

Review article



Variant calling and benchmarking in an era of complete human genome sequences

Nathan D. Olson¹, Justin Wagner¹, Nathan Dwarshuis¹, Karen H. Miga², Fritz J. Sedlazeck¹³, Marc Salit⁴
& Justin M. Zook¹

Abstract

Genetic variant calling from DNA sequencing has enabled understanding of germline variation in hundreds of thousands of humans. Sequencing technologies and variant-calling methods have advanced rapidly, routinely providing reliable variant calls in most of the human genome. We describe how advances in long reads, deep learning, de novo assembly and pangenomes have expanded access to variant calls in increasingly challenging, repetitive genomic regions, including medically relevant regions, and how new benchmark sets and benchmarking methods illuminate their strengths and limitations. Finally, we explore the possible future of more complete characterization of human genome variation in light of the recent completion of a telomere-to-telomere human genome reference assembly and human pangenomes, and we consider the innovations needed to benchmark their newly accessible repetitive regions and complex variants.

Sections

Introduction

Remaining challenges across the genome

Advances in sequencing methods

Advances in variant-calling methods

Benchmarking variant accuracy

Conclusions and perspectives

¹Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA.

²UC Santa Cruz Genomics Institute, University of California Santa Cruz, Santa Cruz, CA, USA. ³Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, USA. ⁴The MITRE Corporation, McLean, VA, USA.

e-mail: jzook@nist.gov

Introduction

Calling variants (sequence differences) in the genome of an individual with respect to a reference genome assembly has been the standard practice for characterizing hundreds of thousands of human genomes since the completion of the Human Genome Project. The human reference genome assembly, first published in 2003, was the basis for a wide variety of methods to ‘map’ (align) reads to the reference and identify differences between those reads and the reference, commonly termed ‘variant calling’. The typical variant-calling process includes sequencing, read mapping or de novo assembly, variant calling, filtering of false positives and, increasingly, phasing. Calling variants across many individuals has enabled understanding of variants associated with disease and clinical diagnostics, as well as evolutionary mechanisms.

Calling small variants from short-read sequencing has become highly accurate in regions of the genome to which short reads (typically 150–250 bp long) can be accurately mapped^{1–3}. This high accuracy has resulted from improved bioinformatics tools, decreases in sequencing error rates (for example, due to PCR-free sequencing) and modest increases in short-read length. However, short reads have limitations in repetitive regions of the genome (Fig. 1). Generally, if a read is shorter than a repeated or duplicated region, it may be difficult to determine the ‘true’ location of the read in the genome, and thus calling variants from it will be ambiguous. Such repetitive regions include segmental duplications, long interspersed nuclear elements (LINEs), short tandem repeats (STRs), variable number tandem repeats (VNTRs), telomeres and satellite DNA repeats (up to 30 Mb long). To deal with these limitations, novel sequencing methods and bioinformatics tools have been invented to enable longer read lengths, specifically ultralong (100 kb to several megabase) reads⁴ and highly accurate long (15–20 kb) reads⁵. These longer reads enabled the complete de novo assembly of an effectively haploid human genome by the Telomere-to-Telomere (T2T) Consortium, correcting some errors in the GRCh38 human reference genome and adding ~7% of highly repetitive sequence that was previously unassembled⁶.

Genomic benchmarks have played an important role in optimizing sequencing methods and analysis pipelines. These benchmarks include well-characterized samples and variant callsets from community efforts such as the Genome in a Bottle Consortium (GIAB; from the US National Institute of Standards and Technology (NIST)) and Sequencing Quality Control Consortium (SEQC), along with benchmarking tools from the Global Alliance for Genomics and Health (GA4GH). Scientists can use these benchmarks by acquiring a well-characterized sample (for example, HG002 from GIAB), run it through their sequencing and analysis pipeline, and compare their results with the accompanying benchmark variant callsets. Thus, performance can be measured across analysis pipelines in a standardized way, enabling selection of optimal methods for a particular application. Furthermore, as benchmarks include increasingly challenging genomic regions or variants, they help technology and method development, providing key insights into what laboratories may miss.

In this Review, we discuss advances in DNA sequencing technologies and bioinformatics and the accompanying benchmarks, which together have made variant calling routine in much of the human genome, as well as common sources of bias and error in variant calls, particularly in repetitive regions. We review the role of benchmarks and benchmarking tools in understanding and improving variant accuracy, and we conclude with our perspective on the future of variant calling and benchmarking in complex, repetitive and highly variable regions.

Remaining challenges across the genome

Over the years, many advancements have been made in identifying variants and creating benchmarks. Nevertheless, it is important to keep in mind that there are multiple challenging regions left across the genome. Here, we introduce the types of challenging regions and provide insights on why they are challenging and why it still matters to resolve these. The challenges are generally related to different types of repetitive regions in the genome, which are often analogized to regions of a puzzle in which the pieces are very similar. Different types of repeats present different challenges in sequencing errors, mapping and variant calling, which we describe below.

Homopolymers and tandem repeats

Homopolymers and tandem repeats are sequences repeated many times next to each other (Fig. 1a). They cause systematic sequencing errors, mapping errors and challenges in variant representation. Homopolymers are a single base repeated many times, which causes errors during PCR amplification and during sequencing, with most errors in A and T homopolymers due to their high prevalence in the human genome (Fig. 1b). Although less common, G and C homopolymers generally have a higher error rate⁷ (Fig. 1c), partly because certain sequence motifs may cause errors in one direction and not the other, called ‘strand bias’ (for example, GGT is sometimes read as GGG by Illumina sequencing platforms)⁸. Variants in tandem repeats (regions with a sequence of two or more bases repeated consecutively) are often filtered (discarded from analyses) with standard short-read variant callers if short reads do not span the repeat, and are sometimes noisy in long reads (Fig. 1d,e). Long-read platforms, as further described below, include high-fidelity (HiFi) reads from Pacific Biosciences (PacBio) and nanopore-based sequencing from Oxford Nanopore Technologies (ONT). GA-rich simple repeats are poorly covered in current PacBio HiFi data⁹. Although reads can generally be mapped or assembled accurately with sufficient reads longer than the full repeat length, large or complex variants may confuse alignment to the reference. Furthermore, even with accurate long reads that span the repeat, the specific representation of a variant may be important to understanding its role in a phenotype or disease, particularly for the horizontally complex variants (multiple variants on the same haplotype) and vertically complex variants (different variants occurring on both haplotypes) (Fig. 2a,b). These challenges can be different for the two general historical categories of tandem repeats: STRs are defined as repeats whose repeating unit is 6 bp or less, and VNTRs have longer units¹⁰.

As an example for STRs, Huntington disease is caused by having 36 or more CAG repeats in the huntingtin (*HTT*) gene, and the Genome Aggregation Database (gnomAD) recently genotyped 59 disease-associated STR loci in >19,000 samples^{11,12}. Without understanding the repeat structure, variant callers may represent these repeats in a naive, repeat-agnostic way, often with multiple variants in the same repeat. This makes the variant difficult to interpret: ideally, one would want a repeat count and sequence from each haplotype to determine phenotype or disease progression. For VNTRs, correctly identifying and genotyping repeat count is even more difficult for two main reasons (Fig. 1d): first, VNTRs can often be much longer than the sequencing read length, sometimes even for long reads (owing to their longer repeat size and total length), and second, VNTRs often have multiple point mutations in their repeats and multiple repeat motifs, making alignment more difficult¹³. Even with VNTRs longer than the read length, these can often be assembled with accurate long reads, as in the lipoprotein A (*LPA*) gene. adVNTR is a short and long read variant

Review article

a

Unit size	Name	Structure	Example
1	Homopolymer	() _n	
2	Dinucleotide	() _n	
3	Trinucleotide	() _n	
...
N	Tandem repeat	(...) _n	

Sequencing errors in homopolymers

b

Correct sequence	
Single-base deletion	
Single-base insertion	

c

Correct sequence	
Two SNVs (forward)	
Two SNVs (reverse)	

Mapping and sequencing errors in tandem repeats

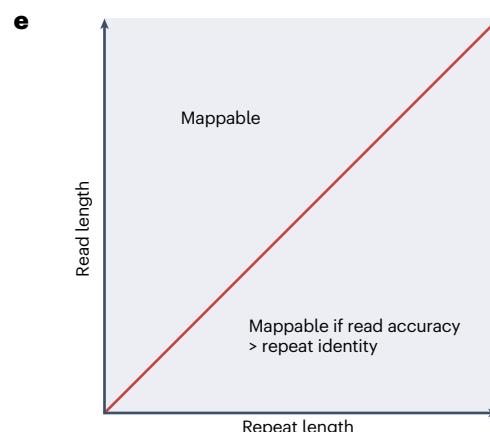
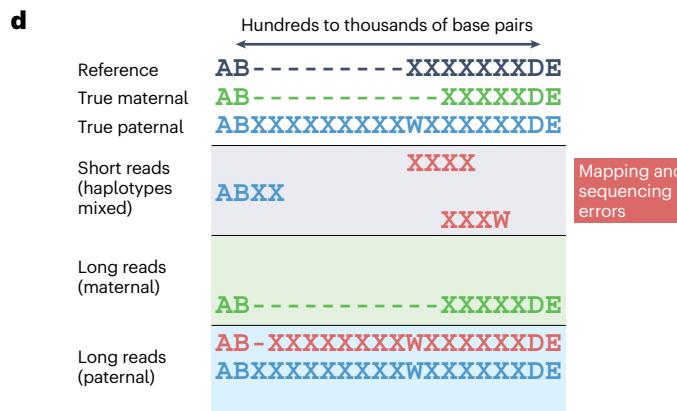


Fig. 1 | Challenges of mapping and variant calling in simple repetitive regions. **a**, Examples of homopolymers and tandem repeats, which are a common source of systematic sequencing and mapping errors. *N* indicates the number of bases within each unit; *n* indicates the number of units in the homopolymer or tandem repeat. **b**, Systematic insertion and deletion (indel) sequencing errors occur in homopolymers and short tandem repeats (STRs), usually with one extra or one missing copy of the repeat unit. In this example, the first read (middle line) has one missing A and the second read (bottom line) has one extra A. This happens most often in sequencing platforms such as 454/Ion Torrent/Ultima and Oxford Nanopore Technologies (ONT), often in raw Pacific Biosciences (PacBio) reads, somewhat often in PacBio high-fidelity (HiFi) and PCR-based short reads, and rarely (except in very long homopolymers) for PCR-free sequencing by synthesis on Illumina and Element platforms. **c**, Systematic errors occur in C and G homopolymers for Illumina platforms owing to sequencing chemistry biases. The single-nucleotide variant (SNV) and indel errors tend to happen after the homopolymer ends, as shown by the errors (black background) in the forward sequencing read (middle line) and reverse sequencing read (bottom line). **d**, Tandem repeats can result in sequencing and mapping errors. X represents the tandem repeat unit, which may be two to hundreds of base pairs in size and is repeated many times, although there frequently are some differences between

units, denoted as W. When tandem repeats are longer than the read length, reads map ambiguously to the repeat sequence, so traditional variant callers miss true variants and sometimes call false positives. For example, because the W sequence is only in the reads and not in the reference, short-read mappers can ambiguously map sequencing reads containing W multiple places within the tandem repeat, resulting in false-positive and/or false-negative variant calls. By contrast, if long reads traverse the entire repeat and flanking sequence, the variants can be accurately called. Although long reads sometimes have systematic errors at tandem repeats (shown in the red long read, which is missing one copy of the X repeat), these can often be averaged to the true variant call because reads can be partitioned by haplotype (except in highly homozygous regions or where the long reads are very noisy). **e**, Mappability of sequencing reads based on the relationship between read length and repeat length along with read accuracy and repeat identity. Reads are generally mappable if they are longer than the repeat length, so that some reads have unique flanking sequence on both ends, or if the read accuracy is higher than the repeat identity and the repeats contain at least one location that differentiates repeat copies. This general rule assumes that variants have negligible impact on mapping. Variants, particularly large or complex variants, can further hinder accurate mapping, but pangenome references can improve mapping for these variants.

caller that specifically focuses on calling VNTRs¹⁴. It uses a hidden Markov model (HMM), which encodes the likelihoods that a base letter will follow a specific position in a repeat or flanking region. Even when accurately assembling or calling variants in VNTRs, representation of variants in these regions is challenging, and new tools need to be developed to compare variants between methods or across individuals.

Segmental duplications

Segmental duplications are nearly identical sequence fragments that are typically defined analytically as being at least 1,000 bp long and occur at least twice throughout the genome, and may be either tandem (adjacent) or interspersed (distant)¹⁵. Because variant calling is inherently dependent on correctly mapped reads, calling small variants in these regions has historically been elusive: variants were frequently filtered or missed even if they were true positives because it was difficult to distinguish them from false positives. Recent advances in both sequencing technology and computation algorithms have begun to unlock these regions. One recent study found that segmental duplications and tandem repeats accounted for >90% of large deletions identified by long reads but missed by short reads¹⁶.

Segmental duplications are challenging for variant calling for several reasons. First, reads may not be long enough to confidently map to the correct copy of the segmental duplication (Fig. 3a), even when using long reads for large, highly identical segmental duplications such as *SMN1* and *SMN2*, associated with spinal muscular atrophy¹⁷. This can result in false-positive and false-negative variant calls, although paralogous sequence variants (PSVs) can be used to distinguish copies in some cases, as described below. Second, large, complex structural variants (SVs) such as inversions and duplications are often mediated by segmental duplications, which presents unmet challenges for variant call representation and benchmarking¹⁸. Third, many segmental duplications differ in copy number between individuals. When an individual has an extra copy of a segmental duplication relative to the reference, reads from the extra copy often map to existing copies in the reference, typically resulting in higher-than-normal coverage and dense false-positive heterozygous variant calls from PSVs (Figs. 2c,3b). In some cases, the GRCh37 and GRCh38 references are missing copies of a segmental duplication, resulting in false positives in all individuals, including medically relevant genes such as *KCNJ18* and *MAP2K3* (refs. 19,20). Although long reads generally have fewer mapping errors than short reads, long reads can result in more false positives than short reads when the individual has an extra copy owing to population variability and/or reference errors. For example, in *KMT2C* in ref. 21, short reads from the extra copy may remain unmapped if the extra copy of the duplication is highly diverged from the reference (has many variants relative to the reference copy). Furthermore, GRCh37 and GRCh38 have gaps around segmental duplications that can cause mapping errors (Fig. 2d). However, improved references such as T2T-CHM13 and pan-genome references can eliminate many but not all of these false positives and mapping errors. Finally, segmental duplications can undergo gene conversion events, in which the sequence in one copy replaces the sequence in another copy relative to the reference (Fig. 3c), resulting in mis-mapping of reads and inaccurate variant calls when these are polymorphic in the human population, such as in the medically relevant genes *RHCE* and *SIGLEC16* (ref. 21). The gene conversions were recently implicated in the increased mutation rate seen in assemblies of segmental duplications²². Multiple methods have been developed to characterize copy numbers of segmental duplications from short reads²³, but only recently have long reads become

sufficiently accurate to characterize both small and large variation using haplotype-resolved de novo assembly, and standards for representing these variants are needed^{9,24,25}.

