Towards population-scale long-read sequencing

Wouter De Coster1,2,5, Matthias H. Weissensteiner3,5 and Fritz J. Sedlazeck4

Abstract | Long-read sequencing technologies have now reached a level of accuracy and yield that allows their application to variant detection at a scale of tens to thousands of samples. Concomitant with the development of new computational tools, the first population-scale studies involving long-read sequencing have emerged over the past 2 years and, given the continuous advancement of the field, many more are likely to follow. In this Review, we survey recent developments in population-scale long-read sequencing, highlight potential challenges of a scaled-up approach and provide guidance regarding experimental design. We provide an overview of current long-read sequencing platforms, variant calling methodologies and approaches for de novo assemblies and reference-based mapping approaches. Furthermore, we summarize strategies for variant validation, genotyping and predicting functional impact and emphasize challenges remaining in achieving long-read sequencing at a population scale.

These studies highlighted that a substantial proportion of hidden variation can be discovered with long-read sequencing. Indeed, recent long-read sequencing studies of Icelandic and Chinese populations have already identified previously undetected variants associated with height, cholesterol level and anaemia11,12. Analysis of 26 maize genomes13 revealed that more SVs are involved in causing diseases than in conferring agronomically important traits. In addition, long-read sequencing is beneficial for improving the continuity, accuracy and range of variant phasing14–16, assessing complex small variants17 and has been applied to find disease-associated alleles18–20. For de novo assemblies, multiple methods have been published over recent years to promote the use of long reads11–25.

Ongoing advances in sequencing technology and bioinformatics have paved the way to achieving long-read sequencing on a population scale26. The two main competitors driving innovation in the field are Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). PacBio high fidelity (HiFi) reads are generated by their Sequel II system; HiFi reads are both long (15–20 kbp) and highly accurate27. The ONT PromethION platform can produce much longer reads (up to 4 Mbp28), has a higher throughput at lower cost, but produces less accurate reads than the Sequel II system. Recent comparisons show an equivalent performance for SV calling with the two platforms29,30 (in-depth technical review and further comparison of long-read sequencing platforms available elsewhere31). Within the past 2 years, multiple studies have applied long-read sequencing to answer various questions in

1 department of biology, Penn State university, Pennsylvania, PA, USA.
2 Human Genome sequencing center, Baylor college of medicine, Houston, TX, USA.
3 These authors contributed equally: Wouter De Coster, Matthias H. Weissensteiner.
4 e-mail: fritz.sedlazeck@bcm.edu
5 These authors contributed equally: Wouter De Coster, Matthias H. Weissensteiner.
6 e-mail: fritz.sedlazeck@bcm.edu

https://doi.org/10.1038/s41576-021-00367-3
Structural variants
(SVs). Genomic alterations that are 50 bp or larger, including deletions, duplications, insertions, inversions and translocations.

Single-nucleotide variants (SNVs). Genomic alterations of 1–50 bp that are present at any frequency in the population. These variants include substitutions, insertions and deletions.

Short-read sequencing
Parallel sequencing of clonally amplified clusters of DNA molecules using optical or electrical methods, ranging from 25 bp to 400 bp per fragment.

Long-read sequencing
Continuous stretch of nucleotides derived from a sequencing machine, which usually exceed 1,000 bp and currently range up to 4 Mb.

Phasing
In this Review, only per sample (physical) phasing is considered, which refers to the detection of co-occurrences of two or more variants on the same DNA molecule by their overlap on the same read. In contrast to statistical inference phasing (using linkage information), this approach can include phasing of private or de novo variants.

PacBio high fidelity (PacBio HiFi). A type of PacBio sequencing that yields reads that are accurate (average 99.9%) and long (15–25 kbp). These reads are produced as a consensus from multiple serial observations of the same DNA molecule in a row. Previous versions of this method are referred to as ‘circular consensus sequencing’ (CCS).

ONT PromethION
A sequencing platform that yields longer (up to 4 Mb) but less accurate (average 5–8%) reads than the PacBio HiFi platform.

Mapping
The alignment of reads (sequences from shotgun sequencing) to a reference genome or de novo assembly.

