has demonstrated the benefits of employing multi-platform genomics technologies for more comprehensive SV detection across the human genome [14]. Chaisson et al. [14] used a suite of long-read, short-read, strand-specific sequencing technologies, coupled with optimal mapping and SV calling algorithms to provide a complete spectrum of haplotype-resolved SVs in human genomes from three sample trios. Specifically, their multi-platform strategy detected three- to

Table 1 A comparison among different sequencing technologies for structural variant detection

Platform	Read length	Cost	Comments	Run time
Short reads (Illumina)	NovaSeq: up to 250 bp	$	Short-read NGS performs well for >1kb regions. It struggles with shorter CNV detection 50-500bp, and in complex genome regions	NovaSeq: 0.15Tb/day
10X Genomics Chromium	Up to ~100 kb	$$	Sparse sequencing rather than true long reads; more complicated to align, with poorer resolution of locally repetitive sequences. However, 10X Genomics Chromium is currently discontinued	-
PacBio SMRT sequencing	10–15 kb (average) and up to 100 kb	$$$	HiFi: long reads (10-20kbp) of high fidelity having a similar error rate as Illumina. CLR: Longer raw reads have high error rates dominated by false insertions; requires new alignment and error correction algorithms	20 Gb/day
Oxford Nanopore	averaging ~10 kb and up to 2 Mb	$$$	Raw reads have ~5% error rates dominated by false deletions and homopolymer errors; often requires new alignment and error correction algorithms	A MinION Flow Cell : ~ 25 Gb/day
Hi-C-based analysis	<100 bp	$$	Sparse sequencing with highly variable genomic distance between pairs (1 kb to 1 Mb or longer); Detection may result from random chromosomal collisions Less than 1% of DNA fragments actually yield ligation products. Due to multiple steps, the method requires large amounts of starting material	Whole analysis within 28 hours
BioNano Genomics optical mapping	~250kb or longer	$	Limited algorithms to discover high-confidence alignment between an optical map and a sequence assembly	100x coverage of 3 human genomes is collected in less than 6 hours

seven-fold more SVs than most standard high-throughput sequencing studies. Moreover, more inversions located within critical genome regions were discovered, associated with rare recurrent microdeletions and microduplication syndromes [60]. Despite these advantages, the use of multiple sequencing technologies to resolve SVs

may be impractical. Furthermore, are there certain biases introduced when sequencing the same sample multiple times?

The SEQC-II adopted multi-platform and multi-lab designs for a comprehensive assessment of reproducibility and accuracy of the detection of SVs. The SV working group set out to investigate the reproducibility and variability of SV calls when the sample was sequenced across multiple sequencing instruments or in different laboratories [31, 32]. Interestingly, current approaches produced a level of variability often associated with false negatives (i.e., missed SVs) in SV calls with current methodologies. The somatic mutation working group used the developed tumor-normal materials coupled with multi-platform sequencing technologies, including short/long/linked-read sequencing and high-throughput chromosome conformation capture (HiC). Moreover, they also performed multi-platform, single-cell RNA sequencing technologies to establish best practice for single-cell RNAseq analysis [61]. The Oncopanel working group employed four commercialized WES panels with multiple library preparations and 10X Genomics linked-read sequencing for the individual cell lines to generated high-coverage sequence data on the developed reference samples. SVs reported by linked-read sequencing data were then compared with gene fusion events detected in RNA sequencing data of UHRR. Meanwhile, the linked-based WGS and array-based SNParray/aCGH data were also generated for investigational and confirmation purposes. The germline working group utilized most Illumina-based short-read sequencings such as HiSeq 2000, NovaSeq to XTen, and the Chinese Quartet were also sequenced using long-read PacBio.

Although the promise of multi-platform genomics technologies for complete SV detection is apparent, it may be more appropriate to use a combination of different sequencing technologies; insight into their different strengths and weaknesses is key for effective deployment. The pros and cons of different genomics technology for SV detection were discussed extensively elsewhere [1, 17, 62]. Here, we focus on key points in optimizing the selection of multi-platform technologies for enhancing the accuracy and comprehensiveness of SV detection.

First, a multi-technology approach is not scalable given the costs and DNA, cell, or sample requirements such as quantity and quality for the clinical samples. One technology that may enable more accurate SV detection is known as long reads [3, 63]. Despite the fact that long reads may have a higher error rate (HiFi: ~0.1–1%, ONT: 3–8%), they have shown remarkable performance and resolution of SVs across the genome. In contrast to short reads, they often identify almost twice as many SVs, many of which include novel sequences (i.e., insertions). Furthermore, long-read sequencing significantly reduces the overall false discovery rate seen with short reads (e.g., translocations that shadow repeat expansions [63]). However, long-read sequencing remains costly, and the DNA requirements (quality and quantity) are not always achievable. Furthermore, the HGSV and others have also demonstrated the limitations of long reads, such as accessing and recovering certain complex regions of the genome (e.g., centromeres or telomers) or other complex SV types where more targeted essays might perform better (e.g., StrandSeq [64]).

Second, given the trend of reducing sequencing costs and improving yield as well as error rate, multiple long-read sequencing projects are underway from the assembly of individual genomes through the study of the genomic architecture of individual human

cells [65]. This trend is likely to continue. However, will it replace short-read sequencing? The answer is a tentative no, given the scalability, the DNA requirements, and the costs.

There are multiple short read-based assays currently available. Linked reads constitute a fast and inexpensive method that provided information across a long molecule (e.g., 100kbp), leading to improvements in mappability and phasing. In theory, this approach could improve SV calling itself, but the software for achieving this was limited. Other interesting concepts such as HiC and StrandSeq also show promises. StrandSeq enables the accurate detection of inversions but requires laborious preparation, and there is not yet a standardized kit available. However, both technologies are limited by the mapping into repeats given their read length of 100–150 paired-end reads. Nevertheless, a combination of StrandSeq and PCR-free WGS may improve the detection of SVs at a cost and sample requirement that outcompetes long reads.

Currently, WGS remains the workhorse of genomics, with million genomes sequenced every year. Therefore, we need to establish robust pipelines, ensure technical reproducibility, and understand the risk of bias. SEQC-II has so far approached this by sequencing samples across multiple centers to improve our understanding of potential variabilities and the impact on SV detection [66].

Establishment of high-quality SV call sets

The evolution of genomics technologies has also resulted in the proliferation of different SV calling algorithms. Over 85 publicly available SV calling algorithms have been developed for different NGS data types [26]. These algorithms aim to identify the divergence between the reference genome and the sample reads, which examine the following read features or combinations: read-pair, read-depth, split-read, and de novo or local assembly [1]. The motivation behind continually producing new SV calling algorithms is to improve on previous shortcomings, resulting in better precision and recall rate, speed and user-friendliness, and handling specific SV types. To establish newly developed SV calling algorithms, researchers compared outcome with previous calling algorithms using simulated and actual NGS data and highlighted their superiority from a particular perspective. Consequently, downstream users have little to guide them when choosing the best "fit-for-purpose" algorithms since they find every algorithm claims to be the right option. As a result, community efforts comparing the SV calling algorithms are being carried out to determine relative advantages and disadvantages and suggest best-practice for SV algorithm selection [25, 26].