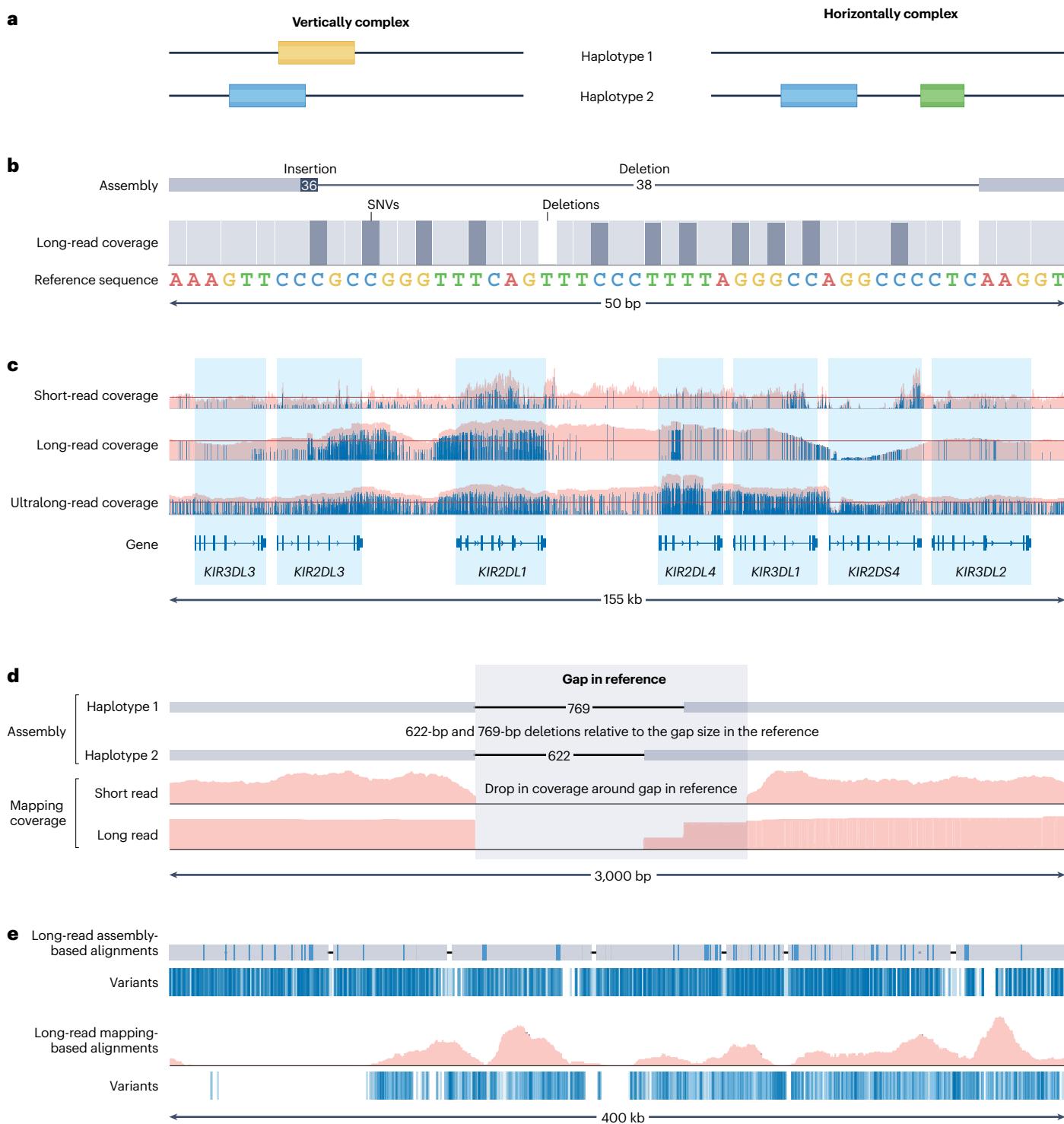
Centromeres, heterochromatin and satellites

Satellite DNAs (long arrays of nearly identical tandem repeats) are enriched within human centromeres, pericentromeric heterochromatin regions and the short, acrocentric arms of some chromosomes. These regions have typically been ignored by variant callers because they were missing or incomplete even in the reference. However, high-resolution maps of these regions were revealed in 2022 by the T2T Consortium in the first complete assembly of a human genome (representing an effectively haploid CHM13 cell line, or T2T-CHM13 v2.0)^{6,26}.

Pairwise alignments of repeat copies within the largest arrays (typically greater than 500,000 bp) are observed to be highly similar to one another (in the range of 98–100% identity). Furthermore, by mechanisms of repeat expansion, highly similar repeats – if not exactly identical repeat units – are often organized in close proximity to one another. Importantly, satellite DNAs are expected to evolve rapidly, with extreme variation in the length of the array (copy number of tandem repeats), differences in repeat unit structural variation (some repeat units within an array or between arrays may vary in length), and sequence variants that distinguish one copy of a repeat from another. These unique genomic features present a fundamental challenge to generating confident short-read or long-read alignments within an array or between homologous arrays. The small number of sparsely organized differences (single-nucleotide variants (SNVs) and insertions/deletions (indels)) can present extensive regions where repeats are indistinguishable from one another. New methods are needed to ensure meaningful assembly-to-assembly comparisons between arrays. A study of diverse centromeric satellite arrays on haploid Xchromosomes²⁶ revealed considerable variation in overall array length, regions of recent duplication (with several arrays reporting large internal duplications greater than 100,000 bp) and exceptional complex variation with local conversion and rearrangement. It is difficult to align two distinct satellite arrays that vary in repeat content in a meaningful way. Alignment strategies that are not repeat-aware (such as minimap2) align the HG002 assembly to CHM13 with many SNVs but few SVs, whereas minimap2 aligns HiFi reads with many copy number variants (CNVs) and fewer SNVs (Fig. 2e). In addition to the alignment challenges, some highly repetitive satellite DNA sequences, such as HSat2 and HSat3 arrays, were found to have strand bias and shorter read lengths for ONT and higher-than-normal coverage for HiFi, and conversely, coverage was lower than normal for both HiFi and ONT in AT-rich HSat1 arrays²⁷. All of these challenges need ongoing methods development, but some analyses of these regions are now possible with the advances in sequencing and variant-calling methods discussed next.

Advances in sequencing methods

Advances in sequencing technology have improved variant-calling accuracy and helped to identify variants in complex regions of the genome and challenging variant types²⁸. Sequencing read length and base-calling accuracy partly determine whether a variant can be correctly identified within a specific genomic context. Highly similar genomic regions are generally inaccessible by sequencing technologies with sequencing accuracy less than the paralogous region similarity, reads shorter than the distance between PSVs, or reads shorter than the paralogous regions²⁹ (Fig. 1e). Insufficient read length or sequence accuracy can cause incorrect or ambiguous mapping of reads



to a reference, as well as inaccurate or broken de novo assemblies^{30–32}. Improper read mapping can cause false-negative variant calls when reads are unmapped or ambiguously mapped, or false-positive variant calls when reads are mapped in the wrong location³³. Highlighting the clinical importance of these errors, one study found that one in seven pathogenic variants can be difficult to detect with standard

short-read sequencing¹⁷. Sequencing methods have evolved to address these limitations by increasing accuracy, read length or both²⁹.

Short-read sequencing technologies were the first next-generation sequencing methods and have made whole-genome sequencing economical³⁴. Short reads are commonly used in large population genomics studies, basic research and clinical laboratories^{35–37}. PCR-free short

Fig. 2 | Remaining challenges in representing and benchmarking complex variants.

Benchmarking challenges are primarily due to a lack of reliable benchmarks and/or comparison tools for benchmarking. **a**, Diagram showing vertically and horizontally complex variants, in which nearby variants occur on opposite haplotypes or the same haplotype, respectively. When complex variants happen in genes, knowing that they are vertically complex can be important for knowing that both copies of the gene have lost function. **b**, A small horizontally complex variant that is represented as an adjacent 36-bp insertion and 38-bp deletion in the assembly alignment and as single-nucleotide variants (SNVs; dark grey bars) and 1-bp deletions (white bars) in the high-fidelity (HiFi) long-read alignment. Comparing different representations of variants requires sophisticated benchmarking tools such as *hap.py* and *vcfeval* for small variants, or *Truvari* for larger variants. **c**, Different copy numbers of a large genomic region in the sample being sequenced, relative to the reference genome used for comparison, can

create mapping and variant-calling errors. Shown are sequencing data for the HG002 human genome killer immunoglobulin-like receptor (KIR) region, which is partially duplicated relative to the GRCh38 reference. As reads from the duplicated sequence are incorrectly mapped, this causes dense false heterozygous variants (thin blue bars in the sequencing tracks), as well as coverage (pink histogram) higher than the average coverage (horizontal red lines) (Fig. 3b). **d**, Differing alignments around a gap in the GRCh38 reference in the *CIR* gene, where short reads do not align to the gap but assemblies and long reads align across the gap with deletions and many SNVs. Benchmarking tools currently do not work robustly in these regions. **e**, When aligning the X chromosome HSat region of the HG002 genome to the X chromosome of the CHM13 reference genome, the assembly (top) is aligned to the reference very differently from HiFi reads with standard mapping methods (bottom). The HiFi reads have highly variable coverage (pink histogram), resulting in highly discordant variant calls (vertical blue bars).

reads reduce indel errors in homopolymers (Fig. 1a,b) and tandem repeats shorter than the read length. New short-read technologies are promising error rates of 1 in 10,000 bp, with particular performance gains in homopolymers³⁸. Higher accuracy may be particularly important in calling somatic and mosaic variants in a small fraction of the reads and in calling variants in regions prone to systematic errors, such as homopolymers. Regardless of base accuracy, the short read lengths hinder variant calling in large tandem repeats (Fig. 1d) and highly homologous regions, such as segmental duplications (Fig. 3), and in the highly variable, medically relevant human leukocyte antigen (HLA) gene region, which encodes several immune system components¹³³. To improve calls in some of these regions, new sequencing library preparation methods were developed. Paired-end sequencing of reads several hundred base pairs apart and mate-pair sequencing of reads several thousand base pairs apart improve mappability and SV calling³⁹. Still, SVs are harder to detect using short-read sequencing data, particularly in repetitive regions enriched for SVs⁴⁰. To further improve mappability and SV detection with short reads, library preparation methods such as linked reads, synthetic long reads, Hi-C and Strand-seq have been developed to incorporate much longer-range information.

Linked-read and synthetic long-read methods leverage high-throughput short-read sequencing technologies but add barcodes identifying reads originating from the same long DNA molecule, further improving mappability and phasing^{41,42}. For these methods, large genomic DNA molecules, tens to hundreds of kilobases in length, are partitioned and barcoded. Barcodes enable read mapping in larger repeat regions such as segmental duplications, resulting in improved variant calling in these regions. Because reads sharing the same barcode within a region are generally from a single molecule, many heterozygous variants can be phased (that is, determining whether they occur on the same haplotype or on opposite haplotypes). Some of the first haplotype-resolved assembly methods were also developed with linked reads, although these methods have substantially lower contiguity than more recent methods using long reads⁴³. Alternative linked-read methods used dense sequencing of each molecule to enable local assembly of reads into ‘synthetic long reads’ (for example, *moleculo* and more recently Complete Long Reads from Illumina). Linked reads generally have more DNA amplification biases in homopolymers and tandem repeats, and mapping or assembly challenges in tandem repeats and tandem duplications. The initial linked-read methods have been discontinued as commercial products, but new methods have been released (such as *TELL-seq*)⁴⁴ or announced (such as Complete Long Reads).

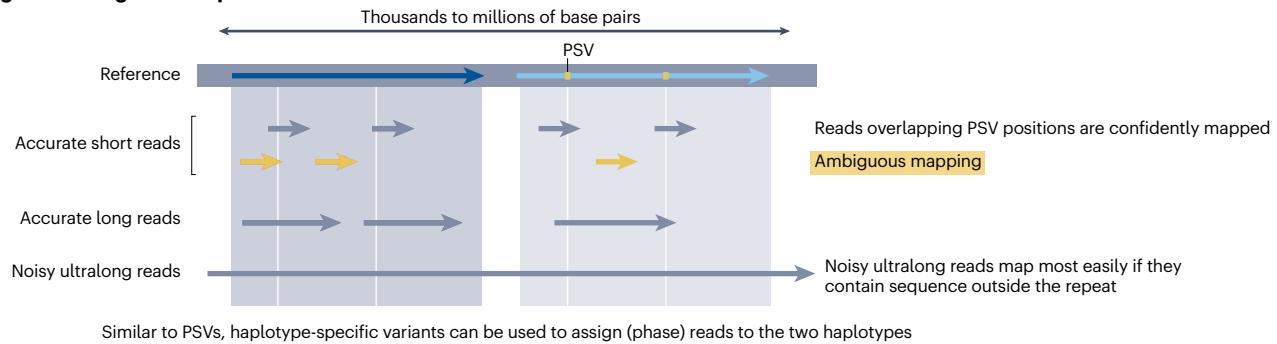
Similar to linked-read sequencing methods, Hi-C and Strand-seq are sequencing methods that combine novel library preparation

methods with the high throughput of short-read sequencing methods to increase the amount of genomic information within a read. Hi-C sequencing methods provide chromosome contact information by crosslinking segments of DNA that are in close physical proximity⁴⁵. Although Hi-C was originally developed and is still used for analyses of 3D genome organization, the chromosome contact information is also widely used for phasing variant calls as well as for scaffolding and phasing of genome assemblies^{46,47}. Strand-seq is a single-cell strand-specific sequencing protocol that tags reads by direction of sequencing (in the same direction as the reference or in the reverse complement direction), and enables clustering of reads by haplotype up to the length of chromosomes⁴⁸. The resulting strand-specific information is particularly useful with the identification of large inversions along with phasing variants and genome assemblies⁴⁹.