Germline variants
Variants that are present in germline cells and therefore occur in every cell of an organism.

multiple different organisms32–35 (FIG. 1; TABLE 1). The largest human-focused long-read sequencing study to date investigated the genomic diversity of 3,622 Icelandic genomes11, with many other studies to follow, such as the NIH All of Us research programme and the NIH Center for Alzheimer’s and Related Dementias (CARD) in the USA and similar efforts in China, Abu Dhabi and Qatar. Long-read sequencing of a global diversity cohort is also being carried out as part of the Human Pangenome project12. Aside from human studies, long-read sequencing has been applied on a population scale to discover structural variation associated with phenotypes in crops32,33, fruitflies34 and songbirds35, and increasingly has a role in metagenomic studies (BOX 1). Here, we restrict our discussion to eukaryotic organisms, as long-read sequencing studies of bacteria and other prokaryotes require specific laboratory and bioinformatics approaches, and the challenges are inherently different.

In this Review, we discuss the approach of long-read, population-scale, whole-genome sequencing and highlight its advantages, point out challenges and provide an overview of different experimental setups. We define population-scale sequencing here as sequencing of more than five genomes, although in the case of more limited genomic diversity in some organisms, a lower number of individual genomes may be sufficient. We focus on technologies that produce continuous sequence reads and do not address other long-range technologies, such as linked reads or optical mapping (for example, Bionano Genomics). However, both these technologies may be useful and applicable in a population setting36,37. When sequencing of the highest number of samples is required, targeted sequencing may be a cost-efficient alternative to whole-genome approaches (BOX 2). Similarly to most population-scale sequencing projects, we focus on germline variants, as somatic variants require higher genome coverage and access to the relevant tissues.

Project strategies
The total number of sequenced individuals (or rather chromosomes) should in general be as high as possible. However, the different underlying questions that motivate population-scale sequencing studies have vastly different sample size requirements. Although estimating the degree of genetic differentiation or ancestral population size is already possible with a sample size as low as ten chromosomes (five individuals of a diploid organism)38, the identification of rare variants (and potentially associated diseases) in a population usually requires sample sizes that are many orders of magnitude higher39. Regardless of the approach taken, it is crucial to keep track of metadata and control for covariates in the cohort selection.

There are multiple commonly applied strategies with specific budget requirements to be considered at the beginning of a large population-scale sequencing project (FIG. 2a). Here, we discuss three main strategies that

---

**Fig. 1 | Overview of population-scale studies using long-read sequencing.** Studies published in 2019–2021 in which five or more samples were sequenced are included. Genome size of study organisms is viewed in three different categories (<500 Mbp, 500–2,000 Mbp and >2,000 Mbp), and the methodological approach taken to investigate genetic variation (comparison of assemblies, read mapping against a reference or both) is illustrated by the different colours. For further details, see TABLE 1.
Table 1 | An overview of long-read-based population studies

| Study                        | Organism and category | Technology* and analysis approach       | Sample size* | Genome size (Mbp) | Ref. |
|------------------------------|-----------------------|----------------------------------------|--------------|-------------------|------|
| Kou et al. (2020)            | Rice                  | PacBio Assembly comparison and read mapping | 15 (LR); 393 (SR) | 430               | 129  |
| Weissensteiner et al. (2020) | Crow                  | PacBio Read mapping                    | 33 (LR); 127 (SR) | 1,300             | 35   |
| Chakraborty et al. (2019)    | Drosophila            | PacBio Assembly comparison             | 14 (LR)      | 180               | 34   |
| Jiao & Schneeberger (2020)   | Arabidopsis           | PacBio Assembly comparison             | 7 (LR)       | 135               | 330  |
| Alonge et al. (2020)         | Tomato                | ONT Read mapping                       | 100 (LR)     | 950               | 32   |
| Beyter et al. (2020)         | Human                 | ONT Read mapping                       | 3622 (LR)    | 3,200             | 31   |
| Tusso et al. (2019)          | Yeast                 | ONT and PacBio Assembly comparison and read mapping | 17 (LR); 161 (SR) | 12               | 10   |
| Liu et al. (2020)            | Soy bean              | PacBio Assembly comparison             | 26 (LR)      | 1,150             | 33   |
| Chawla et al. (2020)         | Rapeseed              | ONT and PacBio Read mapping            | 12 (LR)      | 1,132             | 331  |
| Hiatt et al. (2020)          | Human                 | PacBio Assembly comparison             | 18 (LR)      | 3,200             | 332  |
| Mitsuhashi et al. (2020)     | Human                 | ONT and PacBio Read mapping            | 37 (LR)      | 3,200             | 25   |
| Shafin et al. (2020)         | Human                 | ONT Assembly comparison                | 11 (LR)      | 3,200             | 333  |
| De Roeck et al. (2020)       | Human                 | ONT Read mapping                       | 11 (LR)      | 3,200             | 334  |
| Chaisson et al. (2019)       | Human                 | ONT and PacBio Assembly comparison     | 9 (LR)       | 3,200             | 335  |
| Morena-Barrio et al. (2020)  | Human                 | ONT Read mapping                       | 19 (LR)      | 3,200             | 336  |
| Song et al. (2020)           | Rapeseed              | PacBio Assembly comparison             | 8 (LR)       | 1,132             | 337  |
| Sone et al. (2019)           | Human                 | ONT and PacBio Read mapping            | 17 (LR)      | 3,200             | 338  |
| Kim et al. (2020)            | Drosophila            | ONT Assembly comparison                | 101 (LR)     | 180               | 339  |
| Pauper et al. (2020)         | Human                 | PacBio Read mapping                    | 15 (LR)      | 3,200             | 340  |
| Ebert et al. (2020)          | Human                 | PacBio Assembly comparison             | 64 (LR)      | 3,200             | 46   |
| Quan et al. (2020)           | Human                 | ONT Read mapping                       | 25 (LR)      | 3,200             | 341  |
| Hufford et al. (2021)        | Maize                 | PacBio Assembly comparison             | 26 (LR)      | 2,200             | 342  |
| Hu et al. (2021)             | Maize                 | PacBio Assembly comparison             | 6 (LR)       | 2,200             | 343  |
| Wu et al. (2021)             | Human                 | ONT and PacBio Read mapping            | 405 (LR)     | 3,200             | 344  |