This work started over a decade ago. SV calling is improving but has by no means reached the reliability and quality of SNV calling. While multiple reasons may attribute to this problem, SV alleles are longer and often more complex than SNV which can be contained in a single-short read. Nevertheless, what steps can be taken to improve the field further or spark a new evolution of methodologies and approaches? Here we highlight some potential directions and discuss their promises. These are (1) extending the number and diversity of SV benchmark sets, (2) alignment vs. assembly vs. graph genome approaches, (3) single caller vs. consensus vs. genotyping approaches to improve SV precision, and (4) incorporation of machine learning and AI.

Extending the number and diversity of SV benchmark sets

Over the past few years, multiple methods have been proposed and implemented to simulate SV (e.g., VarSim [67], SURVIVOR [68], etc.) and simulate read data (e.g., NanoSim [69], PBsim [70], etc.) [71]. While these methods are helpful for rapid, early understanding of the utility of an SV caller, they often under- represent the complexity of SVs either at the level of the allele itself or in the regions they tend to occur (e.g., repetitive). Therefore, they are no replacement as yet for high-quality benchmark sets such as HGSV or GIAB, comprising by now 10–20 benchmark sets across multiple individuals and ethnicities. Nevertheless, these SV benchmark sets range in quality and accessibility and thus also in their ability to be leveraged for benchmarking SV callers. For example, for SV, NA12878 from GIAB is an older and lower quality benchmark set compared with the HG002-4 trio, mainly because NA12878 has multiple redundant SV calls and imprecisions. HG002-4, however, is of high-quality spanning over 90% of the human genome, but is only available for hg19 at the moment [22]. HGSV also produced high-quality assemblies, leveraging multiple technologies over the past years. Having the latest releases will improve the phasing, continuity, and accuracy of their assemblies and thus also the SV calls themselves. These samples are derived from different genders and a few different ethnicities, but so far do not cover all ethnicities and are missing disease phenotypic benchmark genomes. At the time of writing the review, no one from our knowledge have done a head-to-head comparison of the GIAB and HGSV released call set also potentially because both seems to be of very high-quality given all the validation and QC steps done. Both efforts focus mainly on insertion and deletions (especially GIAB) as they represent the majority of SV. HGSV also included inversions, but both are not addressing the biggest issue around translocations or other more complex SV types [72]. Short read SV call sets often report in the access of up to 4000 translocations of which ~50–70% are repeat extensions [63]. In contrast most methods can detect 50–300bp insertions these days (e.g., delly and manta), but fail at larger size ranges. The high-quality calling set generated from SEQC-II tumor-reference samples is potentially the first step in the right direction (Fig. 2). Here, we summarized the data generated from SEQC-II consortium efforts for trigger the community's interest to further explore the potential improvement for SV detection (Table 2).

Furthermore, some cancer cell lines (e.g., SKBR3) have already been heavily sequenced and studied, but no official benchmark SV set is available. Cancerous but also other human disease genomes will be important as these include challenging regions (e.g., oncogene amplification) that are not trivial to resolve. Additionally, the need to include different ethnicities and the potential for novel sequences (e.g., in African Americans) suggests that there is still some way to go. In any case, the current existing benchmark sets are already showing an improvement in SV calling methods and methodologies and are also providing standards for wet lab technology development.

Alignment vs. assembly vs. graph genome approaches

There is often a debate about the best approach to identify SVs, either at scale (multiple hundred to thousand samples) or comprehensively (complex structures within a sample). Mapping (i.e., the alignment of reads to an established reference genome) compared to de novo assembly (i.e., the reconstruction of the sample genome without any template)

Table 2 Benchmark datasets generated from SEQC-II consortium efforts and potential application in SV detection

Working group	Reference samples	Benchmark data	Potential benefit for SV detection	Link
Somatic mutation [30, 32, 66, 73–76]	**Tumor-normal** sample: HCC1395BL as normal and HCC1395 as tumor	*Fresh DNA*: **WGS** - HiSeq, NovaSeq, 10X Genomics, and PacBio) **WES** - Hiseq and Ion Torrent **AmpliSeq** - MiSeq **Microarray** -AffyChip CytoScan HD *FFPE/mixed DNA*: **WGS/WES**: Hiseq *Fresh cells*: **scCNV**: 10X Genomics	• Somatic SV benchmark establishment • Low allelic frequency (LOF) somatic SV detection in liquid biopsy or FFPE samples • Deep learning-based somatic SV detection • Reproducibility and repeatability assessment of somatic SV detection based on multiple sample and design	**All raw data** (FASTQ files): NCBI's SRA database (SRP162370) **VCF and source code**: ftp://ftp-trace.ncbi.nlm.nih.gov/seqc/ftp/release/Somatic_Mutation_WG/ **BAM files**: Seven Bridges' s Cancer Genomics Cloud (CGC) platform and license is needed. **scCNV data**: SRA repository under accession code no. PRJNA504037. **Source code**: https://github.com/oxwang/fda_scRNA-seq and https://codeocean.com/capsule/0497386 or https://doi.org/10.24433/CO.1559060.v1.
Oncopanel [28, 29, 54, 56, 77]	**Sample A**: ten cancer cell line mixture **Sample B**: a normal male cell line (Agilent OneSeq Human Reference DNA, PN 5190–8848) **Spike in samples:** 5% AcroMetrix spikes-ins + Sample B	*8 pan-cancer gene panels*: **WES**: HiSeq, NovaSeq, Ion Torrent, Nanopore, Stranded RNAseq **WGS**: 10X Genomics **Microarrays**: SNP array and aCGH	• Reproducibility and repeatability assessment of actionable somatic SV assessment • Benefit of gene fusion detection by integrating DNAseq and RNAseq	**FASTQ or BAM**: BioProject PRJNA677997 - https://www.ncbi.nlm.nih.gov/bioproject/PRJNA677997. **VCF/BED**: https://figshare.com/projects/SEQC2_Oncopanel_Sequencing_Working_Group_-_PanCancer_panel_Study/94520
Germline mutation [36, 38, 39]	Chinese Quartet samples (B-lymphocyte cell line and blood samples) HapMap samples (HG001)	**WGS**: Hiseq, NovaSeq, illumina X10, PacBio **Microarrays**: SNP array	• Influential factors on reproducibility assessment for germline SV detection • Germline SV detection concordance between B-lymphocyte cell line and blood samples • Deep learning-based somatic SV detection	**Raw data**: BioProject PRJNA723125 (HapMap samples) https://www.ncbi.nlm.nih.gov/bioproject/PRJNA723125/ and NODE OEP001896 (Chinese Quartet Samples) https://www.biosino.org/node/project/detail/OEP001896 **Source code**: https://github.com/justwalking2017/SEQC_WG3_Script
	HapMAP Ashkenazi Trio	**WGS**: HiSeq, BGISEQ, MGISEQ, NovaSeq **WES**: Ion proton and Ion S5	• Cross check the best practice of germline SV detection with NIST efforts	**Raw data**: BioProject PRJNA646948 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA646948), within accessions SRR12898279–SRR12898354 **Source code:**https://www.github.com/jfoox/abrfngs2
	Bacterial genomes (ATCC MSA-3001)	Miseq, Ion PGM, Ion S5, MinION, Flongle, and GenapSys		

often has different advantages and disadvantages [3]. In general, de novo assemblies require more coverage to provide a continuous reconstructed sequence without gaps or uncertainties, while mapping strategies rely on the completeness of the reference genome [3, 17]. While these two particular criteria might sound trivial, they present a challenge with multiple implications for the performance of different approaches. For example, mapping an African sample to the human genome (GRCH38) may result in some unaligned reads due to the absence of reference sequences in the human reference genome. SV caller tries to identify this sequence later on as insertions are often missing (e.g., short reads) or at least missing to reconstruct larger insertions (multiple kbp for long reads). Thus, mapping approaches often succeed in scaling and requiring less expensive sequencing technologies or coverages to succeed. Their disadvantages, however, are often that more complex alleles are hard to resolve and the reconstruction of larger (multiple kbp) novel sequences often fails [3].