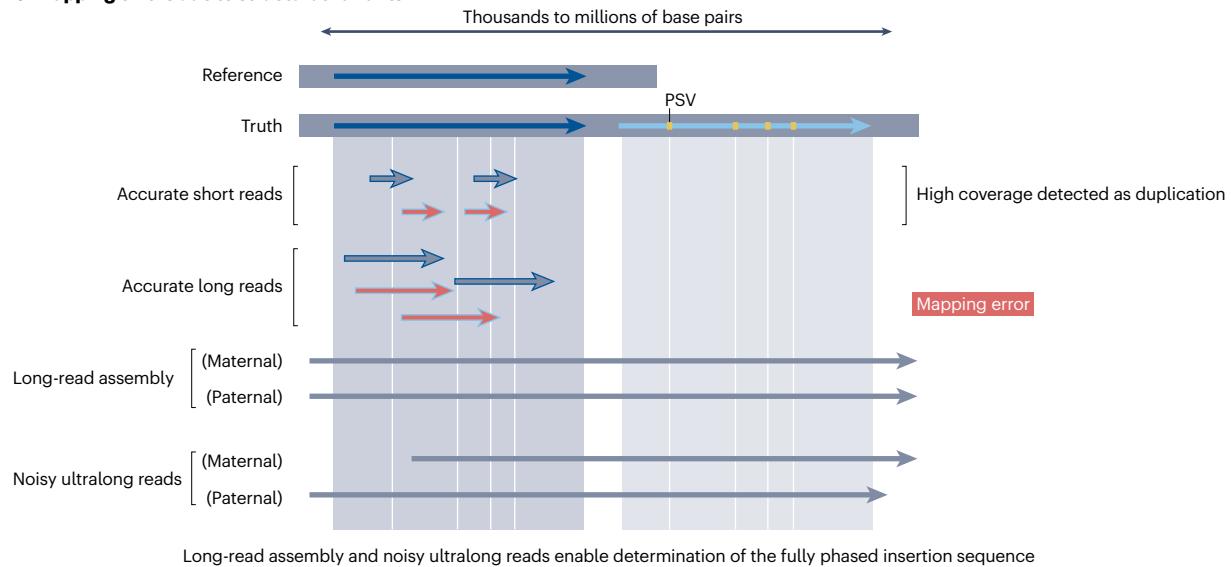
Increased read length improves mappability, expanding callable regions, and increases genome assembly quality. PacBio released the first long-read sequencing product^{50,51} followed by ONT⁴. Often referred to as third-generation sequencing, these new methods initially offered longer reads but with lower sequencing accuracy, lower throughput and higher cost compared to short-read sequencing methods. The lower sequence accuracy limited accuracy of calling small variants, but the longer read lengths substantially improved genome assembly and SV calling relative to short reads^{52–55}. Recent improvements in read accuracy have made calling of small variants possible, particularly with the new PacBio HiFi approach based on circular consensus sequencing⁵. The unique combination of read length and accuracy in the HiFi data has resulted in high-accuracy calling of small variations and SVs as well as diploid genome assembly, as discussed below. Although ONT reads are still less accurate than HiFi reads, their accuracy has been steadily improving through improvements to the sequencing methods (including new pore designs, duplex sequencing reading the same molecule twice, and base-calling methods), enabling accurate SNV calling as well as SV calling^{56–58}. The unique pore-based electrical signal detection method from ONT has allowed for the generation of sequencing data with read lengths >2 Mb, and data sets with read length N50>100 kb (refs. 52,59,60). Combined, ultralong ONT and HiFi data sets have enabled the generation of the first complete (‘telomere-to-telomere’) human genome assembly⁶.

To complement sequencing technologies, optical⁶¹ and electronic mapping⁶² technologies measure the spacing between sequence motifs that are marked on long DNA molecules and were recently reviewed⁶³. These technologies do not give sequence-level information, but can enable de novo assembly and detection of large SVs that are challenging to detect with sequencing, because they start with long DNA molecules⁶¹. Optical mapping has also been used to scaffold and correct

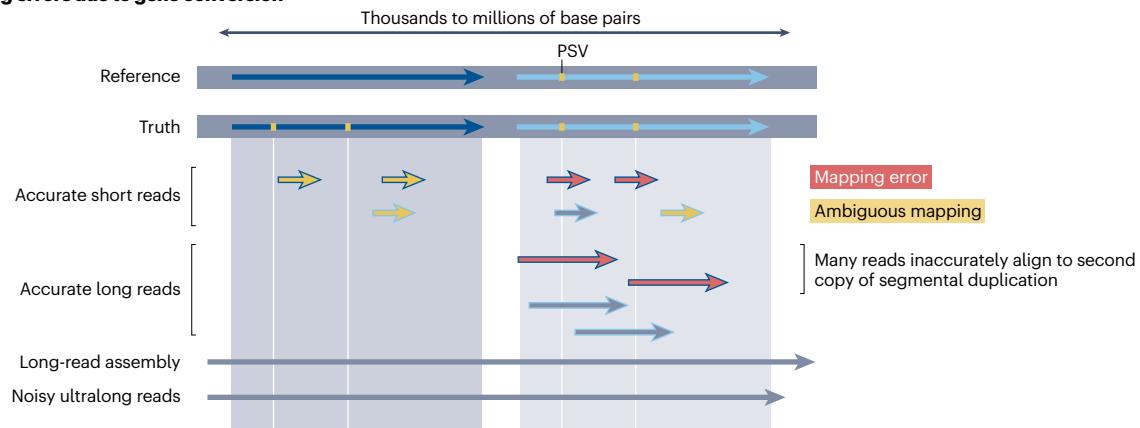
a Mapping errors in segmental duplications



b Mapping errors due to structural variants



c Mapping errors due to gene conversion



assemblies of sequencing reads, as well as for detection of large germline SVs and CNVs associated with diseases⁶⁴ and of somatic structural changes in cancer⁶⁵. We next discuss methods developed to leverage new sequencing technologies to improve variant calling and thus benchmark creation.

Advances in variant-calling methods

Variant-calling pipelines typically have included mapping (or aligning) sequencing reads to the reference genome assembly, and then identifying differences between these reads and the reference genome, typically called variants, and represented in the variant call format

Fig. 3 | Mapping challenges in segmental duplications and large structural variants. **a**, The diagram shows a highly identical segmental duplication that is larger than short or highly accurate long reads but shorter than noisier ultralong reads. In this figure, the segmental duplications are close to each other (tandem duplications), but they can also be distant. Short reads generally cannot map reads or call variants with confidence except very near paralogous sequence variants (PSVs; marked by vertical white lines) that differentiate the duplicated sequence. False positives result from reads mapped to the other copy of the segmental duplication (ambiguously mapped reads are shown as yellow arrows). Sophisticated long-read mappers can use nearby PSVs to align the reads and call variants on the correct copy of the segmental duplication. Ultralong reads can also correctly align across the segmental duplication and flanking sequences despite their higher error rate, but sophisticated variant-calling methods are needed to distinguish true variants from sequencing errors. **b**, A large structural variant, specifically a tandem duplication, is in the individual but not the reference. When short or long reads are shorter than or about the same size as the duplicated region, reads from the duplicated sequence (indicated by red

arrows with light blue outlines) are often mapped to the existing sequence in the reference, resulting in higher-than-normal coverage and denser variants due to PSVs in the new duplication (the true repeat origin of the short and long sequence reads is shown by the outline colour of the arrow). When long-read assemblies or ultralong reads traverse the duplication and flanking sequences, the duplicated sequence can be detected as an insertion of sequence similar, but not identical, to the reference sequence. **c**, Gene conversion, in which the second copy (light blue arrow) of the segmental duplication replaces the sequence of the first copy (dark blue arrow), such that the first copy now contains the PSVs from the second copy. This typically results in no short reads mapping confidently to the first copy of the segmental duplication, as there are no PSVs from it. Long reads that are shorter than the duplication may map confidently to the second copy because they do not contain the PSVs, even if they actually originate from the first copy (red reads). When long-read assemblies or ultralong reads traverse the entire region and flanking sequences, then the variants may be detected accurately across the entire region. Note that this figure is simplified to show only one haplotype, but a genome from an individual generally has two haplotypes that contain different variants.

(VCF)⁶⁶ (Fig. 4, left). These candidate variants may be true variants in the individual or errors in mapping or sequencing, so variant callers typically include a filtering step to remove or flag likely false positives or uncertain variants. Errors in variant calls can arise at each stage of genome analysis, from library preparation to sequencing to mapping to variant calling. Errors and variability introduced in each of these stages are often interrelated, and we detail different sources of errors in Supplementary Box 1, which are summarized in ref. 67. Because the focus of this Review is on germline variant calling, we refer the readers to a recent review of somatic variant calling for a comprehensive examination of challenges specific to somatic variants⁶⁸. A summary of advances in calling de novo mutations, variants in RNA sequencing and other specialized data types is in Supplementary Box 2. In the following sections, we describe recent advances in variant calling that take advantage of new sequencing methods to access challenging genomic regions.

Mapping reads

There are methods specially designed for hard-to-map regions of the genome. As described above, even highly similar segmental duplications have some differences (for example, PSVs) between them. These PSVs can be used to distinguish different mapping locations (Fig. 3a,b), but are not used explicitly by most mapping algorithms. Winnowmap2 solves this problem using minimal confidently alignable substrings, which are substrings in reads used to establish mapping confidence by comparing the highest-scoring location with the second-highest scoring⁶⁹. By requiring this confidence to be above a user-defined threshold, this naturally leverages PSVs between alignment locations. This generally leads to an order-of-magnitude improvement in the false-positive rate and false-negative rate compared to other mappers such as winnowmap v1 (ref. 70), minimap2 (ref. 71) and NGMLR⁷². DuploMap takes a different approach using both *a priori* PSVs (identified from the University of California Santa Cruz (UCSC) Table Browser) and calculation of longest common subsequence between reads and alignment locations (the assumption being that correctly aligned reads should share long and unique sequences with the mapped location)⁷³. The authors show a large improvement for PacBio HiFi and ONT reads in improving mappability to segmental duplications, which also improves variant calling^{1,58}.

Filtering errors with deep learning and other machine learning

Filtering out false-positive variants based on characteristics of the reads and repeats in the genome sequence has been important since

the first variant callers. Filtering methods have become increasingly sophisticated by improving expert-designed features and using new methods such as deep learning. As seen in the precisionFDA Truth Challenge V2 (ref. 1), deep learning implementations are becoming a prominent approach in many variant-calling methods^{74,75}, particularly for newer sequencing technologies. To develop a thorough contextual understanding of these developments, we direct readers to reviews of best-practice considerations⁷⁶ and applications⁷⁷ for machine learning methods in bioscience. Whole-human-genome small-variant calling is amenable to deep learning because of openly available sequencing data and benchmarks for training and testing that cover millions of variants in a range of genome contexts. The dominant architecture for variant-calling deep learning models is currently convolutional neural networks (CNNs). Traditional variant callers use expert-designed features about the sequence of the locus (for example, whether it is a homopolymer), as well as characteristics of the reads from a sample aligned to the locus (for example, whether the variant is strand biased) (Fig. 1c). To reduce the need for expert-designed features, a CNN architecture accounts for information from sequencing reads and sequence of the reference genome at and around the variant. With appropriate training hyperparameters, a CNN can approximate a complex, nonlinear function that classifies loci as homozygous variant, heterozygous variant or homozygous reference (non-variant), often yielding empirically accurate performance metrics for short-read and long-read sequencing technologies.

Although neural networks do not need as many expert-designed features as other methods, it is still important to represent relevant features of alignments in the input data when designing variant callers. For example, in the precisionFDA challenge, the DeepVariant input format includes read base, base quality, mapping quality, strand, reads that support variant, base differs from reference, and insert size, whereas NeuSomatic used a reference sequence along with alignment features¹. Looking beyond CNNs, recurrent neural network architectures that account for the sequential structure of the genome are used in prediction over sequencing reads such as in DeepConsensus⁷⁸. Overall, deep learning techniques have been particularly important in enabling rapid adoption of new and evolving sequencing technologies for variant calling, such as PacBio HiFi, ONT and new short-read technologies^{58,74,79–82}.