*Two main platforms are used in long-read sequencing projects, Pacific Biosciences (PacBio) high fidelity (HiFi) and Oxford Nanopore Technologies (ONT) PromethION. Sample sizes for long-read (LR) and short-read (SR) sequencing are specified.
Box 1 | Long-read metagenomics

Metagenomic studies do not address populations in a traditional sense, yet they nevertheless assess genetic information stemming from separate (organismal) entities and chromosomes. Long-read sequencing is seemingly ideal to study prokaryotic organisms and viruses contained in metagenomic (for example, stool, gut and environmental) samples, since their genomes are usually much smaller than the currently achievable average read length in these technologies\(^1\). However, for metagenomics, factors such as the generally higher amount of required input DNA, high sequence similarity between taxonomic units and higher cost per base pair have thus far hampered the widespread application of long-read sequencing.

Recent improvements in high molecular weight (HMW) DNA extraction specific to metagenomic samples seem to hold the potential to facilitate a more widespread application of long-read sequencing in metagenomics. For example, a workflow to obtain improved yields of HMW DNA from human stool samples and furthermore provide a bioinformatic workflow incorporates base-calling, assembly, error correction and genome circularization with ONT reads\(^2\). Other efforts have been directed at improving the assembly step. metaFlye\(^3\) is the first metagenomics-specific genome assembler, dealing with highly uneven coverage as well as sequence similarity between closely related genomes typical of metagenomic samples, and it seems to greatly enhance the ability to generate bacterial genomes in single contigs. Furthermore, others have sequenced the 16S rRNA gene as a species identifier, benefitting from the closely related genomes typical of metagenomic samples, and it seems to greatly enhance the ability to generate bacterial genomes in single contigs. Furthermore, others have sequenced the 16S rRNA gene as a species identifier, benefitting from the longer read length to improve the classification\(^1\).

To improve cost efficiency, a hybrid approach using both short and long reads seems to be a valid approach for assessing metagenomic samples. Overholt et al.\(^4\) have demonstrated that by combining Illumina and ONT reads, twice and four times more high-quality assemblies were recovered from a water column sample than by using each technology alone, respectively. Although these hybrid approaches will continue to be used, long-read-only approaches are likely to succeed in the long run\(^5\).