Alternatively, for assemblies, there are still risks of over-merging or splitting regions due to high (e.g., immune regions) or low (e.g., LOH) heterogeneity. This is mainly due to the central paradigm of an assembly of when are two genomic regions the same and some sequencing artifact altered them slightly or if these are indeed two different regions. Prominent assembler for short reads (e.g., SPAdes [78]) or even long reads (e.g., Canu [79], hiFiasm [80], Shasta [81]). Furthermore, the current bottleneck is often caused by the accuracy of genomic alignment [82]. Still, de novo assembly is currently the way to establish new benchmark SV sets [83]. The suggestion that de novo assembly will replace mapping for large-scale genomics is currently not foreseen.

Another option that might be able to combine these advantages over time is a graph genome approach. Here multiple genomes (e.g., representing different ethnicities, cancer samples, or other groups) are first assembled into high-quality genomes and then utilized to improve mapping and thus variant detection. As one can imagine, if this graph genome carries a novel sequence or complex SV, it enables the detection of these in the analysis of other samples. However, there are apparent limitations such that there needs to be a balance between comprehensiveness (i.e., including every private allele) and complexity (too many alternative SNVs and SVs making the graph ambiguous). While these problems are being addressed, graph genomes have already shown significant promises for SV calling: Paragraph [84] and VGTool [85]. As such graph genomes are still seen as a novelty and not many robust and established methods exist. Nevertheless, this will likely change in the near future with ever more high-quality assemblies being made available [82].

Single caller vs. consensus vs. genotyping approaches to improve SV precision

The SV detection limit differs across different types (e.g., deletions vs. insertions), sizes (e.g., 50 bp vs. 50 kb), and genomic regions (e.g., repeats). SV calling algorithms were developed based on different heuristics, hypotheses, and thresholds, resulting in a considerable divergence. This is illustrated by comprehensive benchmark papers across SV callers [25, 26]. Both authors found no single SV calling algorithms could perform at the top across different SV types and sizes. The high divergence on the precision and recall rate between the simulated data and actual data highlights the low commutability of simulated data. Long run time and increased memory requirement do not guarantee

better SV detection performance. Importantly, the assembly-incorporating callers such as GRIDSS [86] and Manta [87] outperformed other callers with the actual data, consistent with the conclusion reached by Kosugi et al. [25]. We suggest further investigating whether SV caller performance findings drawn from these studies could be extrapolated to the different sequence depth and tumor purity through SEQC-II multi-platform and multi-lab sequencing data. One aspect to consider is of course also the size of the variants that are of interest. For example, very large alterations of multiple Mbp or even chromosome arms are often better detected by coverage approaches than SV callers (i.e., utilizing alignment signals) themselves [3]. That is often the case since these large CNV events do not have resolved breakpoints. Thus, a CNV approach over coverage or B-allelic (i.e., using SNP called) often performs better to characterize these.

The use of consensus calls generated from multiple SV callers can achieve better precision and recall rate [25, 26]. Nevertheless, unions of all SVs calls across multiple SV calling methods might result in a high false-positive rates than a single caller has despite often having a higher sensitivity [88, 89]. This is as the new SV set will incorporate also the falsely identified SV calls from the individual SV callers. On the other hand, a too restrictive approach to require, for example, three or more SV caller to agree leads often to a low sensitivity but high precision [88, 89]. Thus, the balance is often hard to achieve. Meta-callers, created by combining multiple SV calling algorithms have been proposed [88, 90, 91]. The proposed meta-callers could be mainly divided into rule-based strategies and machine learning with discriminating features based on the adopted ensemble strategy. However, these meta-callers either focus on a few SV-calling algorithms or need a "true set" to train the optimized model, leaving room for further improvement. The primary problem with rule-based strategies is the absence of unified criteria to prioritize SV calling results from different callers. For example, a few meta-callers such as Parliament2 [91] and SVMerge [92] focus on coordinate overlap through soft clip comparison and adjustment with local assembly. Parliament2 incorporates the usage of five pre-defined and optimized SV calling methods and allows users to decide on an optimized SV caller combination through quality score determined by SURVIVOR [68] and genotyping with SpeedSeq [93]. Alternatively, MetaSV adopts a priority-based combination strategy, exemplifying the weight of read pair-based callers over the split read-based ones. For the employed ensemble strategy, a true set was utilized for training SV callers either through machine learning strategies (e.g., CN-Learn [94]) or statistical measurements for each different SV type and size condition (e.g., Parliament2 [91] and FusorSV [88]). All three methods (Parliament2, FusorSV, and MetaSV) either incorporate or require a certain set of SV calling methods over short reads to improve their runtime and consensus performance. Noteworthy is that SURVIVOR offers a generalizable framework; however, this might need to be optimized further for a comparison across different technologies or methodologies (e.g., assembly vs. mapping-based SV calls) as the representation of SV are often different. Thus far, SURVIVOR has been used across different sequencing technologies successfully [95].

Another less-studied area is the use of a SV genotyper after initial SV calling. Here, a SV genotyper uses a VCF with predefined SVs and summarizes the evidence for these SVs across a given sample or bam file. There are currently multiple SV genotyping methods available such as STIX, Paragraph, SVTyper, GraphTyper, and others. Their

performance varies similar to SV callers over sizes and types of SV [89]. Still, one potential benefit is that SV genotypers have, in general, fewer constrains and thresholds to identify SV as they operate on a given list of SVs. Thus, this might be able to improve recall while retaining high precision compared to other consensus methods that rely only on SV discovery methods.

Incorporation of machine learning and AI

Machine learning algorithms especially neural networks have impacted almost every facet of our daily life and have revolutionized genomics [96–99]. For example, deep learning has shown merits in enhancing SNV/small indel detection, outperforming most well-established SNV callers [100]. Inspired by DeepVariant and others, we believe that machine learning has a great potential to improve every step of the proposed roadmap towards accurate and reproducible SV detection.

The fundamental difference between conventional SV calling algorithms and deep learning approaches lies in retrieving the SV information from the BAM file. Traditional methods aim to detect the divergence between the sample sequence and reference at a per read level first. In contrast, deep learning approaches such as convolutional neural networks (CNNs) transform variant detection as a classification problem by considering the BAM file or its genomic regions as an image [100].

A few initial attempts to apply deep learning algorithms for SV detection have been conducted, and some encouraging results were obtained [101–103]. One example is DeepSV, which utilized the CNN model for training the visualized sequence in the 1000 Genomes Project for deletion/detection and yielded better accuracy than the conventional SV callers [102]. The proposed DeepSV could be extended to detect other SV types using the SEQC-II high-quality calling set and NIST germline SV calling set [22]. Another way to utilize the power of machine learning is for the assessment of the quality or accuracy of an SV. It is interesting to note that multiple visualization methods around SV enabled a fast and reliable filtering of SV as a replacement for manual assessments. However traditional SV callers lack this intuition of the human eye or mind. Samplot-ML is a method that combined the visualization of SV and was trained on false calls from SV methods to better distinguish true from false SV candidate calls. It will be interesting to observe future developments that use different machine learning approaches for the field of SV calling, all of which could potentially improve the field. Here, we propose a few deep learning frameworks illustrating the potential of AI to enhance SV detection that is intended as a discussion point within the community (Fig. 3).