Other machine learning techniques are also commonly used to filter potential false-positive variant calls. Examples include the Gaussian Mixture Model used in Variant Quality Score Recalibration in the

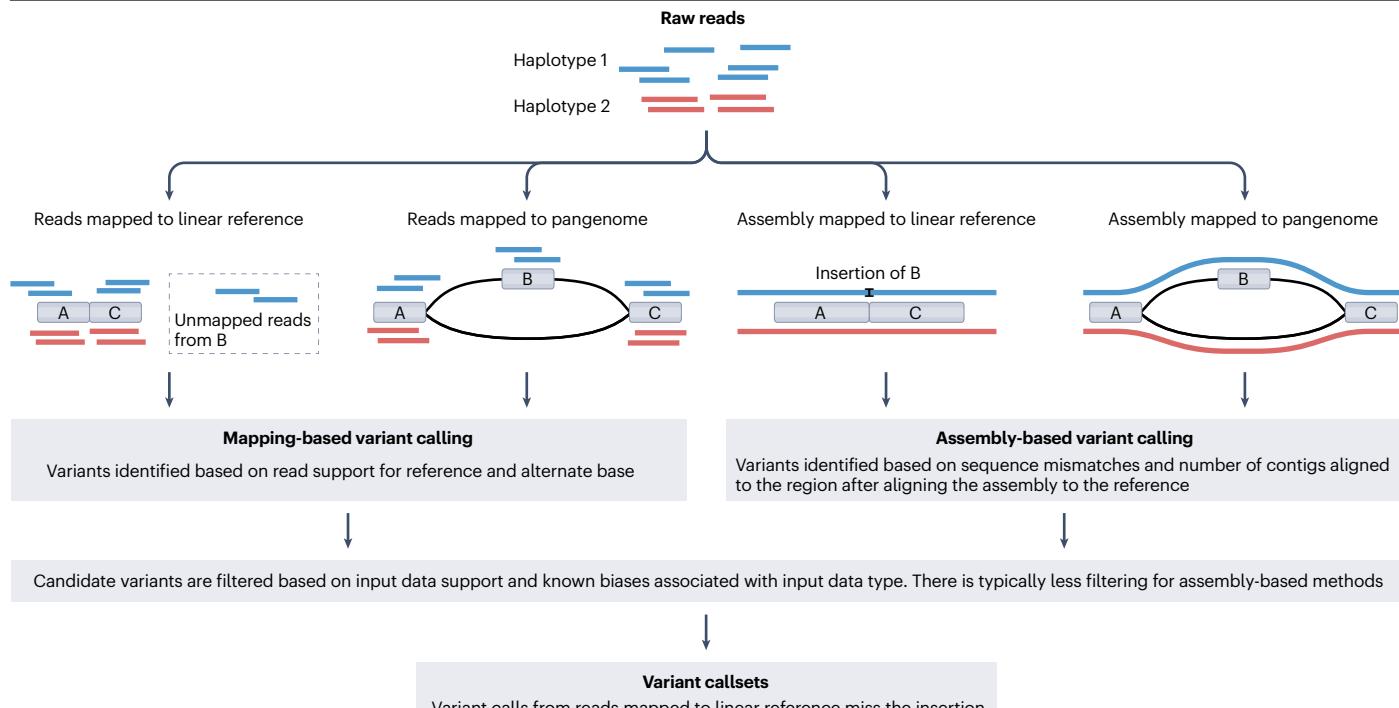


Fig. 4 | Workflow strategies for variant calling. Diagram of four variant-calling workflows. Workflows are distinguished by read alignment method – read mapping (the two routes on the left) versus de novo assembly (the two routes on the right) – and reference genome structure (linear versus pangenome). Raw reads or a de novo assembly are mapped to the linear or pangenome reference.

in this case depicted as a graph. The alignments are then used for variant calling and subsequent filtering. Raw reads are from two haplotypes and align to genome segments A, B and C. Segment B is present in the pangenome but not in the linear reference. In this case, the large insertion of segment B is missed by reads mapping to the linear reference because the reads from segment B remain unmapped.

Genome Analysis Toolkit (GATK)⁸³, the optional random forest classifier in Octopus⁸⁴ and Gradient Boosting Machines in DNAscope⁸⁵. These models typically require feature engineering, so features associated with errors in each new technology need to be designed. Although engineering features can be challenging, these models can also give an indication of which features are important in filtering a variant – a task that is more challenging with deep learning models.

Given the complexities of variant calling, it is important that method developers summarize the model, training and benchmarking for users to understand the intended use and limitations of a method. We recommend that variant detection method developers adopt some of the transparency approaches used in machine learning, as described in Box 1. This transparency is useful for all variant callers, even when not based on deep learning, and it is particularly important for clinical applications, as one recent study showed improved variant detection with deep learning-based methods⁷⁵. Also, other machine learning disciplines use model zoos for distributing trained models, and genomic variant callers could similarly benefit from using Kipoi or a similar mechanism⁸⁶.

Tandem repeat-specific callers

A few specialized methods are designed to cope with STRs and their repetitiveness, because these regions are often ignored or mis-called by standard small-variant and SV callers owing to the size of alternative alleles, repeat structure, or complexity. ExpansionHunter is a short-read variant caller that genotypes STRs by using a predefined variant catalogue, encoding the structure of the repeated loci in question using

a regular-expression-like syntax⁸⁷. ExpansionHunter has been shown to outperform other STR-specific variant callers such as HipSTR⁸⁸, gangSTR⁸⁹ and TREDPARSE⁹⁰. However, ExpansionHunter matches STRs based on user-defined patterns, requiring the user to know which variants they are targeting, and it is also less flexible regarding point mutations in the repeat units themselves. Other tandem repeat-specific callers have been developed for forensic STRs⁹¹, ONT reads⁹² and PacBio HiFi reads⁹³.

Phasing and haplotyping

Phasing entails assigning heterozygous variants, reads or assembled contigs (contiguously assembled sequences of tens to thousands of kilobases) to the haplotype coming from the father or mother. When sequencing data are available from the parents, haplotypes can often be assigned as originating from the mother or father, or common variants can be phased based on large population panels. Otherwise, nearby variants can be phased locally when heterozygous variants are within the read length or paired-end distance for standard sequencing methods. As described above, specialized library preparations, including linked reads, Hi-C and Strand-seq, have been developed to phase variants and assembled contigs at longer scales up to entire chromosomes. Several tools, including WhatsHap and HapCut2 (refs. 94,95), have been developed to phase variants using a variety of types of sequencing data and pedigree information. WhatsHap can also use phased variants to assign long reads to each haplotype, which is often helpful for calling variants and visualizing read support for variants⁹⁴. Other tools include phasing steps to improve variant calling or assemblies^{25,46,49,96,97}.

It is important to note that short-read-based phasing often only locally assigns variants to haplotype blocks (that is, the sub-region where variants are phased together), with long reads and linked reads producing larger haplotype blocks than short reads, but there may be switches between blocks. Local phasing can particularly be important for clinical applications – for example, to understand whether two loss-of-function variants in a gene occur on the same haplotype, so that only one copy of the gene is non-functional, or on opposite haplotypes, so that both copies are non-functional.

Haplotype-resolved de novo assembly

De novo assembly is an increasingly possible alternative to read mapping that involves stitching together reads independently of a reference genome. From short reads, only relatively short contigs and nonrepetitive regions can be assembled, so it was rarely used except for regional assembly of large variants. Long reads enable assembly of much longer contigs. However, before the advent of highly accurate long reads in 2019 (ref. 5), even the best assemblies collapsed haplotypes in most regions of the genome and had many small indel errors from the noisy reads, so they were not useful for small-variant calling^{98,99}. With the advent of accurate long reads from PacBio HiFi, haplotype-resolved (or ‘diploid’) assembly across much of the genome became possible^{9,25,46,49,100} and, combined with new methods, enabled accurate small-variant calling. Initially, assemblies still collapsed many segmental duplications, resulting in missed and inaccurate variants. However, many segmental duplications were resolved accurately by the best methods submitted to a 2020 comparison of diploid assemblies for HG002. This comparison showed the progress in assembly metrics and established assembly-based variant calling as a leading approach to characterize small and large variation of a genome, closely matching curated benchmarks¹⁰¹. In 2022, the T2T Consortium published a complete effectively haploid human assembly as well as complete chromosome X assemblies in males^{6,26}, and they are currently working towards complete haplotype-resolved assemblies of diploid humans. The Human Pangenome Reference Consortium (HPRC) recently released high-quality diploid assemblies for the purpose of building a pangenome reference, as described in the next section¹⁰². Diploid assemblies currently are the best method to resolve complex variants in the most repetitive regions of the genome, such as for segmental duplications and satellites, but much work remains to understand and benchmark these variants, with pangenome alignments providing one path.

Pangenomes and graph-based variant calling

Traditionally, variant calling in human sequences has been performed by aligning reads to a single linear reference such as GRCh37, GRCh38 or, most recently, T2T-CHM13. However, this approach is limited in regions where an individual differs substantially from the reference, such as large indels, SVs, CNVs and other highly variable regions such as the medically relevant HLA (reviewed elsewhere¹⁰³) and killer immunoglobulin-like receptor (KIR) regions. To address this challenge, approaches have been developed to map reads to pangenomes, often using graph-based references that incorporate variants from many individuals as different reference paths. For small variants, read alignments to the graph-based pangenome reference are typically translated into alignments to a linear reference so that normal variant-calling tools can be used to generate variant calls on the linear reference. Pangenome approaches were recently reviewed¹⁰⁴. Strengths and weaknesses of linear reference and pangenome reference approaches are shown in Table 1, and Fig. 4 depicts an example

of a large insertion in an individual relative to the linear reference, and how mapping to a pangenome enables reads to be mapped to the inserted sequence.

The first human graph-based references incorporated small variants and/or SVs from short-read population sequencing projects such as the 1000 Genomes Project and showed improvements in variant calls particularly for larger indels and SVs¹⁰⁵. Recently, this approach was shown to improve mapping statistics and increase the number of variant calls in individuals of African ancestry¹⁰⁶. Individuals with African ancestry generally have more variants and higher diversity than other populations, so it is plausible that graph-based references may particularly improve accuracy of variant calls for African individuals, but the lack of benchmarks for African samples makes it challenging to understand the accuracy of new variant calls detected by pangenome methods. With the advent of long reads applied to diverse samples,

Box 1

Transparency for variant detection methods and pangenome construction

It is increasingly important to summarize attributes of variant detection methods using transparency techniques such as Model Cards, Transparency Notes and AI360 (ref. 160). Historically, variant detection methods relied on classic statistical models, but they increasingly rely on models trained using machine learning techniques. Benchmarking allows for comparison between variant-calling methods based on performance metrics, including stratifying by genomic context to understand strengths and weaknesses. A complementary approach to compare methods is evaluating the algorithmic performance and modelling approach¹⁶¹. Modelling inherently relies on assumptions about the data and estimated function characteristics as well as hyperparameter selection, which impacts the results of machine learning solutions. We expect that adopting prominent transparency approaches from the machine learning community could enable improved comparison of variant detection models. We propose that developers of both statistical and machine learning-based variant detection methods use these transparency techniques to explicitly summarize characteristics and limitations of the training and test data, model attributes, hyperparameter search space explored, known biases or limitations of the method, data used in graph-based reference genomes, and expected use cases. Making transparent both how benchmark sets are used and the attributes of the model helps users to determine the best approach for their application. Similarly, transparency for samples used in pangenome graph construction as well as sequencing data and parameters used to generate input haplotypes will be increasingly important moving forwards. For example, samples used for benchmarking should be excluded from the graph to avoid biases. A minimum communication method such as the discussed transparency techniques will mitigate potential issues regarding interpretation and reproducibility of resulting variants when moving from a linear reference to pangenome graphs.

Table 1 | Comparison of variant-calling process from mapping, graph reference and diploid assembly

Variant-calling process			
Input sample data	Raw/preprocessed whole-genome sequencing or targeted sequencing reads		De novo assembly
Reference type	Linear	Graph/pangenome	Linear or pangenome
Sequence alignment: strategy	Read–reference genome alignment (mapping)		Assembly–reference genome alignment
Sequence alignment: example tools	bwa-mem ¹⁵⁷	Seven Bridges GRAF ¹⁰⁵ Dragen graph variant-calling pipeline ¹ Giraffe ¹⁰⁸	minimap2 (ref. 71) MUMmer ¹⁵⁸
Variant detection: strategy	Variants identified based on read support for reference and alternate base		Variants identified based on assembly-to-reference alignment, including sequence differences and large structural changes
Variant detection: example tools	GATK ⁸³ DeepVariant ⁸²	Seven Bridges GRAF ¹⁰⁵ Dragen Giraffe-DV ¹⁰⁸ GraphTyper2 (ref. 159)	dipcall ¹²³ PAV ⁵⁵ MUMmer ¹⁵⁸ SVAnalyzer (structural variant calling) ¹¹⁷
Variant filtering	Candidate variants are filtered based on input data support and known biases associated with input data type. There is typically less filtering for assembly-based methods		
Strengths	Works with short or long reads Less computationally intensive High accuracy for easy regions Mature infrastructure Extensive reference annotations	Works with short or long reads High accuracy for easy regions and some structural variants	Phased small-variant and structural variant calls (for diploid assemblies) Ability to call small variants and complex structural variants in very difficult regions, although still limited by insufficient standards for representing complex variants and copy number variants
Limitations	Low accuracy for difficult regions of the genome Limited accuracy for structural variants	More computationally intensive Infrastructure and tools still being developed No standard reference graph genome Information may be lost when translating variants to a linear reference genome	Requires long reads More computationally intensive Variant-calling accuracy is dependent on assembly quality, particularly for homopolymers and tandem repeats Currently worse in highly homozygous regions

several approaches were developed to incorporate SVs discovered with long reads into the graph, which then enables genotyping of many of these SVs with short reads, although genotyping the majority of SVs is still challenging because they are located in tandem repeats. Recent graph-based references were made from long-read de novo assemblies by the Human Genome Structural Variation Consortium (HGSVC) and the HPRC. The 2019 HGSVC assemblies were based on long reads with ~10% error rate, so did not enable full haplotype separation, full assembly of segmental duplications or accurate SNV calling, but they still substantially improved SV genotyping by short reads^{55,107}. The 2022 HPRC Phase 1 assemblies used more accurate HiFi reads, which enabled higher resolution of small and large variants, as well as better assembly of each haplotype and segmental duplications. This, along with refined graph-based variant-calling methods (Giraffe-DeepVariant)¹⁰⁸ and genotype inference methods (PanGenie)¹⁰⁷, enabled further improvements in variant calling, particularly for large indels, SVs, highly polymorphic regions and regions with errors in the GRCh38 reference^{107–109}. Although a pangenome reference can include new segmental duplications and complex structural variation missing from a linear reference, rare variation may not be represented even in future pangenome reference resources. Calling rare complex variants, such as those causing rare disease, may still require assembly-based approaches, although interpretation of the resulting assemblies often benefits from alignment to the pangenome reference.

Alternatives to graphs have been proposed when using pangenesomes to improve variant calling. Low-frequency variants (that is, variants observed with <1% allele frequency) are typically observed to be regional or specific to a given population¹¹⁰. Therefore, there could be an advantage in using a reference that contains major alleles in the population¹¹¹ or that most closely corresponds to the ancestry of the sequenced individual, so that existing mapping and variant calling tools designed for a linear reference can be used¹¹². This approach may not be sufficient for individuals with highly admixed ancestries or for regions of the genome that are highly variable between individuals in a population, such as the HLA and KIR regions, and some segmental duplications. A ‘reference flow’ approach progressively aligns reads to multiple reference genomes to improve variant calling, which is faster than graph-based approaches and therefore may enable the use of a larger number of reference genomes¹¹³.