allow for different scaling and budgeting and thus have an impact on the level of resolution in detecting genetic variation. Across virtually all sequencing technologies, the cost per sequenced base pair is consistently decreasing. To be able to compare the strategies discussed below, we use the required long-read sequencing output as a proxy for costs (Supplementary table 1). Although we assume a diploid genome with a size similar to the haploid human genome (3.2 Gbp), we note that for genomes with higher ploidy (for example, hexaploid plants), the overall coverage must be adapted to the ploidy of the organism (that is, the number of homologous chromosomes). Furthermore, we assume a sample size of ~2,500 individuals, similar to that of the 1000 Genomes project\(^6\). At the time of writing (early 2021), the least expensive option to generate long-read data is the ONT PromethION platform, with a yield of roughly 100–150 Gbp per flow cell at a price between US$650 and US$2,100, depending on the discount obtained when multiple flow cells are purchased simultaneously. Of note, PacBio HiFi reads are of adequate length and high accuracy, and although not formally assessed, it is reasonable to expect that lower coverage would be sufficient with this technology. However, at the time of writing (early 2021) this still equates to a higher cost than with the ONT PromethION platform, as one PacBio single-molecule real-time (SMRT) cell costs ~US$1,300 and yields ~500 Gbp (continuous long reads) or ~30 Gbp (HiFi) of data.

A full coverage approach. Although the most expensive of the three approaches, the highest level of resolution is obtained with a strategy that aims to sequence every sample of the population with medium to high coverage (a ‘full coverage’ approach; FIG. 2a). The main criterion for deciding on the coverage required per sample is whether a de novo assembly (>40-fold coverage required) or reference-based alignment approach (>12-fold coverage required\(^7\)) is planned. The advantage of this strategy is its comprehensiveness, the simplicity of the study design and the relatively straightforward computational workflow. Furthermore, samples receive similar coverage and are therefore equally well studied, and rare variations in each sample can be easily detected. Sequencing all 2,500 individuals at 20-fold coverage requires 150 Tbp of sequencing data.

A mixed coverage approach. In the ‘mixed coverage’ approach (FIG. 2a), a subset of samples that are representative of the subgroups in the cohort (for example, ethnicities or subpopulations) are sequenced at high coverage (for example, 30-fold) and the remaining samples at low coverage (for example, >5-fold). Although this approach is generally less expensive than the full coverage approach, it still achieves high overall detection sensitivity and is thus particularly suitable for studies with a high number of individuals or a limited budget. However, several analytical challenges remain, especially in achieving high accuracy of genotypes across multiple samples or differentiating somatic from heterozygous germline variants, which is further complicated by regions exhibiting recurrent mutations. In addition, there will certainly be a bias towards common alleles with this mixed coverage approach, as many rare alleles can be missed, especially if a locus is heterozygous and the alternative allele is thus sparsely covered. Assuming that in this second strategy 200 individuals are sequenced at 30-fold coverage and the remainder of the cohort at 8-fold coverage, this approach requires 73 Tbp of data and is thus potentially half as expensive as the full coverage strategy.

A mixed sequencing approach. The ‘mixed sequencing’ approach (FIG. 2a) involves long-read sequencing of just a few samples (for example, 10–20% of all samples) and short-read sequencing of the remaining samples to genotype variants that are discovered by long-read sequencing. The rationale behind this approach, similar to the selection of individuals for high coverage in the mixed coverage strategy, is to identify a small subset of samples (either randomly or by known diversity\(^8\), ethnicity or phenotype) and sequence only these to higher coverage. This mixed sequencing approach was effective in elucidating germline SVs that predispose to cancer, whereby short-read sequencing was used to identify evidence of SVs followed by long-read sequencing of selected samples\(^9\). Phylogenetic analysis of variants detected by short-read sequencing has also been used to select a representative set of soybean accessions for long-read sequencing and de novo assembly\(^10\). Other studies have used SVCollector\(^11\) to automatically select samples (this is done over iterations by selecting the most diverged sample and re-ranking remaining samples based on non-selected variation) for long-read sequencing to complement existing short-read sequencing data\(^12\). Once a subset of samples have been sequenced with long-read technologies, yielding a set of identified SVs, their

Somatic variants

Variants that can occur in any tissue cells but not in the germline cells. They often vary in frequency because they usually occur in only a subset of cells.

De novo assembly

A method for constructing genomes from a large number of short-read or long-read DNA fragments, with no a priori knowledge of the correct sequence or order of the fragments.
breakpoint coordinates can be genotyped (for example, insertions) across the short-read sequence data sets. In this way, robust allele frequencies for the identified variants can be obtained, albeit with a bias towards variants identified by long-read sequencing, which means that rare variants contained in other samples may be missed. It may not be possible to directly genotype all types of SV using short reads, especially in repetitive regions, but knowledge of the haplotypes on which the SVs of interest are found will enable imputation of these variants based on short-read SNV genotypes. This strategy has already been applied using diversity panels of human SVs to discover novel expression quantitative trait loci (eQTLs) and signatures of evolutionary adaptation. If for this strategy no additional short-read data need to be generated, then this approach is likely to be the most affordable, as sequencing 200 of the 2,500 individuals to 30-fold coverage only requires 18 Tbp of data.