Besides SV detection, machine learning and AI may also be a good option for SV pathogenicity prediction to facilitate clinical application. Although NGS has tremendously improved the resolution of SV detection, it also poses a significant challenge to prioritize and pinpoint a small subset of SV that is clinically relevant. Therefore, accurately discerning the pathogenicity of the SVs identified through NGS testing is profound for its clinical adoption [112]. The guideline for the interpretation and reporting of constitutional CNVs have been jointly proposed by the ACMG and the Clinical Genome Resource (ClinGen), suggesting scoring metrics by integrating reported cases, consistency of phenotype, the pattern of inheritance, and the pathogenic mechanisms of variants to prioritize the CNV pathogenicity [113].

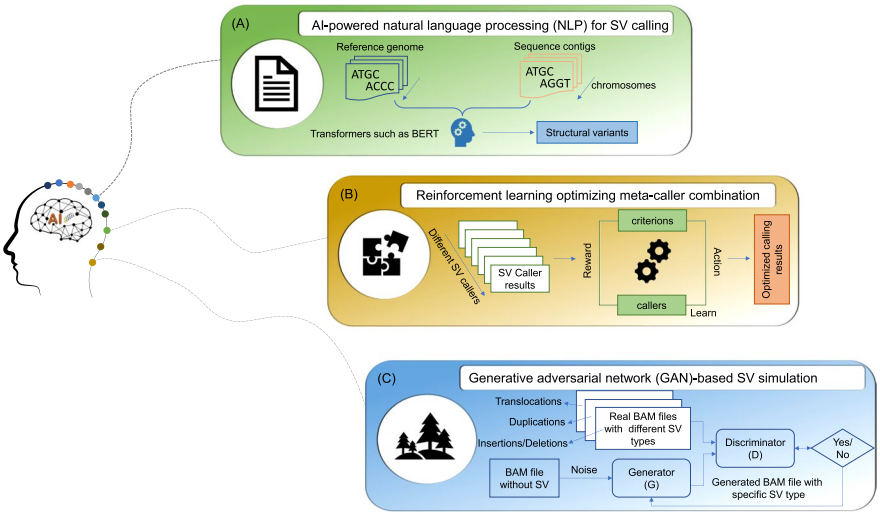

Fig. 3 The role of AI in promoting SV detection. AI-powered natural language processing (NLP) for SV calling. Considering the suboptimal performance of different SV calling algorithms concerning completeness and accuracy, we suggest that deep learning may be an alternative worth further exploration. CNNs are the primary deep learning algorithm investigated for SV detection, which considers the BAM files as image. Rapid development of deep learning algorithms such as AI-powered NLP not only provided unprecedented innovation for information retrieve from free-text documents but repositioned in other type of biological information such as chemical structures and protein sequences [104]. Here, we developed a hypothesis by resembling chromosomes as paragraphs, the sequence reads as sentences, and different A, T, G, C combinations (e.g., tandem repeats and microsatellite) as vocabularies. Subsequently, the AI-powered language models such as different transformers [105–107] could be utilized to digest genome sequence as human beings read a book. The difference between the sample genome and reference genome (i.e., variants) could be extracted by compared transformer-based genome embedding, which is very similar to the rationale behind de novo assembly (**A**). Reinforcement learning optimizing meta-caller combination. There is the potential to integrate multiple callers using more sophisticated approaches than simple heuristic union/intersection rules for improving SV detection. Artificial intelligence (AI) may be a solution. The rapid evolution of emerging genomics technologies suggests that improved SV detection should be taking place. The ideal combination strategy for combining different SV callers is to take advantage of each SV caller and eliminate the false positives, which fits well the concept of reinforcement learning. Reinforcement learning is a branch of deep learning that focuses on how intelligent agents ought to take actions in an environment to maximize cumulative reward [108]. AlphaGo is an excellent example of reinforcement learning applications [109]. Reinforcement learning could be utilized to develop the intelligent ensemble SV callers to maximize SV detection performance (**B**). For each type of SV, the combination for each SV caller could be learned by minimizing the loss function that measures the divergence between called SV and ground truth. Ultimately, reinforcement learning-based ensemble SV callers allow the integration of any individual caller, incorporate different SV types, and incorporate the advantages of newly developed technologies. Generative adversarial network (GAN)-based SV simulation. A true set is the key to investigate the accuracy and reproducibility of SV detection. Unfortunately, the complexity of the SV events and associated genome properties are central to the whole picture of SV events in the sample, hampering objective evaluation. Many reports suggest that simulated ground truth cannot recapitulate genome and SV characteristics and mimic the actual patient situation. Therefore, the simulated SV truth sets with high commutability are urgently needed. The generative adversarial network (GAN) is a deep neural network framework integrating a generative model and discriminative model to generate new data similar to the statistical distribution of the training set [110]. GANs have been widely applied in image generation in fashion, art, and advertising and have attracted much attention in the scientific community. For example, one type of GAN model, named DeepFake, has been utilized to predict cell type-specific transcriptional states induced by drug treatment [111]. Here, we envision a generative adversarial network (GAN) to simulate the binary alignment map (BAM) file spiked in different SV types based on the actual data (**C**). The proposed GAN model collects the high-quality BAM files with varying SV and length types from the real data, such as SEQC-II and other consortium efforts, as a training set to generate the target SV spiked BAM file. The potential benefit of the proposed GAN model is the simulated SV spiked BAM file could maximize the preservation of the original matrix effect of real data such as VAF level and tumor purity

Some efforts have been made for pathogenicity prediction of SV, such as strategies by aggregation of SNP pathogenicity within the SV intervals (e.g., SVscore [114] and AnnotSV [115]) and rule-based approaches for pathogenicity assessment based on the ACMG guideline (e.g., ClassifyCNV [116]). The guideline's execution relies heavily on domain experts and opinion specific, limiting its application to tackle a large number of CNVs detected by NGS testings [117]. ClassifyCNV is the first tool that automates the implementation of the updated ACMG guidelines to classify CNVs, which is suitable for integration into NGS analysis pipelines. However, current SV pathogenicity scoring tools mainly focus on protein-coding regions, and no single approach integrates the distribution of CNVs across ethnic groups to more precisely predict CNV pathogenicity.

Machine learning-based approaches by integrating different genome features and ethnic information across the population could effectively improve the prediction performance of SV pathogenicity in the whole genome scale. As part of the FDA-led SEQC II effort, we developed a novel machine learning-based framework X-CNV (www.unimd.org/XCNV) by integrating over 30 informative features correlating with SNV pathogenicity, to quantitively predict the pathogenicity of CNVs across various ethnic groups across approximately 93% of the human genome [118]. The proposed X-CNV outperformed all the current CNV pathogenicity tools with an AUC of 0.94. Moreover, we developed a meta-voting prediction (MVP) score to quantitively measure the pathogenic effect aligned with the ACMG guideline to enhance clinical application. In the current version of X-CNV, we employed the XGBoost algorithm, and further investigations on advanced deep learning algorithms may promise enhanced performance.