Looking forward, alignments between pangenome assemblies are likely to be important for understanding the complex variation that frequently occurs in segmental duplications, VNTRs, satellite DNA and other repetitive regions. Standardizing alignments between genomes in complex regions and representation of variants in these regions will be important for benchmarking accuracy and for understanding the clinical relevance of complex variants. Many pangenome methods are under active development, and more comprehensive benchmarks and sophisticated benchmarking tools will be needed to assess their improvements in the most challenging regions, which we discuss in the following sections.

Benchmarking variant accuracy

Introduction to benchmark sets

Shared benchmark sets play a crucial role in advancing genomic science, in wet-lab and dry-lab technology development, and in confident application of genome sequencing. Little progress can be made without the ability to compare performance metrics from different approaches. These widely available benchmarks are the foundation for such comparison and can be the basis for translation to clinical use in regulated applications. A few extensively characterized genomes have been developed into ‘benchmarks’ to understand the performance of variant-calling methods. These widely available genomic resources are composed of both the genomic DNA or cell lines containing the genome and extensive data from multiple DNA sequencing technologies. These data are used to form a ‘benchmark set,’ the preferred term from GIAB and GA4GH to describe the set of variants (represented in VCF) and the regions in which practically all variants have been characterized (represented in browser extensible data (BED) format)^{114–117}.

‘Benchmark set’ reflects the intended use, but synonyms in the literature include ‘truth set’¹¹⁸, ‘high-confidence variants and regions’^{119,120}, ‘baseline variants’¹²¹ and ‘gold standard’^{53,121,122}. These benchmark variants and regions enable users to identify true positives (correctly called variants), false positives (incorrectly called variants) and false negatives (missed variants) (Fig. 5a,b).

There are multiple reasons why companies or scientists use benchmarks, including evaluating DNA library preparation, sequencing and bioinformatics methods. Developers of sequencing technology might use benchmarks to ensure that a new instrument is working as expected, and clinical laboratories might use them to ensure that a targeted protocol is capturing all of the expected variants in clinically relevant regions. In addition to testing library preparation and sequencing methods, the benchmarks are highly valuable for developing new computational methods to identify or filter genomic variants. Here, a scientist can download existing data from benchmark samples (sometimes already mapped) or separately run their method on their own data

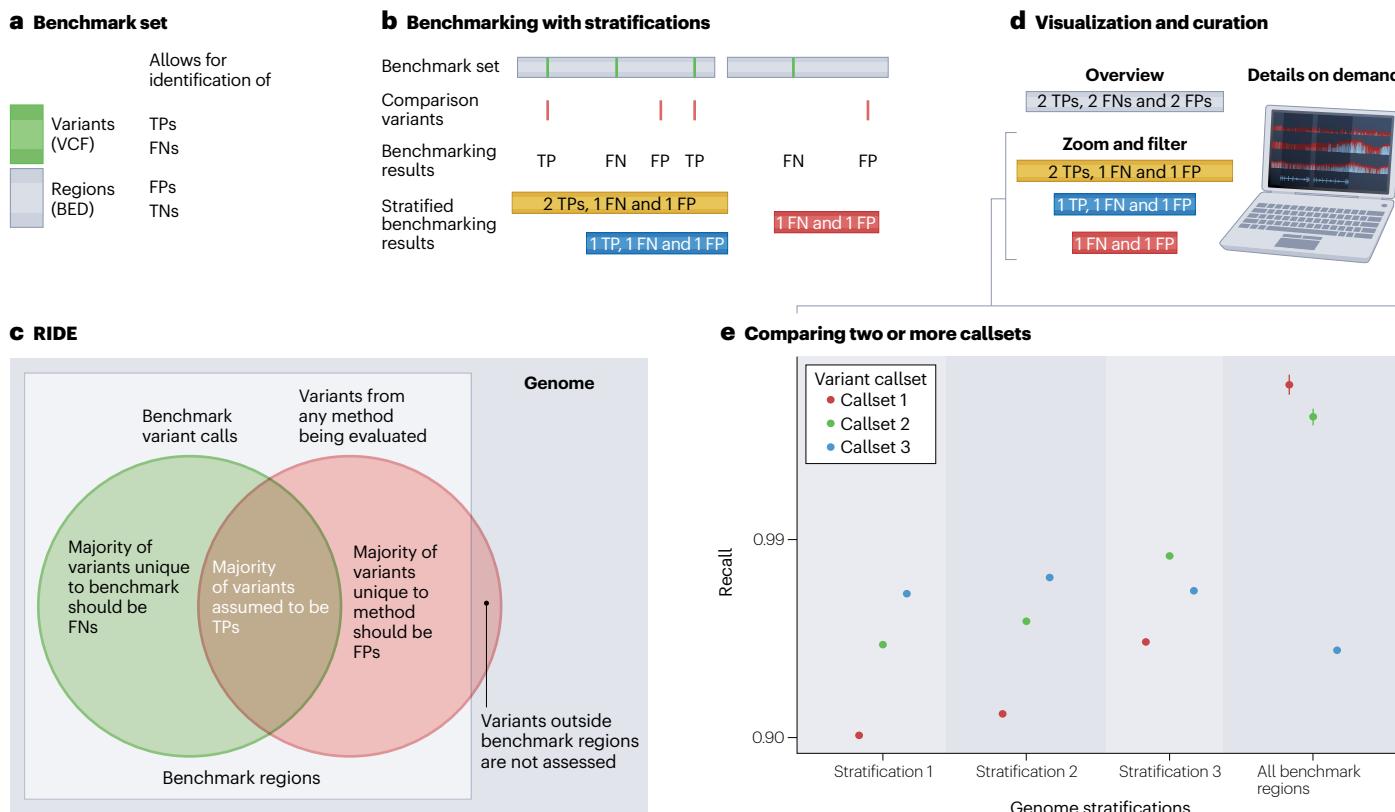


Fig. 5 | Considerations when generating and using benchmark sets for evaluating variant-calling methods. **a**, Primary components of a benchmark set, specifically the benchmark variants (in variant call format (VCF)) and the benchmark regions (in browser extensible data (BED) format). The benchmark regions exclude variants and regions that are unclear with current methods. The benchmark set can be used to identify false negatives (FNs; missed variants) and false positives (FPs; incorrectly called variants). **b**, Benchmarking variant calls using genome stratifications to provide the variant callset performance in different types of repetitive genomic contexts. **c**, Diagram of the reliable identification of errors (RIDE) principle for determining whether a benchmark is fit for purpose. Typically, when variants from the method being evaluated (pink) are compared against the benchmark (green), matching variants are assumed to

be accurate (true positives (TPs)). Variants unique to the benchmark are curated to ensure that most are actually FNs. Variants unique to the method being evaluated inside the benchmark regions are curated to ensure that most are actually FPs. **d**, Visualization and curation of benchmarking results, with three layers of analysis: (1) high-level overview (overall summary statistics), (2) zoom and filter (performance in stratifications representing different types of genomic repeats) and (3) details on demand (visualization of aligned read support for variant calls, typically in a genome browser such as the Integrative Genomics Viewer (IGV)). **e**, Example of zoom and filter, showing strengths and weaknesses of three callsets in different stratifications: callset 1 has the highest recall overall (all benchmark regions), but callset 3 has the highest recall in stratifications 1 and 2, and callset 2 has the highest recall in stratification 3. TN, true negative.

Glossary

Acrocentric arms

Short arms of human chromosomes 13, 14, 15, 21 and 22, which are known to be enriched with satellite DNA, segmental duplications and transposable element insertions. They also contain long tracts of ribosomal DNAs. They are highly similar in repeat structure and sequence content.

Admixed ancestries

Individuals with ancestors coming from multiple populations that had previously diverged.

Benchmarking variants

The process of comparing a variant callset (the query callset) to the benchmark callset in the benchmark regions in order to identify true positives, false positives and false negatives.

Benchmark sets

Set of variants and regions defined to reliably identify false positives and false negatives, also sometimes called ‘high-confidence’, ‘truth’, ‘baseline’ and ‘gold standard’.

Centromeres

Genomic regions, one per chromosome, that map the location of kinetochore assembly, typically marked as a primary constriction on a metaphase chromosome.

Circular consensus sequencing

A sequencing method in which a single molecule is circularized and sequenced multiple times to improve accuracy (for example, in Pacific Biosciences HiFi sequencing).

De novo assembly

Analysis of DNA reads to produce the genome sequence of an individual without mapping individual reads to a reference genome. Increasingly, human genome assemblies can be haplotype-resolved (phased), such that separate assembled sequences are produced for the copies of each chromosome coming from the mother and father.

Genome in a Bottle Consortium

(GIAB). A public-private-academic consortium formed by the US National Institute of Standards and Technology (NIST) in 2013, involving a broad community from government, academia, commercial technology developers and clinical laboratories. Its aim is to develop authoritatively characterized genomes that can be used to benchmark human genome variant calls.

Germline variant

A variant attributed to the initial sequence of an organism at conception, and typically found in all the cells in an individual.

Haplotype

A region of DNA containing multiple variants (or alleles) that are frequently inherited together.

Indels

Variants that are insertions and deletions of sequence, typically 1 to 49 bp in size.

Long interspersed nuclear elements

(LINEs). A family of transposons, with approximately 100,000 truncated copies and a few thousand full-length 6,000-bp copies in the human genome, causing mapping challenges.

N50

A summary measure of read length distribution: 50% of the bases in the reads are in reads longer than the N50 value. Similarly, for de novo assemblies, 50% of the bases in the assembled contigs are in contigs longer than the N50 value.

Pangenome references

Collection of many genomes used as references (sometimes, but not always, represented as graphs) in addition to the standard linear genome reference assemblies.

Pericentromeric heterochromatin regions

Typically multi-megabase-sized regions directly adjacent to centromeres that are enriched with satellite DNA, segmental duplications and transposable elements. These regions are associated with darkly staining constitutive heterochromatin.

Phasing

The process of assigning heterozygous variants to the same haplotype (for example, the maternal copy of the chromosome contains both variants) or to opposite haplotypes (one variant is on the maternal copy and the other is on the paternal copy).

Precision

The fraction of query variants in the benchmark regions that match the benchmark variants, or true positives/(true positives+false positives).

Read mapping

Aligning a given read to a reference.

Reads

Small sequence fragments from larger molecules generated by a given sequencing technology; the length can range from 100 bp to >1 million bp, depending on the sequencing method.

Recall

The fraction of benchmark variants that are matched by query variants, or true positives/(true positives+false negatives).

Reference genome assembly

A haploid genome assembly to which sequencing reads are mapped and variants are called. The current versions in common use are GRCh37 (also known as hg19), GRCh38 (also known as hg38) and T2T-CHM13.

Reference material

A material that is sufficiently stable (over time) and homogeneous (between vials) for its applications. For example, genomic reference materials from the US National Institute of Standards and Technology (NIST) are extensively characterized to develop benchmark variants and regions to reliably identify false positives and false negatives.

Satellite DNA

Highly repetitive regions that originally were defined by their density owing to a unique composition of A, C, G and T bases. Satellite DNA regions are often characterized by tandem repeats organized in very long arrays and are embedded in regions known to be enriched in silent, constitutive heterochromatin.

Scaffolding

The process of connecting assembled contigs even when the intervening sequence is unknown.

Segmental duplications

Long DNA sequences that are highly similar to each other in the reference genome assembly, typically at least 1,000 bp in length and not a transposable element, tandem repeat or satellite DNA. There is some overlap between variable number tandem repeat (VNTR) and segmental duplication annotations, particularly for tandem repeat unit sizes longer than 1,000 bp, as occurs in the medically relevant genes *LPA* and *CR1*.

Sequencing Quality Control Consortium

(SEQC). A consortium formed by the US Food and Drug Administration (FDA) to compare sequencing methods and understand sources of variability.

Short tandem repeats

(STRs). Many consecutive repeats of 2-bp to 6-bp sequence units.

Glossary (continued)

Single-nucleotide variants (SNVs). Variants that are single-base substitutions. They are also commonly called single-nucleotide polymorphisms (SNPs) when they occur at an appreciable frequency (typically >1%) in the germ lines of the wider population.	Somatic variant A variant attributed to a mutation after conception. Only some cells in the organism will have this variant; they are most frequently detected in cancer tissues or blood.	Structural variants (SVs). Typically defined as variants of at least 50 bp in size.	Variable number tandem repeats (VNTRs). Many consecutive repeats of >6-bp sequence units.
---	--	---	---

set for the benchmark samples to measure both the fraction of variants that are re-identified from the benchmark set (that is, recall or sensitivity) and the number of extra variants identified in the benchmark regions but not matching the benchmark variants (false positives). The lower the number of false positives, the better the precision or specificity of the method. Of note, benchmark samples generally should have defined benchmark regions to enable identification of false positives.

Ongoing advances in genome-wide variant calling required iterative refinement and development of the GIAB benchmarks. Developing and maintaining these benchmarks is unprecedented in the field of reference material and data development, with more than 10^9 properties in each benchmark. Complementary benchmark development efforts have used family pedigree information¹²⁰ and assemblies of two effectively haploid cell lines¹²³. Similarly, the Sequencing Quality Control Phase 2 (SEQC2) Consortium developed benchmarks for somatic variants^{124–126}. Each of these benchmarks has strengths and limitations for understanding different aspects of performance, as described below.

Principles for benchmark set design

GIAB has developed principles for designing and evaluating genomic benchmark sets. The reliable identification of errors (RIDE) evaluation process developed by GIAB ensures that its benchmarks reliably identify false positives and false negatives across a variety of methods¹²⁷ (Fig. 5c). Specifically, when GIAB develops a draft benchmark, the RIDE evaluation includes recruiting experts in a diverse set of methods from around the world to evaluate the utility of the benchmark for their particular callset. Each callset is compared against the draft benchmark, and a random set of potential false positives and false negatives from different categories are selected for manual curation. The external experts visually curate aligned short-read, linked-read and long-read data, along with annotations of repeats, and determine whether the benchmark is correct on both alleles in the region, the particular callset is correct on both alleles in the region, or if the correct answer is unclear given current technologies. They are asked to be critical of the benchmark, selecting ‘unclear’ if the benchmark is not clearly supported by the data. Then, the NIST team that developed the benchmark re-curates any locations that were not determined to be correct in the benchmark and incorrect in the query. Finally, the NIST team and external experts come to a consensus about any sites with differing opinions. All of these curations are made available with the benchmarks to transparently highlight limitations of the benchmark and particular errors or unclear regions. For variant benchmarks, it is crucial that the benchmark accurately calls both haplotypes in the region surrounding any variants, including any nearby variants and variants in the same homopolymer or tandem repeat. Otherwise, benchmarking tools may inaccurately identify false positives and false negatives. As discussed in the next section, it is also important to choose appropriate benchmarking tools and parameters for a particular benchmark set, which is why GIAB selects robust benchmarking tools as part of the RIDE evaluation.