### Sequencing logistics

Efficiently operating long-read sequencers at scale, from logistics to sample preparation, loading optimizations and run monitoring, is not a trivial task. ONT and PacBio have different advantages but also challenges in almost every step in this process given their different designs of flow cells and sequencing instruments (FIG. 2b). The per-sample sequencing process and the characteristics of each technology are reviewed elsewhere.

A substantial amount of high molecular weight DNA (HMW DNA) and highly pure input DNA is of crucial importance in these methods. Achieving this DNA quality requires specific extraction methods and is often challenging for samples for which only limited or degraded material is available (for example, non-contemporary samples or samples from very small organisms). Amplification-free low-input DNA kits exist for both PacBio and ONT (https://nanoporetech.com/products/kits) sequencing platforms, with a minimum input DNA amount of 150 ng and 400 ng, respectively. However, these machines frequently require much more DNA to produce optimal sequencing yields. At the time of writing, it is often necessary to perform a nuclease flush and library reloading on an ONT flow cell to recover blocked pores to obtain the highest yield, which is an additional preparation step that is not necessary for PacBio cells. Importantly, ONT flow cells and PacBio SMRT cells have a limited shelf life, which is logistically challenging when sequencing many samples. Depending on the organism and its features, such as its physical size, the presence of a cell wall or secondary metabolites, high-quality DNA extraction can be a major constraint. Variability in DNA quality and molecular weight is a common issue and pre-sequencing quality control is necessary to ensure that inadequate samples are omitted and other technical covariates are recorded to be taken into account in downstream statistical analysis.

ONT sequencers store the raw data as hdf5 files (in the fast5 format), requiring base calling to obtain the more commonly used and much smaller fastq and BAM formats. Currently, incremental updates to the ONT base-calling algorithm regularly improve the read accuracy, which suggests that repeating the base calling of older data is valuable. This reanalysis requires long-term storage of the fast5 files, which can be up to 1.5 TB for a single PromethION flow cell, although further compression is possible. By contrast, the PacBio base-calling process is highly mature, and BAM files containing unaligned reads are produced directly from the sequencing machine. For HiFi reads, post-processing of the subreads is essential to collapse consecutive sequenced DNA molecules down to a high-quality consensus sequence, which is also done on the latest version of the machine (Sequel IIe system), and thus the overall data storage requirement is much reduced.

### Analytical considerations

Arguably the main challenge in population-level studies is a scalable and streamlined analysis. Multiple recent reviews have discussed approaches at the single sample level. Table 3 lists computational tools that are commonly used in long-read sequencing projects and these are reviewed in-depth elsewhere. Of note, in this very rapidly developing area of genomics, new tools are introduced constantly while established ones quickly become outdated. As we do not assume that matching

---

**Box 2 | Targeted sequencing**

Sample numbers can be scaled up at a lower cost using target enrichment approaches. Several methods have been introduced to enrich for a particular region of a genome, ranging from traditional capture and PCR amplicons to using the Cas9 system and an in silico sequencer-based selection (for example, Uncalled or Readfish). These approaches typically can target 10–20 kb regions, although sequencer-based selection methods potentially enable larger targets to be sequenced. The Cas9 system can enrich a region without amplification and thus also enables the assessment of methylation patterns and sequences that are hard to target, such as repeats. All these laboratory enrichment methods work for both long-read sequencing platforms, namely Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). However, the in silico enrichment is unique to the ONT platform and is of interest for many future applications, as it does not require laboratory enrichment. Both Uncalled and Readfish sequence the first ~1 kbp of every read and if this read does not overlap with a targeted region, the DNA molecule is ejected and the next molecule is read. However, if the read matches the sequence of the targeted region, sequencing continues, resulting in a modest on-target enrichment.