Although machine learning and AI shed light on SV detection and interpretation, more comprehensive evaluation and further investigation on the context of use are highly recommended towards a robust, secured, privacy-preserving, and explainable machine learning and AI solution for clinical application. The U.S. Food and Drug Administration (FDA) recently issued the "Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan" to enable the FDA and manufacturers to evaluate and monitor a software product in a lifecycle-based regulatory framework for machine learning and AI technologies and allow for modifications to be made from real-world learning and adaptation (https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device). Crowdsourcing efforts led by government agencies such as NIST and FDA will be tremendously helpful to prioritize fundamental and translational AI research consistent with regulatory priorities for robust, safe, secure, and privacy-preserving machine learning in different real-world applications. One example is PrecisionFDA, led by the U.S. FDA, which is a secure, collaborative, high-performance computing platform to advance precision medicine, inform regulatory science, and enable improvements in health outcomes. Several PrecisionFDA challenges have been launched, such as calling variants from short and long reads in difficult-to-map regions to standardize NGS testing in the precision medicine practice (https://precision.fda.gov/challenges) [119]. The SEQC-II consortium will release a new challenge on machine learning and AI powering genetic variant identification in the PrecisionFDA to take advantage of crowdsourcing efforts to better understand AI's pros and cons in the context of NGS testing in a clinical setting.

Towards clinical implementation of NGS-based SV detection

Although NGS has tremendously improved the resolution of SV detection and facilitated our understanding of disease etiology and pathogenesis, substantial challenges remain for its full clinical implementation [34]. The influential critical aspects of the clinical implementation of NGS-based SV detection are multifactorial, and the significant factors include pathogenicity assessment, reporting, accreditation, analytical, and clinical validation. The current first-line clinically accredited laboratory genetic tests are still array-based technologies such as array comparative genome hybridization (aCGH) or SNP arrays, which are only capable to detect CNV down to a resolution of ~ 50 kb and unable to detect other types of SV events, such as inversions or balanced translocations. NGS has shown high sensitivity, specificity, and reproducibility for the SV less than 50 bp [36]. Besides, NGS shows its promise in the large SV detection, although substantial space for further improvement for enhanced sensitivity and specificity.

The strategies of promoting NGS towards its clinical implementation and adoption have been intensively discussed elsewhere [120, 121]. One of the outstanding prerequisites of NGS application in a clinical setting is whether the NGS-based SV detection is superior to the current first-line clinically accredited genetic testing. Gross et al. [122] conducted a comparative analysis between NGS-based CNV detection and clinically accredited array-based strategy across a clinical cohort of 79 rare and undiagnosed cases. The study showed CNV calls from NGS are at least as sensitive as those from microarrays while only creating a modest increase in the number of variants interpreted (~10 CNVs per case). Encouragingly, 15% of these incidental or secondary findings (ISFs) from NGS could be confirmed with an orthogonal approach.

The potential high false-positive rate of NGS-based SV detection is one of the major concerns for its clinical applications compared to currently clinically accredited laboratory genetic tests. Since the NGS pipeline for SV detection involves multiple steps, a standard framework for managing and standardizing the NGS-based SV detection is urgently needed for clinical application. To fill the gap, a ClinSV was recently proposed to provide a "one-stop" solution for WGS based SV integration, annotation, prioritization, and visualization [123]. The proposed ClinSV achieved a low false rate (1.5~4.5%) and high reproducibility (95~99%), and high sensitivity (99.8%) for simulated pathogenic ClinVar CNVs > 10 kb and 100% from clinically accredited array-based testing. More importantly, ClinSV identified actionable variants in 22 of 485 patients (4.7%), 35~63% of which were not identified by current clinical microarray designs.

Future directions

The accurate and reproducible detection of SVs is required for use in clinical applications. As a direct benefit of the rapid evolution of genomic technologies coupled with the work of different stakeholders, an increased appreciation of the contribution of SVs in genetic diversity and disease etiology continues to emerge [4, 124–126]. A strategic and thoughtful application of genomics technologies can drive seamless harmonization of diverse strategies to provide reliable and robust SV detection. Here, we have made several proposals to improve the accuracy and reproducibility of SV detection. Some aspects remain to be addressed such as the reliable estimation for SV pathogenicity [115, 116, 118] and the association between SV and complex trait loci

[127]. Furthermore, our perspective is only focused on the human genome; however, SV events are also widely distributed in other species [128, 129]. One notable example is SV events in SARS-CoV-2, which may contribute to COVID-19 transmission and severity [130, 131]. Additionally, we did not discuss RNA-seq based gene fusion detection. Gene capture strategies, genomics technologies selection and calling algorithms divergence have been extensively studied under SEQC II. Specifically, the SEQC Oncopanel working group generated multi-platform and multi-lab long/short-read RNA sequencing data along with conventional genomics technologies such as stranded/polyadenylated/ribosome-depleted sequencing [28, 29]. This enables a comprehensive assessment of gene fusion detection capability using the common reference material UHRR.

The field of SV detection continues to expand with advances in genome technologies, the establishment of and more reference standards and the release of high-quality calling sets. These resources allow us to further investigate some of the critical questions that remain and facilitate novel genomics technology development towards reliable and comprehensive SV identification (see Outstanding Questions). Notably, regulatory attention is on standardizing structural variant detection and on classification for convenient clinical adoption [132]. We suggest establishing a bridge among different stakeholders to establish best-practice recommendations and quality control to encourage the uptake of SV diagnosis into clinical practice. Additionally, we argue that innovations in AI may help in tackling key challenges in SV detection and provide alternative options to develop more accurate and robust approaches. However, these proposed deep learning frameworks still require close examination and feasibility analysis. For example, there are still multiple challenges to fully implement deep learning framework in SV calling such as lack of training data, divergency of performance on complexity of SV and regional differences.

Conclusions

The ongoing development of many genomics technologies sees a period of excitement, huge investment and perhaps disappointment that, in turn, may trigger a new wave of innovation. We foresee that more accurate and reproducible SV detection approaches will emerge soon, generating a more complete picture of the landscape of the human genome. Next steps will require different stakeholders to work towards a common goal of further resolving the difficulty in SV detection and accelerating clinical application.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-022-02636-8.

Additional file 1. Review history.

Acknowledgements
FJS was in part funded over the US National Institutes of Health (UM1 HG008898).

Peer review information
Anahita Bishop and Barbara Cheifet were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history
The review history is available as Additional file 1.

Disclaimer
The views presented in this article do not necessarily reflect current or future opinion or policy of the US Food and Drug Administration. Any mention of commercial products is for clarification and not intended as endorsement.

Authors' contributions
Z.L. and W.T. conceived and designed the study. Z.L., F.S., and W.T. wrote the article. Z.L., F.S., R.R., T.M., J.S., and W.T. revised the article. The authors read and approved the final article.

Funding
Fritz J. Sedlazeck was supported by the NIH (UM1 HG008898 & 1U01HG011758-01).

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
RR is co-founder and co-director of ApconiX, an integrated toxicology and ion channel company that provides expert advice on non-clinical aspects of drug discovery and drug development to academia, industry, and not-for-profit organizations. FJS received travel reimbursements from Pacbio and Oxford Nanopore.