Germline variant benchmarks

Since GIAB was formed in 2012, it has released several versions of benchmarks for seven human genomes. These benchmarks use data from multiple sequencing technologies and bioinformatics methods, taking advantage of the strengths of each technology, ignoring technologies at genomic locations where they appear biased, and delineating benchmark regions to exclude regions that are biased in all current technologies. As sequencing technologies and variant-calling methods have improved, GIAB benchmarks have grown to include more challenging variants and regions of the genome, from 77% of the autosomal GRCh37 bases in 2014 (ref. 119) to 88% in 2016–2019 (ref. 115) and 94% in 2020–2021. The latter versions also added two son–father–mother trios of Ashkenazi Jewish and Han Chinese ancestry, as well as benchmarks on GRCh38.

Based on these benchmarks, PrecisionFDA and GIAB held two ‘Truth Challenges’ to inspire development of improved small-variant callers and provide a baseline for ongoing improvements. The first challenge was held in 2016 before releasing the first benchmark for the second GIAB sample, HG002 (ref. 114). When comparing each submission to the v3.2 HG002 benchmark released after the challenge, it demonstrated that a variety of short-read-based variant callers have accuracy >99.9% for SNVs and >99% for indels in the regions covered by the v3.2 GIAB benchmarks. However, many challenging variants and regions were excluded from this benchmark, and concordance between two of the highest performing variant callers outside the v3.2 benchmark regions was <80% for SNVs and indels¹¹⁴. With the advent of long reads with >99% accuracy, GIAB developed a new small-variant benchmark covering 76 million bp of segmental duplications and the highly polymorphic HLA gene region^{116,127}, which was used in the 2020 PrecisionFDA Truth Challenge V2 (ref. 1). Results from the Truth Challenge V2 demonstrated substantial improvements in sequencing technologies, variant-calling methods and benchmark sets. The most accurate submissions from the first challenge had substantially lower accuracy with respect to the new v4.2.1 benchmark, with SNV accuracy decreasing as much as 10-fold¹. Still, v4.2.1 excludes 8% of the sequence in GRCh38, in addition to the 7% of sequence missing from GRCh38, and variants are likely to be enriched in the remaining sequence. This result highlights the importance of understanding the limitations of any benchmark, particularly any challenging regions that are not included in the benchmark.

In complementary benchmark sets, the Illumina Platinum Genomes and Real Time Genomics benchmarks used a 17-member, 3-generation family pedigree to develop phased benchmark sets of variants for the mother and father^{120,128}. In this approach, the 11 grandchildren enabled robust phasing of the variants in their mother and father, and variants that were not inherited as expected according to this phasing were removed as potential errors. This fully phased set of variants has also been widely used to benchmark phasing methods.

The mother (NA12878) is the same as the pilot genome from GIAB, enabling cross-comparisons to validate and improve each benchmark. In general, these different approaches to forming benchmarks are highly concordant inside their benchmark regions, with ~4 differences per million matching variants after excluding differences near the edge of the Platinum Genomes benchmark regions¹¹⁵. An advantage of the Platinum Genomes benchmark is that it contained substantially more variants than the short read-based GIAB benchmarks available at the time. However, it had some limitations around complex variants, particularly in tandem repeats, owing to its fragmented benchmark regions¹¹⁵. In addition, because it relied on short reads, mapping errors caused some inaccurate variants in segmental duplications even though they were phased and inherited as expected¹²⁷.

The first benchmark to use long reads was based on assemblies of two effectively haploid human hydatidiform mole cell lines (CHM1 and CHM13)¹²³. These two assemblies were then aligned to the reference to call variants together as a synthetic diploid benchmark, covering some regions that were difficult to map with short reads. Because highly accurate long reads were not yet available, small indels could not be benchmarked owing to small errors in the assembly, and some segmental duplications could not yet be assembled. Nevertheless, this resource served as an important benchmark for more challenging variants, particularly before the recent advent of diploid assemblies based on highly accurate long reads, and it highlighted an important limitation of existing benchmarks excluding more difficult regions.

For SVs, benchmarks are less mature, but several resources have been developed. First, the Parliament method was developed to integrate SV calls from multiple technologies and variant callers¹²⁹. The HGSVC performed a focused analysis of a subset of the 1000 Genomes samples, developing an initial resource of phased SVs for 3 trios⁵³, and later a resource of SVs for 15 trios⁵⁵, both of which can be useful for testing sensitivity of variant callers. Another resource was developed for small variants and SVs using the unique shotgun Sanger sequencing data set available for HuRef^{130,131}. The GIAB Consortium developed a set of benchmark insertions and deletions larger than 50 bp along with benchmark regions that exclude complex SVs in HG002, making it the first SV benchmark to enable assessment of both sensitivity and precision¹¹⁷. Still, multiple challenges remain, especially for more complex SVs and when comparing SVs across many samples.

Recent efforts can identify SVs from thousands to millions of genomes with short or long reads. One challenge is to accurately identify SVs across many samples, particularly because the number of false calls can be amplified by the sample number. Some methods achieve this by jointly analysing all samples simultaneously per region^{132,133}, whereas other approaches revisit each SV per sample individually, reconciling different SV representations during the merging^{134–137}. Another approach focuses on building a database of discordant reads first to enable exact querying of SVs across the samples¹³⁸. For both benchmarking and population-scale analyses, substantial challenges remain for comparing different representations of complex SVs, especially when combining different sequencing technologies or SV-calling methods.

Many of these benchmarks did not include variants on the X and Y chromosomes in male samples because they are effectively haploid (except for the pseudoautosomal regions). Chromosome Y is especially challenging owing to its high fraction of satellite and segmental duplications, so GIAB and T2T currently have a focused effort to develop a benchmark from the first complete assemblies of these chromosomes in HG002, as well as from complete assemblies of the entire genome^{6,139}.

Somatic variant benchmarks

Somatic variants are those that arise after conception, and are often relevant in the context of understanding cancer. Benchmarks for tumour genomes are more challenging and currently have substantial limitations, but some initial benchmarks are now published. A DREAM Challenge was organized for somatic variant detection, using a tool that modifies real sequencing data to simulate somatic variants^{122,140}. Additional tools have been developed to simulate somatic variants of different types^{141–144}, which have been recently reviewed¹⁴⁵. Alternatively, DNA from normal and/or tumour cell lines like those from GIAB can be mixed to simulate somatic mutations, but some variant callers filter out the germline mutations¹²⁴. Synthetic DNA can also be added to normal cell lines to mimic a smaller number of mutations¹⁴⁶. Griffith et al. published analyses of deep sequencing data from an acute myeloid leukaemia (AML) patient sample; this information can be used to benchmark bioinformatics tools, although a cell line is not available¹⁴⁷. A unique benchmark data set used cell lineage information from cell sorting to develop a benchmark from a cell line that accumulates somatic mutations¹⁴⁸. Several efforts have developed multiple sequencing data sets and benchmarks for paired tumour and normal cell lines from the same individual, including COLO-829/COLO-829BL with 35,543 SNVs, 446 indels and 6,500 genes with copy number changes¹⁴⁹. The SEQC2 somatic working group recently published extensive interlaboratory sequencing data and benchmarks for another tumour–normal cell line pair with 37,398 SNVs and 1,754 indels assigned as high-confidence somatic mutations^{125,126}. Although existing studies provide important information, no current benchmark tumour–normal cell line pairs are explicitly consented for public release of genomic data, hence development of appropriately consented cell lines with corresponding public data sets is a crucial need for future somatic benchmarks (Box 2). Existing somatic benchmarks also generally include fewer challenging regions than germline benchmarks, so ongoing work is needed to benchmark somatic variants in repetitive regions.

Robust variant comparison tools for benchmarking

Although reliable benchmark sets have been important in advancing our understanding of the performance of variant callers, appropriate tools to compare variant calls from any method against the benchmark are essential for reliable use of these benchmarks. To develop best practices for benchmarking germline small variants, the GA4GH Benchmarking Team brought together benchmark set developers, benchmarking tool developers, and clinical and other bioinformatics users of benchmarking tools. This team standardized definitions for performance metrics, optimized benchmarking tools to account for different representations of variants, and developed methods to stratify performance by variant type and genome context. This team developed the hap.py framework with the vcfeval engine as best practice for comparing germline SNVs and small indels. This framework compares the query VCF (variants from the method being evaluated) to the benchmark VCF within the benchmark regions. We point the reader to the GA4GH best practices paper for details on benchmarking small variants¹¹⁴, as well as the recent precisionFDA Truth Challenge V2 for an example implementation of these best practices¹.

Several outstanding challenges remain for benchmarking complex variants, SVs, segmental duplications and satellites. No standards exist for representing many types of complex variants, so sophisticated benchmarking tools have been developed to reconcile differing representations of small-variant calls in a query and benchmark VCF, as long as the variants are called completely accurately in the region (Fig. 2b).

If any part of the variant is filtered, missed or incorrectly genotyped, particularly in homopolymers and tandem repeats, then other parts of the variant may be counted as errors if they are represented differently from the benchmark²¹.

SVs pose even more challenges for benchmarking due to imprecision in detection and lack of standard representations of the numerous types of complex SVs that occur. In addition, accuracy of SVs can be measured at different stringencies; for example, ‘least stringent’ would require that only SV type and rough location are correct (for example, SURVIVOR)¹⁵⁰, and ‘most stringent’ might require that the exact sequence change is correct and that it is annotated correctly (for example, the most stringent mode of Truvari)¹⁵¹. Isolated insertions and deletions in nonrepetitive sequences are generally the easiest to detect and to benchmark, but these make up only a small fraction of all SVs^{35,117}. Because the majority of SVs are located in tandem repeats, methods such as Truvari are designed to compare not just their reference location (start and end) but also their length (for example, the insertion could have a larger length than the reference location) and their sequence content¹⁵¹. Although these complex SVs can increasingly be resolved accurately by phased, long-read assemblies, even the most sophisticated SV benchmarking tools such as Truvari and hap-eval are just starting to be able to compare different representations if they are not represented as a single isolated insertion or deletion. For example, a recent GIAB benchmark for challenging medically relevant genes excluded genes such as *CR1* and *LPA* that were accurately resolved because the variation they contained was too complex for current benchmarking tools²¹. The kringle repeats in the gene *LPA* can either be represented as CNVs, or as one or more large insertions or deletions and small variants. In addition, SV callers may represent tandem duplications as insertions (SVTYPE = INS), as duplications (SVTYPE = DUP) or even as breakends (SVTYPE = BND) or translocations (SVTYPE = TRA) in VCF. The variant representation often depends both on the mapping methods and on the variant caller used. Other challenges in variant representation occur around gaps in the reference (Fig. 2d). Complex variants are increasingly characterized, even in large short-read studies³⁵, but robust benchmarking of these complex variants is an outstanding challenge that will need new standards for sequence alignment, variant representation, and comparison, possibly building on pangenome alignments.

Segmental duplications are increasingly being resolved accurately by phased, long-read assemblies, and these pose additional challenges in variant representation and benchmarking due to their structural and copy number variation (Fig. 2c). Current benchmark sets and benchmarking tools enable comparisons when the sample matches the reference in copy number, but new benchmarking tools and standards for representation need to be developed to benchmark the small variants and SVs that occur in additional copies of segmental duplications that are not in the reference. Graph-based and/or assembly-based representations of these regions may provide a path towards variant calling and benchmarking variants.

When benchmarking, variant call accuracy can differ by several orders of magnitude depending on the type of variant and genome context. GA4GH and GIAB provide methods for stratifying performance by variant type and in the different types of repetitive sequences that occur in the genome (Fig. 2). These tools also indicate variant types and regions with a high fraction of variants not assessed because they are not included in the benchmark regions. These variants outside the benchmark regions tend to be more challenging, so that variant error rates in the whole genome will be higher than those estimated from the benchmark¹¹⁴, and performance metrics can vary by an order of

magnitude between benchmarks that are more or less conservative^{1,123}. As benchmarks include increasingly challenging regions of the genome, stratifying performance by genome context and variant type becomes increasingly important for interpreting the results and powerful for understanding the strengths and weaknesses of any variant-calling method. Stratifying performance can also help to predict potential false positives or variants needing further confirmation in clinical samples. Measuring how false negatives are associated with genomic repeats can also help to predict clinically relevant variants that might be missed by a method⁷. In addition to improvements in accuracy, there have also been efforts to improve the speed of variant-calling methods (reviewed previously¹⁵²).