Multiple projects that use this more cost-efficient methodology to study specific diseases with known gene targets have been published. The analysis of these data sets is often very similar to full genome analysis, but is computationally less demanding. The coverage per target typically exceeds that of whole-genome approaches, achieving hundreds of fold coverage for the targeted regions. Furthermore, off-target reads (sequences that have not been fully depleted) must be taken into account and filtered out so that they do not affect the analysis. Depending on the type of targeted sequence (for example, amplicon versus the Cas9 approach), these off-target reads can occur more frequently than others owing to the different efficiencies in off-target depletion. For example, a Cas9 system often has off-target effects as well as sequencer-based targeting of regions (~30% enrichment on target). By counting the reads within and outside the targeted region, it is possible to assess the efficiency of the chosen method.

Another very common application of these targeted sequencing approaches that has recently become very important is enriching for a specific pathogen or virus, such as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus responsible for the coronavirus disease 2019 (COVID–19) global pandemic. The most commonly applied protocol in this context is ARTIC, which aims to amplify ~200 bp RNA segments of the virus. In addition, loop-mediated isothermal amplification (LAMP) and/or capture methods have been very effective in studying the diversity of SARS-CoV-2 isolates. Another interesting development from ONT is a targeted approach to detect the presence of SARS-CoV-2 using the LAMP-based assay LamPORE. LamPORE targets three regions of the viral genome (ORF1a and the E and N genes) and a control (human actin), which allows testing of ~96 patients in a single MinION run in ~1 h (REF.166).
Segmental duplications
DNA sequences (typically >1 kbp in length) that are highly identical (90–100%) in sequence content and exist in multiple locations in a genome. They can also be considered a special form of duplication.

short-read sequencing data are available for every individual, the integration of long-read and short-read data is not discussed. Nevertheless, we highlight the important role of short reads for the polishing of long reads and assemblies or in fine-scale resolution of SV breakpoints. These applications may lose their relevance as the accuracy of long-read sequencing improves, as is already the case for PacBio HiFi data.

For population-scale projects, the choice of analytical tools often involves balancing sensitivity and computational efficiency. Before downstream analysis, it is crucial to perform quality control of experimental factors that directly affect the performance of assembly, SV detection and read phasing, such as DNA fragment length and sequencing yield. Multiple tools have been developed for this purpose. Changes in sequencing chemistry or technical equipment during the project may lead to artefacts in the analysis and can thus potentially affect the findings. As such, it is important to randomly assign samples to batches, for example, sequencing runs, to reduce technical covariates.

Two main strategies for downstream analysis are available: aligning reads from individual samples to a single reference genome or comparing de novo assemblies (Fig. 2c). These two approaches are very different in their computational and coverage requirements, which in turn depend to a large extent on genome size and complexity. For both approaches, the goal is to apply the same set and versions of methods to all samples. The results need to be generated in a consistent way using correct version control and reproducible pipelines to avoid additional artefacts in the analysis. In the following sections, we discuss alignment-based and de novo assembly approaches and graph genome-based methods.

Read alignment-based analysis. Alignment-based approaches are often the method of choice for population-scale studies, as they facilitate the comparison of all samples against a common coordinate system (that is, the reference genome), which is illustrated by the fact that more than half of population studies (Table 1) employ these approaches. Furthermore, these approaches are often less computationally demanding and require substantially less coverage than assembly-based methods. Alignment-based approaches rely on matching sequencing reads with a reference genome, the overall correctness of which will affect the analysis of read data. If the reference genome is incomplete, incorrect, fragmented or too divergent from the focal sample, it will lead to biases in the downstream analysis.

The software for long-read sequence data analysis is under constant development, and alignment methods in particular have become much faster in recent years (Table 2). The NGMLR and LAST methods speed up the alignment process and improve the accuracy of long-read alignment. The minimap2 aligner is considerably faster than its competitors while often delivering similar results, and thus it is currently the most popular, widely accepted long-read aligner.

Two noteworthy recent innovations are Winnowmap, which improves alignments (specifically in repetitive regions), and Ira, which improves the alignment in the presence of SVs.

The choice of tools for the detection of genetic variation is arguably of equal importance. For SVs, several tools are currently available, such as Sniffles, SVIM, PBHoney, CuteSV and pbsv (Table 2). One of the remaining challenges is the accurate representation of SV breakpoints, which is particularly difficult in the context of more complex events involving multiple variants in repetitive regions, such as segmental duplications or large tandem repeat arrays (SV detection methods are comprehensively reviewed elsewhere). Recently developed tools are removing the need for high sequencing coverage by enabling SV calling and genotyping at lower coverage, although the associated risk of incomplete or erroneous SV detection and genotyping cannot be ignored.