Author details
[1]National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR 72079, USA. [2]ApconiX, BioHub at Alderley Park, Alderley Edge SK10 4TG, UK. [3]University of Birmingham, Edgbaston, Birmingham B15 2TT, UK. [4]Australian Institute for Bioengineering and Nanotechnology, University of Queensland, Brisbane QLD, Australia. [5]Garvan Institute of Medical Research, Sydney, NSW, Australia. [6]St Vincent's Clinical School, University of New South Wales, Sydney, NSW, Australia. [7]Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA.

Received: 27 August 2021 Accepted: 15 February 2022
Published online: 03 March 2022

References

1. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. Nat Rev Genet. 2020;21:171–89.
2. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et al. DNA sequencing at 40: past, present and future. Nature. 2017;550:345–53.
3. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. 2019;20:246.
4. Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, et al. Patterns of somatic structural variation in human cancer genomes. Nature. 2020;578:112–21.
5. Macintyre G, Ylstra B, Brenton JD. Sequencing structural variants in cancer for precision therapeutics. Trends Genet. 2016;32:530–42.
6. Hadi K, Yao X, Behr JM, Deshpande A, Xanthopoulakis C, Tian H, et al. Distinct classes of complex structural variation uncovered across thousands of cancer genome graphs. Cell. 2020;183:197–210.e132.
7. Gardner EJ, Prigmore E, Gallone G, Danecek P, Samocha KE, Handsaker J, et al. Contribution of retrotransposition to developmental disorders. Nat Commun. 2019;10:4630.
8. Sherman MA, Rodin RE, Genovese G, Dias C, Barton AR, Mukamel RE, et al. Large mosaic copy number variations confer autism risk. Nat Neurosci. 2021;24:197–203.
9. D'Abate L, Walker S, Yuen RKC, Tammimies K, Buchanan JA, Davies RW, et al. Predictive impact of rare genomic copy number variations in siblings of individuals with autism spectrum disorders. Nat Commun. 2019;10:5519.
10. Halvorsen M, Huh R, Oskolkov N, Wen J, Netotea S, Giusti-Rodriguez P, et al. Increased burden of ultra-rare structural variants localizing to boundaries of topologically associated domains in schizophrenia. Nat Commun. 1842;2020:11.
11. Levy-Sakin M, Pastor S, Mostovoy Y, Li L, Leung AKY, McCaffrey J, et al. Genome maps across 26 human populations reveal population-specific patterns of structural variation. Nat Commun. 2019;10:1025.
12. Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, et al. Characterizing the major structural variant alleles of the human genome. Cell. 2019;176:663–675.e619.
13. Abel HJ, Larson DE, Regier AA, Chiang C, Das I, Kanchi KL, et al. Mapping and characterization of structural variation in 17,795 human genomes. Nature. 2020;583:83–9.
14. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019;10:1784.
15. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526:75–81.
16. Santos M, Niemi M, Hiratsuka M, Kumondai M, Ingelman-Sundberg M, Lauschke VM, et al. Novel copy-number variations in pharmacogenes contribute to interindividual differences in drug pharmacokinetics. Genet Med. 2018;20:622–9.

17. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat Rev Genet. 2018;19:329–46.
18. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006;7:85–97.
19. De Coster W, Van Broeckhoven C. Newest methods for detecting structural variations. Trends Biotechnol. 2019;37:973–82.
20. Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, et al. Best practices for benchmarking germline small-variant calls in human genomes. Nat Biotechnol. 2019;37:555–60.
21. Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, Heaton H, et al. An open resource for accurately benchmarking small variant and reference calls. Nat Biotechnol. 2019;37:561–6.
22. Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, Xiao C, et al. A robust benchmark for detection of germline large deletions and insertions. Nat Biotechnol. 2020;38:1347–55.
23. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, et al. A structural variation reference for medical and population genetics. Nature. 2020;581:444–51.
24. Gong T, Hayes VM, Chan EKF. Detection of somatic structural variants from short-read next-generation sequencing data. Brief Bioinform. 2020;22:bbaa056.
25. Cameron DL, Di Stefano L, Papenfuss AT. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. Nat Commun. 2019;10:3240.
26. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol. 2019;20:117.
27. Shi L, Kusko R, Wolfinger RD, Haibe-Kains B, Fischer M, Sansone S-A, et al. The international MAQC Society launches to enhance reproducibility of high-throughput technologies. Nat Biotechnol. 2017;35:1127–8.
28. Jones W, Gong B, Novoradovskaya N, Li D, Kusko R, Richmond TA, et al. A verified genomic reference sample for assessing performance of cancer panels detecting small variants of low allele frequency. Genome Biol. 2021;22:111.
29. Gong B, Li D, Kusko R, Novoradovskaya N, Zhang Y, Wang S, et al. Cross-oncopanel study reveals high sensitivity and accuracy with overall analytical performance depending on genomic regions. Genome Biol. 2021;22:109.
30. Chen W, Zhao Y, Chen X, Yang Z, Xu X, Bi Y, et al. A multicenter study benchmarking single-cell RNA sequencing technologies using reference samples. Nat Biotechnol. 2020:39:1–12.
31. Xiao W, Ren L, Chen Z, Fang LT, Zhao Y, Lack J, Guan M, Zhu B, Jaeger E, Kerrigan L, et al: Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. Nat Biotechnol. 2021;39:1141–50.
32. Fang LT, Zhu B, Zhao Y, Chen W, Yang Z, Kerrigan L, et al. Establishing reference samples for detection of somatic mutations and germline variants with NGS technologies. bioRxiv. 2019.
33. Hardwick SA, Deveson IW, Mercer TR. Reference standards for next-generation sequencing. Nat Rev Genet. 2017;18:473–84.
34. Liu Z, Zhu L, Roberts R, Tong W. Toward clinical implementation of next-generation sequencing-based genetic testing in rare diseases: where are we? Trends Genet. 2019;35:852–67.
35. Fang LT, Zhu B, Zhao Y, Chen W, Yang Z, Kerrigan L, et al. Establishing community reference samples, data and call sets for benchmarking cancer mutation detection using whole-genome sequencing. Nat Biotechnol. 2021;39:1151–60.
36. Khayat MM, Sahraeian SME, Zarate S, Carroll A, Hong H, Pan B, et al. Hidden biases in germline structural variant detection. Genome Biol. 2021;22:347.
37. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.
38. López S, Tarekegn A, Band G, van Dorp L, Bird N, Morris S, et al. Evidence of the interplay of genetics and culture in Ethiopia. Nat Commun. 2021;12:3581.
39. Pan B, Ren L, Onuchic V, Guan M, Kusko R, Bruinsma S, et al. Assessing reproducibility of inherited variants detected with short-read whole genome sequencing. Genome Biol. 2022;23:2.
40. Ball MP, Thakuria JV, Zaranek AW, Clegg T, Rosenbaum AM, Wu X, et al. A public resource facilitating clinical use of genomes. Proc Natl Acad Sci. 2012;109:11920–7.
41. Foox J, Tighe SW, Nicolet CM, Zook JM, Byrska-Bishop M, Clarke WE, et al. Performance assessment of DNA sequencing platforms in the ABRF next-generation sequencing study. Nat Biotechnol. 2021;39:1129–40.
42. Huo Z, Tu J, Lee D-F, Zhao R. Engineering mutation clones in mammalian cells with CRISPR/Cas9. In: Immune mediators in cancer. Methods Mol Biol. 2020;2108:355–69.
43. Suzuki T, Tsukumo Y, Furihata C, Naito M, Kohara A. Preparation of the standard cell lines for reference mutations in cancer gene-panels by genome editing in HEK 293T/17 cells. Genes Environ. 2020;42:8.
44. Lee AY, Ewing AD, Ellrott K, Hu Y, Houlahan KE, Bare JC, et al. Combining accurate tumor genome simulation with crowdsourcing to benchmark somatic structural variant detection. Genome Biol. 2018;19:1–15.
45. Raymond VM, Gray SW, Roychowdhury S, Joffe S, Chinnaiyan AM, Parsons DW, et al. Germline findings in tumor-only sequencing: points to consider for clinicians and laboratories. J Natl Cancer Inst. 2016;108:djv351.
46. Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med. 2013;15:565–74.
47. Green RC, Goddard KAB, Jarvik GP, Amendola LM, Appelbaum PS, Berg JS, et al. Clinical sequencing exploratory research consortium: accelerating evidence-based practice of genomic medicine. Am J Hum Genet. 2016;98:1051–66.
48. Alioto TS, Buchhalter I, Derdak S, Hutter B, Eldridge MD, Hovig E, et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat Commun. 2015;6:10001.
49. Gazdar AF, Kurvari V, Virmani A, Gollahon L, Sakaguchi M, Westerfield M, et al. Characterization of paired tumor and non-tumor cell lines established from patients with breast cancer. Int J Cancer. 1998;78:766–74.
50. Staaf J, Lindgren D, Vallon-Christersson J, Isaksson A, Göransson H, Juliusson G, et al. Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. Genome Biol. 2008;9:1–18.