Visualizing and curating variants to understand errors

Visualizing and curating sequencing data from multiple technologies at and near challenging variants is often particularly valuable to understand differences in variant calls between methods or false positives

Box 2

Importance of broad consent for benchmark samples

Guidelines regarding informed consent for sharing human genomic data have evolved over time as potential risks are understood, and explicit consent for broad public sharing of genome data is particularly important for benchmark samples that will be widely used by the community. The family trios who provided samples characterized by the Genome in a Bottle Consortium (GIAB) are broadly consented under the Personal Genome Project for genome data sharing and commercial distribution of products based on their cell lines. This broad consent has enabled several applications, including adding spike-in DNA to test accuracy for particularly challenging variants and somatic variants, as well as mimicking clinical samples such as formalin-fixed paraffin-embedded (FFPE) and cell-free and circulating tumour DNA. Other benchmark samples, such as NA12878 (pilot genome of GIAB and sample from Platinum Genomes) were consented for public release of genome data but not for commercial redistribution. The 11 children of NA12878 used in the Platinum Genomes analysis¹²⁰ were not consented for public release of genomic data, so their data are in the restricted-access database of Genotypes and Phenotypes (dbGaP), making them less accessible as benchmark data. The CHM1 and CHM13 cell lines used in the synthetic diploid sample¹²³ are not in a public repository, so the existing public data can be used to benchmark bioinformatics pipelines, but limited access hinders experimental work to benchmark the sequencing method of a particular laboratory. Similarly, data from the COLO829/COLO829BL somatic variant benchmark study from Craig et al.¹⁴⁹ and the deeply sequenced Washington University acute myeloid leukaemia (AML) cohort¹⁴⁷ were deposited in the dbGaP restricted-access repository in accordance with the terms of the consent. No current benchmark tumour-normal cell lines are explicitly consented for public genome sequence release, so new cell lines are needed for broad use as genomic reference samples.

or false negatives relative to a benchmark. Sometimes, this curation helps to explain why a method has some types of errors, or it might show that the benchmark is incorrect or questionable. This visualization process is not currently available in a single application that provides the utilities described by Ben Shneiderman in his seminal work with the mantra “Overview first, zoom and filter, then details on demand”¹⁵³. Currently, a workflow requires the use of several different software tools, with visualization typically occurring primarily for ‘details on demand’. Specifically, `hap.py` and `vcfeval` provide an ‘overview first’ of variants identified as matching or not matching a benchmark (Fig. 5b). The [GA4GH/GIAB stratification](#) regions provide ‘zoom and filter’ to specific ‘clustering of errors’ correlated with genome context, such as different repetitive region types (Fig. 5d,e). Finally, a genomics viewer program such as Integrative Genomics Viewer (IGV) can provide ‘details on demand’ to inspect read alignments along with other sequencing or genomic features, including repeats that are not apparent from stratification, to determine sources of possible bias or error (Fig. 5d). Within the ‘details on demand’ step, the current visualization workflow from GIAB starts at the window size default for IGV of 40 bp around a variant, or sufficient to view an entire homopolymer or tandem repeat, if applicable. After identifying any local details, the user can zoom out to approximately 10 kb to inspect larger sequence contexts such as segmental duplications, nearby SVs or other notable features that might impact read mapping. Visualizing read alignments around a patient’s clinically important variants was also recommended in the Association for Molecular Pathology bioinformatics guidelines³⁷. Newer tools enable faster curation of SVs^{154,155}. We expect that visualization of the evidence for variants will remain particularly important for complex variants and challenging regions as methods access these areas of the genome^{18,156}. We show some examples of visualizing complex variants in Fig. 2, illustrating how ongoing work is needed to call and benchmark the most complex variants.

Conclusions and perspectives

The best methods now produce highly accurate variant calls for much of the human genome, but variant calling is far from being a fully solved problem. It will continue to be an area of active development as new sequencing technologies and analysis methods are enabling characterization of the most challenging variants and regions of the genome for the first time. The assembly by T2T Consortium of the first complete human genome opens the door to analysing human variation in extremely repetitive regions of the genome such as highly identical segmental duplications and satellite DNA in the centromeres. Pangenome alignments of assemblies from the HPRC highlight complex structural changes that are excluded from current benchmarks, including those in known medically relevant genes and gene families such as *CYP2D6*, *RHCE*, *DAZ*, *LPA* and *HLA*^{18,102}. To translate these developments to broad research and clinical applications, innovations are needed in many areas including sequencing technologies, assembly and variant-calling methods, variant representation, benchmarking and variant comparison tools, and expanded benchmark sets. As methods push into these challenging regions, it will be increasingly important to stratify performance metrics by genome context and variant type.

Performance metrics can differ by orders of magnitude depending on the composition of the benchmark set, genome context and variant type^{1,123}; advances in applications will depend on trustworthy benchmarking. Although methods can achieve >99% accuracy in existing benchmark regions, benchmarks exclude about 15% of the genome (when counting the 7% of sequence added in the newly completed human

genome sequence), and benchmarks probably exclude even more than 15% of all variants. Every benchmark set has limitations, and understanding these limitations is crucial. Variant benchmarks tend to lag behind the first methods to call a class of difficult genomic regions, until these methods are tested and used to expand the benchmark. Current benchmarks exclude many long homopolymers, tandem repeats, segmental duplications and satellite DNA, or some types of variants, such as large indels or SVs (for example, rearrangements). When challenging regions and variants are excluded, performance metrics for these cannot be estimated, and performance metrics are generally overestimated when looking at aggregate statistics. The most useful benchmarks are formed using high-coverage data from multiple technologies not all used by the method being tested. With the advent of complete genome assemblies from high-coverage long-read data, we are poised to be able to benchmark even the most challenging variants and regions. For example, current work in GIAB includes developing new benchmark sets and benchmarking tools for tandem repeats, as well as more comprehensive benchmarks from complete human genome assemblies with the T2T Consortium. However, even benchmarks formed from perfect assemblies will require new methods and standards for aligning assemblies to a reference and for comparing and representing complex variants.

Solutions for remaining challenges in variant analysis and representation could follow several different trajectories in this new age of complete human genome sequences. These options include continuing to align reads to a common reference such as T2T-CHM13 or GRCh38 and call variants, aligning phased assemblies to a common reference to call variants, aligning reads to a population-specific reference to call variants, aligning reads to a graph-based pangenome reference, progressively aligning reads to many reference genomes with ‘reference flow’, or aligning assemblies to a graph-based pangenome reference. We expect that each of these approaches and others yet to be developed will be active areas of research and that new benchmarking approaches will be needed to evaluate these methods, as they provide the greatest benefits for the most challenging variants and regions of the genome. Importantly, benchmarks representing challenging regions and variants across different ancestries will be needed to assess the strengths and weaknesses of each approach.

Pangenome tools are being built to translate variants between pangenesomes and linear reference genomes, which should help to enable research studies and clinical laboratories to use innovations in pangenome references, even if these pangenesomes continue to evolve. However, some of the most complex regions and variants in the large repeats discussed above are likely to only be represented in new formats such as a graph, and may even be best envisioned as something other than variants with respect to a reference genome. Although these complex variants affect a small fraction of the genome, they affect known medically relevant regions, and new important regions are likely to be discovered now that these regions can be accurately sequenced. Sequencing technologies are likely to continue to advance, including cheaper and more accurate short and long reads, along with analysis methods to characterize and benchmark increasingly difficult genomic regions and variants at scale. As new sequencing and analysis methods are developed, a positive feedback loop exists between technological innovation and benchmarking. Technological advances in sequencing and bioinformatics enable improved benchmark sets, and improved benchmarking (tools and sets) promote technology development and clinical translation – a cycle we expect to continue for years to come.

Published online: 14 April 2023

References

- Olson, N. D. et al. PrecisionFDA Truth Challenge V2: calling variants from short and long reads in difficult-to-map regions. *Cell Genom.* **2**, 100129 (2022).
- The latest iteration of the precisionFDA Truth Challenge, which serves as a baseline for variant call performance from short and long reads in easy versus more difficult regions using the GIAB v4.2.1 benchmark.**
- Pan, B. et al. Assessing reproducibility of inherited variants detected with short-read whole genome sequencing. *Genome Biol.* **23**, 2 (2022).
- Foxx, J. et al. Performance assessment of DNA sequencing platforms in the ABRF Next-Generation Sequencing Study. *Nat. Biotechnol.* **39**, 1129–1140 (2021).
- Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
- Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
- Initial demonstration of the value of accurate long reads for variant calling and assembly.**
- Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
- Dwarshuis, N. et al. StratMod: predicting sequencing and variant calling errors with interpretable machine learning. Preprint at bioRxiv <https://doi.org/10.1101/2023.01.20.524401> (2023).
- Meacham, F. et al. Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* **12**, 451 (2011).
- Nurk, S. et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
- Hannan, A. J. Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.* **19**, 286–298 (2018).
- Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
- Public resource of allele frequencies from 141,456 individuals using short reads, made available through the gnomAD genome browser.**
- Weisburd, B., VanNoy, G. & Watts, N. The addition of short tandem repeat calls to gnomAD. gnomAD <https://gnomad.broadinstitute.org/news/2022-01-the-addition-of-short-tandem-repeat-calls-to-gnomad/> (2022).
- Ren, J., Gu, B. & Chaisson, M. J. P. vnavos: VNTR annotation using efficient motif sets. Preprint at bioRxiv <https://doi.org/10.1101/2022.10.07.511371> (2022).
- Bakhtiari, M., Shleizer-Burko, S., Gymrek, M., Bansal, V. & Bafna, V. Targeted genotyping of variable number tandem repeats with adVNTR. *Genome Res.* **28**, 1709–1719 (2018).
- Vollger, M. R. et al. Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).
- Initial analysis of complex segmental duplication variation using the T2T-CHM13 reference.**
- Zhao, X. et al. Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *Am. J. Hum. Genet.* **108**, 919–928 (2021).
- Lincoln, S. E. et al. One in seven pathogenic variants can be challenging to detect by NGS: an analysis of 450,000 patients with implications for clinical sensitivity and genetic test implementation. *Genet. Med.* **23**, 1673–1680 (2021).
- Results from a large clinical laboratory showing that one in seven pathogenic variants are challenging for short reads owing to low mappability or variant type.**
- Chin, C.-S. et al. Multiscale analysis of pangenome enables improved representation of genomic diversity for repetitive and clinically relevant genes. Preprint at bioRxiv <https://doi.org/10.1101/2022.08.05.502980> (2022).
- Behera, S. et al. FixItFelix: improving genomic analysis by fixing reference errors. *Genome Biol.* **24**, 31 (2023).
- Aganezov, S. et al. A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022).
- Initial analysis showing that a complete human genome reference improves variant calling by fixing reference errors and adding new sequences.**
- Wagner, J. et al. Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat. Biotechnol.* **40**, 672–680 (2022).
- Latest benchmark from GIAB, demonstrating that diploid assembly can be used to form reliable small-variant and SV benchmarks for a set of 273 challenging medically relevant genes, and providing a prototype for future assembly-based benchmarks.**
- Vollger, M. R. et al. Increased mutation rate and interlocus gene conversion within human segmental duplications. Preprint at bioRxiv <https://doi.org/10.1101/2022.07.06.498021> (2022).
- Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
- Vollger, M. R. et al. Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019).
- Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
- Altemose, N. et al. Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).
- Mc Cartney, A. M. et al. Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nat. Methods* **19**, 687–695 (2022).
- Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
- Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* **19**, 329–346 (2018).
- Derrien, T. et al. Fast computation and applications of genome mappability. *PLoS ONE* **7**, e30377 (2012).
- Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2011).
- Ou, S. et al. Effect of sequence depth and length in long-read assembly of the maize inbred NC358. *Nat. Commun.* **11**, 2288 (2020).
- Ebbert, M. T. W. et al. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biol.* **20**, 97 (2019).
- Hardwick, S. A., Deveson, I. W. & Mercer, T. R. Reference standards for next-generation sequencing. *Nat. Rev. Genet.* **18**, 473–484 (2017).
- Byrska-Bishop, M. et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440.e19 (2022).
- Haldorsen, B. V. et al. The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
- Roy, S. et al. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J. Mol. Diagn.* **20**, 4–27 (2018).
- Arslan, S. et al. Sequencing by avidity enables high accuracy with low reagent consumption. Preprint at bioRxiv <https://doi.org/10.1101/2022.11.03.514117> (2022).
- Vergult, S. et al. Mate pair sequencing for the detection of chromosomal aberrations in patients with intellectual disability and congenital malformations. *Eur. J. Hum. Genet.* **22**, 652–659 (2014).
- Mahmoud, M. et al. Structural variant calling: the long and the short of it. *Genome Biol.* **20**, 246 (2019).
- Review of variant-calling methods for SVs, to complement our more general review of variant calling.**
- Marks, P. et al. Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res.* **29**, 635–645 (2019).
- Peters, B. A. et al. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487**, 190–195 (2012).
- Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res.* **27**, 757–767 (2017).
- Chen, Z. et al. Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information. *Genome Res.* **30**, 898–909 (2020).
- Belton, J.-M. et al. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).
- Garg, S. et al. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat. Biotechnol.* **39**, 309–312 (2021).
- Rhie, A. et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
- Sanders, A. D., Falconer, E., Hills, M., Spierings, D. C. J. & Lansdorp, P. M. Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nat. Protoc.* **12**, 1151–1176 (2017).
- Porubsky, D. et al. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat. Biotechnol.* **39**, 302–308 (2021).
- Rhoads, A. & Au, K. F. PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* **13**, 278–289 (2015).
- Eid, J. et al. Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
- Shafin, K. et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat. Biotechnol.* **38**, 1044–1053 (2020).
- Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
- Chaisson, M. J. P. et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
- Audano, P. A. et al. Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663–675.e19 (2019).
- Analysis using assemblies to show the prevalence of structural variation in the human genome.**
- Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 129 (2019).
- Xu, Z. et al. Fast-bonito: a faster deep learning based basecaller for nanopore sequencing. *Artif. Intell. Life Sci.* **1**, 100011 (2021).
- Shafin, K. et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat. Methods* **18**, 1322–1332 (2021).
- Recent iteration of the deep learning-based tool DeepVariant to call small variants from noisy long reads.**
- Payne, A., Holmes, N., Rakyan, V. & Loose, M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* **35**, 2193–2198 (2019).
- Logsdon, G. A. et al. The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).
- Cao, H. et al. Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *Gigascience* **3**, 34 (2014).
- Kaiser, M. D. et al. Automated structural variant verification in human genomes using single-molecule electronic DNA mapping. Preprint at bioRxiv <https://doi.org/10.1101/140699> (2017).
- Yuan, Y., Chung, C. Y.-L. & Chan, T.-F. Advances in optical mapping for genomic research. *Comput. Struct. Biotechnol. J.* **18**, 2051–2062 (2020).