51. Stephens PJ, McBride DJ, Lin M-L, Varela I, Pleasance ED, Simpson JT, et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature. 2009;462:1005–10.
52. Popova T, Manié E, Rieunier G, Caux-Moncoutier V, Tirapo C, Dubois T, et al. Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. Cancer Res. 2012;72:5454–62.
53. Hayeems RZ, Dimmock D, Bick D, Belmont JW, Green RC, Lanpher B, et al. Clinical utility of genomic sequencing: a measurement toolkit. NPJ Genom Med. 2020;5:56.
54. Deveson IW, Gong B, Lai K, LoCoco JS, Richmond TA, Schageman J, et al. Evaluating the analytical validity of circulating tumor DNA sequencing assays for precision oncology. Nat Biotechnol. 2021;39:1115–28.
55. Almarri MA, Bergström A, Prado-Martinez J, Yang F, Fu B, Dunham AS, et al. Population structure, stratification, and introgression of human structural variation. Cell. 2020;182:189–199.e115.
56. Willey JC, Morrison TB, Austermiller B, Crawford EL, Craig DJ, Blomquist TM, et al. Advancing NGS quality control to enable measurement of actionable mutations in circulating tumor DNA. Cell Rep Methods. 2021;1:100106.
57. Park C-Y, Sung JJ, Kim D-W. Genome editing of structural variations: modeling and gene correction. Trends Biotechnol. 2016;34:548–61.
58. Boroviak K, Doe B, Banerjee R, Yang F, Bradley A. Chromosome engineering in zygotes with CRISPR/Cas9. Genesis (New York, NY: 2000). 2016;54:78–85.
59. Zhang X-H, Tee LY, Wang X-G, Huang Q-S, Yang S-H. Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. Mol Ther Nucleic Acids. 2015;4:e264.
60. Watson CT, Marques-Bonet T, Sharp AJ, Mefford HC. The genetics of microdeletion and microduplication syndromes: an update. Annu Rev Genomics Hum Genet. 2014;15:215–44.
61. Chen W, Zhao Y, Chen X, Yang Z, Xu X, Bi Y, et al. A multicenter study benchmarking single-cell RNA sequencing technologies using reference samples. Nat Biotechnol. 2021;39:1103–14.
62. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17:333.
63. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15:461–8.
64. Sanders AD, Falconer E, Hills M, Spierings DCJ, Lansdorp PM. Single-cell template strand sequencing by strand-seq enables the characterization of individual homologs. Nat Protoc. 2017;12:1151–76.
65. Hård J, Mold JE, Eisfeldt J, Tellgren-Roth C, Häggqvist S, Bunikis I, et al. Long-read whole genome analysis of human single cells. bioRxiv. 2021. https://doi.org/10.1101/2021.04.13.439527.
66. Chen X, Yang Z, Chen W, Zhao Y, Farmer A, Tran B, et al. A multi-center cross-platform single-cell RNA sequencing reference dataset. Sci Data. 2021;8:1–11.
67. Mu JC, Mohiyuddin M, Li J, Bani Asadi N, Gerstein MB, Abyzov A, et al. VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications. Bioinformatics (Oxford, England). 2015;31:1469–71.
68. Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bähler J, Sedlazeck FJ. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat Commun. 2017;8:14061.
69. Yang C, Chu J, Warren RL, Birol I. NanoSim: nanopore sequence read simulator based on statistical characterization. GigaScience. 2017;6(4):1–6.
70. Ono Y, Asai K, Hamada M. PBSIM: PacBio reads simulator—toward accurate genome assembly. Bioinformatics. 2013;29:119–21.
71. Escalona M, Rocha S, Posada D. A comparison of tools for the simulation of genomic next-generation sequencing data. Nat Rev Genet. 2016;17:459–69.
72. Carvalho CMB, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet. 2016;17:224–38.
73. Zhao Y, Fang LT, Shen TW, Choudhari S, Talsania K, Chen X, Shetty J, Kriga Y, Tran B, Zhu B, et al: Whole genome and exome sequencing reference datasets from a multi-center and cross-platform benchmark study. Sci Data. 2021;8:296.
74. Xiao W, Ren L, Chen Z, Fang LT, Zhao Y, Lack J, et al. Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. Nat Biotechnol. 2021;39:1141–50.
75. Chen Y-C, Seifuddin F, Nguyen C, Yang Z, Chen W, Yan C, et al. Comprehensive assessment of somatic copy number variation calling using next-generation sequencing data. bioRxiv. 2021. https://doi.org/10.1101/2021.02.18.431906.
76. Xiao C, Chen Z, Chen W, Padilla C, Fang L-T, Liu T, et al. Personalized genome assembly for accurate cancer somatic mutation discovery using cancer-normal paired reference samples. bioRxiv. 2021. https://doi.org/10.1101/2021.04.09.438252.
77. Group SOSW, Zhang Y, Blomquist TM, Kusko R, Stetson D, Zhang Z, et al. Deep oncopanel sequencing reveals fixation time- and within block position-dependent quality degradation in FFPE processed samples. bioRxiv. 2021. https://doi.org/10.1101/2021.04.06.438687.
78. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455–77.
79. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27:722–36.
80. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18:170–5.
81. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Nat Biotechnol. 2020;38:1044–53.
82. De Coster W, Weissensteiner MH, Sedlazeck FJ. Towards population-scale long-read sequencing. Nat Rev Genet. 2021;22:572–87.

83. Wagner J, Olson ND, Harris L, McDaniel J, Cheng H, Fungtammasan A, Hwang Y-C, Gupta R, Wenger AM, Rowell WJ, et al: Curated variation benchmarks for challenging medically relevant autosomal genes. Nature Biotechnology. 2022. https://doi.org/10.1038/s41587-021-01158-1.
84. Chen S, Krusche P, Dolzhenko E, Sherman RM, Petrovski R, Schlesinger F, et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. Genome Biol. 2019;20:291.
85. Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga J, et al. Genotyping structural variants in pangenome graphs using the vg toolkit. Genome Biol. 2020;21:35.
86. Cameron DL, Schröder J, Penington JS, Do H, Molania R, Dobrovic A, et al. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res. 2017;27:2050–60.
87. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016;32:1220–2.
88. Becker T, Lee WP, Leone J, Zhu Q, Zhang C, Liu S, et al. FusorSV: an algorithm for optimally combining data from multiple structural variation detection methods. Genome Biol. 2018;19:38.
89. Chander V, Gibbs RA, Sedlazeck FJ. Evaluation of computational genotyping of structural variation for clinical diagnoses. Gigascience. 2019;8(9):giz110.
90. Mohiyuddin M, Mu JC, Li J, Bani Asadi N, Gerstein MB, Abyzov A, et al. MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. Bioinformatics. 2015;31:2741–4.
91. Zarate S, Carroll A, Mahmoud M, Krasheninina O, Jun G, Salerno WJ, et al. Parliament2: accurate structural variant calling at scale. GigaScience. 2020;9(12):giaa145.
92. Wong K, Keane TM, Stalker J, Adams DJ. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. Genome Biol. 2010;11:R128.
93. Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat Methods. 2015;12:966–8.
94. Pounraja VK, Jayakar G, Jensen M, Kelkar N, Girirajan S. A machine-learning approach for accurate detection of copy number variants from exome sequencing. Genome Res. 2019;29:1134–43.
95. Aganezov S, Goodwin S, Sherman RM, Sedlazeck FJ, Arun G, Bhatia S, et al. Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. Genome Res. 2020;30:1258–73.
96. Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. Genome Med. 2019;11:70.
97. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16:321–32.
98. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019;51:12–8.
99. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet. 2019;20:389–403.
100. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36:983–7.
101. Hill T, Unckless RL. A deep learning approach for detecting copy number variation in next-generation sequencing data. G3 (Bethesda, Md). 2019;9:3575–82.
102. Cai L, Wu Y, Gao J. DeepSV: accurate calling of genomic deletions from high-throughput sequencing data using deep convolutional neural network. BMC Bioinformatics. 2019;20:665.
103. Park H, Chun S-M, Shim J, Oh J-H, Cho EJ, Hwang HS, et al. Detection of chromosome structural variation by targeted next-generation sequencing and a deep learning application. Sci Rep. 2019;9:3644.
104. Liu Z, Roberts RA, Lal-Nag M, Chen X, Huang R, Tong W. AI-based language models powering drug discovery and development. Drug Discov Today. 2021;26:2593–607.
105. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv preprint arXiv:170603762; 2017.
106. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805; 2018.
107. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv preprint arXiv:200514165; 2020.
108. Hu J, Niu H, Carrasco J, Lennox B, Arvin F. Voronoi-based multi-robot autonomous exploration in unknown environments via deep reinforcement learning. IEEE Trans Veh Technol. 2020;69:14413–23.
109. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. Nature. 2017;550:354–9.
110. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. arXiv preprint arXiv:14062661; 2014.
111. Umarov R, Li Y, Arner E: DeepCellState: An autoencoder-based framework for predicting cell type specific transcriptional states induced by drug treatment. PLoS Comput Biol. 2021;17:e1009465–e1009465.
112. Marian AJ. Clinical interpretation and management of genetic variants. JACC Basic Transl Sci. 2020;5:1029–42.
113. Riggs ER, Andersen EF, Cherry AM, Kantarci S, Kearney H, Patel A, et al. Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). Genet Med. 2020;22:245–57.
114. Ganel L, Abel HJ, FinMetSeq C, Hall IM. SVScore: an impact prediction tool for structural variation. Bioinformatics. 2017;33:1083–5.
115. Geoffroy V, Herenger Y, Kress A, Stoetzel C, Piton A, Dollfus H, et al. AnnotSV: an integrated tool for structural variations annotation. Bioinformatics. 2018;34:3572–4.
116. Gurbich TA, Ilinsky VV. ClassifyCNV: a tool for clinical annotation of copy-number variants. Sci Rep. 2020;10:20375.
117. Rivera-Muñoz EA, Milko LV, Harrison SM, Azzariti DR, Kurtz CL, Lee K, et al. ClinGen variant curation expert panel experiences and standardized processes for disease and gene-level specification of the ACMG/AMP guidelines for sequence variant interpretation. Hum Mutat. 2018;39:1614–22.

118. Zhang L, Shi J, Ouyang J, Zhang R, Tao Y, Yuan D, et al. X-CNV: genome-wide prediction of the pathogenicity of copy number variations. Genome Med. 2021;13:132.
119. Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, et al. precisionFDA truth challenge V2: calling variants from short- and long-reads in difficult-to-map regions. bioRxiv. 2021. https://doi.org/10.1101/2020.11.13.380741.
120. Colomer R, Mondejar R, Romero-Laorden N, Alfranca A, Sanchez-Madrid F, Quintela-Fandino M. When should we order a next generation sequencing test in a patient with cancer? EClinicalMedicine. 2020;25:100487.
121. Singh RR, Luthra R, Routbort MJ, Patel KP, Medeiros LJ. Implementation of next generation sequencing in clinical molecular diagnostic laboratories: advantages, challenges and potential. Expert Rev Precis Med Drug Dev. 2016;1:109–20.
122. Gross AM, Ajay SS, Rajan V, Brown C, Bluske K, Burns NJ, et al. Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease. Genet Med. 2019;21:1121–30.
123. Minoche AE, Lundie B, Peters GB, Ohnesorg T, Pinese M, Thomas DM, et al. ClinSV: clinical grade structural and copy number variant detection from whole genome sequencing data. Genome Med. 2021;13:32.
124. Sanchis-Juan A, Stephens J, French CE, Gleadall N, Mégy K, Penkett C, et al. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. Genome Med. 2018;10:95.
125. Zhang Y, Yang L, Kucherlapati M, Chen F, Hadjipanayis A, Pantazi A, et al. A pan-cancer compendium of genes deregulated by somatic genomic rearrangement across more than 1,400 cases. Cell Rep. 2018;24:515–27.
126. Eichler EE. Genetic variation, comparative genomics, and the diagnosis of disease. N Engl J Med. 2019;381:64–74.
127. Chakraborty M, Emerson JJ, Macdonald SJ, Long AD. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. Nat Commun. 2019;10:4872.
128. Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell. 2020;182:145–161.e123.
129. Chen L, Chamberlain AJ, Reich CM, Daetwyler HD, Hayes BJ. Detection and validation of structural variations in bovine whole-genome sequence data. Genet Sel Evol. 2017;49:13.
130. Portelli S, Olshansky M, Rodrigues CHM, D'Souza EN, Myung Y, Silk M, et al. Exploring the structural distribution of genetic variation in SARS-CoV-2 with the COVID-3D online resource. Nat Genet. 2020;52:999–1001.
131. Lauring AS, Hodcroft EB. Genetic variants of SARS-CoV-2—what do they mean? JAMA. 2021;325:529–31.
132. Brandt T, Sack LM, Arjona D, Tan D, Mei H, Cui H, et al. Adapting ACMG/AMP sequence variant classification guidelines forsingle-gene copy number variants. Genet Med. 2020;22:336–44.

Publisher's Note