64. Mantere, T. et al. Optical genome mapping enables constitutional chromosomal aberration detection. *Am. J. Hum. Genet.* **108**, 1409–1422 (2021).
65. Gerding, W. M. et al. Optical genome mapping reveals additional prognostic information compared to conventional cytogenetics in AML/MDS patients. *Int. J. Cancer* **150**, 1998–2011 (2022).
66. Coster, W. D., De Coster, W., Weissensteiner, M. H. & Sedlazeck, F. J. Towards population-scale long-read sequencing. *Nat. Rev. Genet.* **22**, 572–587 (2021).
- Recent review of how long-read sequencing is increasingly used to study variation in large numbers of samples.**
67. Poplin, R., Zook, J. M. & DePristo, M. Challenges of accuracy in germline clinical sequencing data. *JAMA* **326**, 268–269 (2021).
68. Cortés-Ciriano, I., Gulhan, D. C., Lee, J. J.-K., Melloni, G. E. M. & Park, P. J. Computational analysis of cancer genome sequencing data. *Nat. Rev. Genet.* **23**, 298–314 (2022).
- Recent review of somatic variant calling, to complement the focus on germline variants in this Review.**
69. Jain, C., Rhee, A., Hansen, N. F., Koren, S. & Phillippy, A. M. Long-read mapping to repetitive reference sequences using Winnnowmap2. *Nat. Methods* **19**, 705–710 (2022).
70. Jain, C. et al. Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36**, i111–i118 (2020).
71. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
72. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
73. Prodanov, T. & Bansal, V. Sensitive alignment using paralogous sequence variants improves long-read mapping and variant calling in segmental duplications. *Nucleic Acids Res.* **48**, e114 (2020).
74. Zheng, Z. et al. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nat. Comput. Sci.* **2**, 797–803 (2022).
75. AlDubayan, S. H. et al. Detection of pathogenic variants with germline genetic testing using deep learning vs standard methods in patients with prostate cancer and melanoma. *JAMA* **324**, 1957–1969 (2020).
76. Whalen, S., Schreiber, J., Noble, W. S. & Pollard, K. S. Navigating the pitfalls of applying machine learning in genomics. *Nat. Rev. Genet.* **23**, 169–181 (2022).
77. Sapoval, N. et al. Current progress and open challenges for applying deep learning across the biosciences. *Nat. Commun.* **13**, 1728 (2022).
78. Baid, G. et al. DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nat. Biotechnol.* **41**, 232–238 (2023).
79. Almogy, G. et al. Cost-efficient whole genome-sequencing using novel mostly natural sequencing-by-synthesis chemistry and open fluidics platform. Preprint at bioRxiv <https://doi.org/10.1101/2022.05.29.493900> (2022).
80. Sahraeian, S. M. E. et al. Deep convolutional neural networks for accurate somatic mutation detection. *Nat. Commun.* **10**, 1041 (2019).
81. Luo, R., Sedlazeck, F. J., Lam, T.-W. & Schatz, M. C. A multi-task convolutional deep neural network for variant calling in single molecule sequencing. *Nat. Commun.* **10**, 998 (2019).
82. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
83. Van der Auwera GA & O'Connor BD. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra* 1st edn (O'Reilly, 2020).
84. Cooke, D. P., Wedge, D. C. & Lunter, G. A unified haplotype-based method for accurate and comprehensive variant calling. *Nat. Biotechnol.* **39**, 885–892 (2021).
85. Freed, D. et al. DNAscope: high accuracy small variant calling using machine learning. Preprint at bioRxiv <https://doi.org/10.1101/2022.05.20.492556> (2022).
86. Avsec, Ž. et al. The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat. Biotechnol.* **37**, 592–600 (2019).
87. Dolzenko, E. et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* **35**, 4754–4756 (2019).
88. Willems, T. et al. Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods* **14**, 590–592 (2017).
89. Mousavi, N., Shleizer-Burko, S., Yanicky, R. & Gymrek, M. Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res.* **47**, e90 (2019).
90. Tang, H. et al. Profiling of short-tandem-repeat disease alleles in 12,632 human whole genomes. *Am. J. Hum. Genet.* **101**, 700–715 (2017).
91. Hall, C. L. et al. Accurate profiling of forensic autosomal STRs using the Oxford Nanopore Technologies MinION device. *Forensic Sci. Int. Genet.* **56**, 102629 (2022).
92. Fang, L. et al. DeepRepeat: direct quantification of short tandem repeats on signal data from nanopore sequencing. *Genome Biol.* **23**, 108 (2022).
93. PacificBiosciences. Tandem repeat genotyping and visualization from PacBio HiFi data. GitHub <https://github.com/PacificBiosciences/trgt> (2023).
94. Patterson, M. et al. Whatshap: weighted haplotype assembly for future-generation sequencing reads. *J. Comput. Biol.* **22**, 498–509 (2015).
95. Edge, P., Bafrna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* **27**, 801–812 (2017).
96. Garg, S. et al. A haplotype-aware de novo assembly of related individuals using pedigree sequence graph. *Bioinformatics* **36**, 2385–2392 (2020).
97. Chin, C.-S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
98. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
99. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
100. Chin, C.-S. & Khalak, A. Human genome assembly in 100 minutes. Preprint at bioRxiv <https://doi.org/10.1101/705616> (2019).
101. Jarvis, E. D. et al. Semi-automated assembly of high-quality diploid human reference genomes. *Nature* **611**, 519–531 (2022).
102. Liao, W.-W. et al. A draft human pangenome reference. Preprint at bioRxiv <https://doi.org/10.1101/2022.07.09.499321> (2022).
- First manuscript from the Human PanGenome Reference Consortium about their initial pangenome formed from accurate diploid assemblies, which can be used to improve variant calling.**
103. Kulski, J. K., Suzuki, S. & Shiina, T. Human leukocyte antigen super-locus: nexus of genomic supergenes, SNPs, indels, transcripts, and haplotypes. *Hum. Genome Var.* **9**, 49 (2022).
104. Sherman, R. M. & Salzberg, S. L. Pan-genomics in the human genome era. *Nat. Rev. Genet.* **21**, 243–254 (2020).
- Review of pangenomes, including how past work on pangenomes for other species can inform work on human pangenomes.**
105. Rakocewic, G. et al. Fast and accurate genomic analyses using genome graphs. *Nat. Genet.* **51**, 354–362 (2019).
106. Tetikol, H. S. et al. Pan-African genome demonstrates how population-specific genome graphs improve high-throughput sequencing data analysis. *Nat. Commun.* **13**, 4384 (2022).
107. Ebler, J. et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.* **54**, 518–525 (2022).
108. Sirén, J. et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374**, abg8871 (2021).
109. Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* **21**, 265 (2020).
110. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
111. Dewey, F. E. et al. Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genet.* **7**, e1002280 (2011).
112. Shumate, A. et al. Assembly and annotation of an Ashkenazi human reference genome. *Genome Biol.* **21**, 129 (2020).
113. Chen, N.-C., Solomon, B., Mun, T., Iyer, S. & Langmead, B. Reference flow: reducing reference bias using multiple population genomes. *Genome Biol.* **22**, 8 (2021).
114. Krusche, P. et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555–560 (2019).
- Primary product of the GA4GH Benchmarking Team, including a summary of best practices for benchmarking variant calls.**
115. Zook, J. M. et al. An open resource for accurately benchmarking small variant and reference calls. *Nat. Biotechnol.* **37**, 561–566 (2019).
116. Chin, C.-S. et al. A diploid assembly-based benchmark for variants in the major histocompatibility complex. *Nat. Commun.* **11**, 4794 (2020).
117. Zook, J. M. et al. A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38**, 1347–1355 (2020).
118. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).
119. Zook, J. M. et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).
120. Eberle, M. A. et al. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res.* **27**, 157–164 (2017).
121. Cleary, J. G. et al. Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. Preprint at bioRxiv <https://doi.org/10.1101/023754> (2015).
122. Ewing, A. D. et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat. Methods* **12**, 623–630 (2015).
123. Li, H. et al. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat. Methods* **15**, 595–597 (2018).
124. Jones, W. et al. A verified genomic reference sample for assessing performance of cancer panels detecting small variants of low allele frequency. *Genome Biol.* **22**, 111 (2021).
125. Zhao, Y. et al. Whole genome and exome sequencing reference datasets from a multi-center and cross-platform benchmark study. *Sci. Data* **8**, 296 (2021).
126. Fang, L. T. et al. Establishing community reference samples, data and call sets for benchmarking cancer mutation detection using whole-genome sequencing. *Nat. Biotechnol.* **39**, 1151–1160 (2021).
127. Wagner, J. et al. Benchmarking challenging small variants with linked and long reads. *Cell Genom.* **2**, 100128 (2022).
128. Cleary, J. G. et al. Joint variant and de novo mutation identification on pedigrees from high-throughput sequencing data. *J. Comput. Biol.* **21**, 405–419 (2014).
129. English, A. C. et al. Assessing structural variation in a personal genome — towards a human reference diploid genome. *BMC Genomics* **16**, 286 (2015).
130. Mu, J. C. et al. Leveraging long read sequencing from a single individual to provide a comprehensive resource for benchmarking variant calling methods. *Sci. Rep.* **5**, 14493 (2015).

131. Zhou, B. et al. Extensive and deep sequencing of the Venter/HuRef genome for developing and benchmarking genome analysis tools. *Sci. Data* **5**, 180261 (2018).
132. Jun, G. et al. muCNV: genotyping structural variants for population-level sequencing. *Bioinformatics* **37**, 2055–2057 (2021).
133. Collins, R. L. et al. A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
134. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
135. Chen, S. et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol.* **20**, 291 (2019).
136. Chiang, C. et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods* **12**, 966–968 (2015).
137. Kirsche, M. et al. Jasmine and Iris: population-scale structural variant comparison and analysis. *Nat. Methods* **20**, 408–417 (2023).
138. Chowdhury, M., Pedersen, B. S., Sedlazeck, F. J., Quinlan, A. R. & Layer, R. M. Searching thousands of genomes to classify somatic and novel structural variants using STIX. *Nat. Methods* **19**, 445–448 (2022).
139. Rhee, A. et al. The complete sequence of a human Y chromosome. Preprint at *bioRxiv* <https://doi.org/10.1101/2022.12.01.518724> (2022).
140. Lee, A. Y. et al. Combining accurate tumor genome simulation with crowdsourcing to benchmark somatic structural variant detection. *Genome Biol.* **19**, 188 (2018).
141. Samadian, S., Bruce, J. P. & Pugh, T. J. BamEngine: introduction of simulated allele-specific copy number variants into exome and targeted sequence data sets. *PLoS Comput. Biol.* **14**, e1006080 (2018).
142. Li, Z. et al. VarBen: generating in silico reference data sets for clinical next-generation sequencing bioinformatics pipeline evaluation. *J. Mol. Diagn.* **23**, 285–299 (2021).
143. Xia, L. C. et al. SVEngine: an efficient and versatile simulator of genome structural variations with features of cancer clonal evolution. *Gigascience* **7**, (2018).
144. Duncavage, E. J. et al. A model study of in silico proficiency testing for clinical next-generation sequencing. *Arch. Pathol. Lab. Med.* **140**, 1085–1091 (2016).
145. Duncavage, E. J. et al. Recommendations for the use of in silico approaches for next generation sequencing bioinformatic pipeline validation: a joint report of the Association for Molecular Pathology, Association for Pathology Informatics, and College of American Pathologists. *J. Mol. Diagn.* **25**, 3–16 (2023).
146. Reis, A. L. M. et al. Using synthetic chromosome controls to evaluate the sequencing of difficult regions within the human genome. *Genome Biol.* **23**, 19 (2022).
147. Griffith, M. et al. Optimizing cancer genome sequencing and analysis. *Cell Syst.* **1**, 210–223 (2015).
148. Shand, M. et al. A validated lineage-derived somatic truth data set enables benchmarking in cancer genome analysis. *Commun. Biol.* **3**, 744 (2020).
149. Craig, D. W. et al. A somatic reference standard for cancer genome sequencing. *Sci. Rep.* **6**, 24607 (2016).
150. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
151. English, A. C., Menon, V. K., Gibbs, R. A., Metcalf, G. A. & Sedlazeck, F. J. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol.* **23**, 271 (2022). **Describes the Truvari tool, which has been important for benchmarking SVs and tandem repeats by comparing different representations of variants.**
152. Alser, M. et al. From molecules to genomic variations: accelerating genome analysis via intelligent algorithms and architectures. *Comput. Struct. Biotechnol. J.* **20**, 4579–4599 (2022).
153. Shneiderman, B. in *The Craft of Information Visualization* (eds. Bederson, B. B. & Shneiderman, B.) 364–371 (Morgan Kaufmann, 2003).
154. Belyeu, J. R. et al. SV-plaudit: a cloud-based framework for manually curating thousands of structural variants. *Gigascience* **7**, giy064 (2018).
155. Chapman, L. M. et al. A crowdsourced set of curated structural variants for the human genome. *PLoS Comput. Biol.* **16**, e1007933 (2020).
156. Guaracino, A., Heumos, S., Nahnsen, S., Prins, P. & Garrison, E. ODGI: understanding pangenome graphs. *Bioinformatics* **38**, 3319–3326 (2022).
157. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.1303.3997> (2013).
158. Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
159. Eggertsson, H. P. et al. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* **10**, 5402 (2019).
160. Mitchell, M. et al. in *Proceedings of the Conference on Fairness, Accountability, and Transparency* 220–229 (Association for Computing Machinery, 2019).
161. Medvedev, P. The theoretical analysis of sequencing bioinformatics algorithms and beyond. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2205.01785> (2022).

Acknowledgements

The authors thank the members of the Genome in a Bottle Consortium, Human Pangenome Reference Consortium and Telomere to Telomere Consortium for helpful discussions about the strengths and limitations of the various technologies and bioinformatics methods. Certain commercial equipment, instruments or materials are identified to specify adequately experimental conditions or reported results. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the equipment, instruments or materials identified are necessarily the best available for the purpose.

Author contributions

All authors contributed to all aspects of the manuscript.

Competing interests

F.J.S. has received support from Oxford Nanopore Technologies, Pacific Biosciences, Illumina and Genentech. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41576-023-00590-0>.

Peer review information *Nature Reviews Genetics* thanks Kai Ye and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Related links

GA4GH/GIAB stratifications: <https://github.com/genome-in-a-bottle/genome-stratifications>

Genome in a Bottle Consortium: <http://www.genomeinabottle.org/>

gnomAD: <https://gnomad.broadinstitute.org/>

Human Pangenome Reference Consortium: <https://humanpangenome.org/>

T2T-CHM13: <https://github.com/marbl/CHM13>

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2023