Owing to the different error profiles of long reads, naive pile-up approaches or SVN and small insertion-deletion (indel) calling methods that were developed for short-read sequencing are usually inadequate or suboptimal for long reads. Over the past few years, multiple strategies have been developed to improve the detection of small variants with sophisticated machine learning models for each of the long-read sequencing technologies (Table 2). Current methods include, for example, DeepVariant, Pepper, Clair (both using deep learning) and LongShot (which explicitly requires alleles to be concordant with the haplotype structure), which also outperforms Illumina-based SNV calling. PacBio HiFi, in contrast to ONT, is also competitive with Illumina for small indels.

Expansions and contractions of tandem repeat arrays are a highly challenging and frequent type of variation. As these repetitive DNAs, which include short tandem repeats (1–6 bp repeat unit) and minisatellites (>6 bp repeat unit), are known to contain disease-causing alleles, accurate characterization of them is crucial. Some tools have been developed specifically for this purpose, such as tandem-genotypes and TRiCoLOR. Similar challenges remain for accurate characterization of other repeats. For example, the LPA locus (encoding apolipoprotein(a)) consists of 8 kbp tandem repeat units (encoding kringle IV domains) that are repeated 5–10 times in human genomes, making it notoriously difficult to assess.

To date, most reference genomes consist of a haplotype-collapsed representation, in which two or more chromosomal haplotypes are collapsed during assembly to a single artificial consensus sequence. Phased genome assemblies, in which the haplotype structure of each chromosome is fully resolved, have the potential to more accurately represent the genome. The human Telomere-to-Telomere (T2T) consortium effort aims to produce the first full chromosome assembly of the human genome from the essentially haploid complete hydatidiform mole (CHM13) genome and has already completed assembly of chromosome 8 (Ref. 79) and chromosome X (Ref. 79). In another example, a single haplotype from a haplotype-resolved de novo assembly was used as the reference for read alignment in a population genetic study in crows.
Population-scale de novo assemblies. Many reference genomes based on short-read sequencing are incomplete or highly fragmented with many gaps. Furthermore, hundreds of megabases of population- and individual-specific sequences are absent from the human reference genome. These missing sequences are often repetitive, but also include coding sequences. As a consequence, a fraction of reads derived from a sample cannot be aligned to the reference genome or they align to paralogous sequences, leading to tens of thousands of false-positive and false-negative variants for each individual. Therefore, creating and comparing de novo assemblies is desirable. The increased availability and affordability of long-read sequencing data have led to an explosion of faster and more accurate genome assembly tools.
which haplotype-resolved de novo assembly is commonly considered the most comprehensive representation of a genome. This competition to produce improved de novo assembly methods has led to the rapid development of new tools, usually focusing on either computational demand, contiguity, completeness or correctness, indicating that genome assembly represents (at present) a trade-off between these key parameters. De novo assembly-based approaches are often more sensitive and better for reconstructing highly diverse regions of the genome than alignment-based approaches, but can also lead to a collapse of highly similar segmental duplications\(^5\). For such duplicated regions, specific algorithms have been developed that leverage SNVs that differentiate multiple copies of repeats and thereby can recover medically relevant duplicated genes\(^8\),\(^9\). The dependence of de novo assembly on high read coverage and more computationally demanding methods has made it historically very challenging for large population-scale sequencing. However, the ever-increasing yield of sequencing technologies will enable the sequencing of each sample to sufficient coverage to obtain a high-quality de novo assembly\(^9\) (FIG. 1; TABLE 1).

Single-genome projects iteratively test multiple parameters or different methods to optimize a de novo assembly, which is neither realistic nor desirable in a population context. Multiple projects have integrated proximity-ligation or strand-specific short-read sequencing methods for substantial improvements of the contiguity of the assemblies\(^5\),\(^6\),\(^7\), but such approaches do not scale well to large populations. De novo assembly-based approaches are typically also more computationally demanding, which becomes especially relevant for large numbers of samples. Large cloud storage infrastructures might improve the scalability, but the computing cost will rise substantially. The recent development of less computationally demanding assemblers may be able to mitigate this limitation\(^7\).

Another important consideration is the scalability of the downstream computational approaches. Although the process of genome assembly already requires considerable computational resources, these demands increase linearly with the addition of more individuals. To infer genomic variation, de novo assemblies are usually compared with a chosen reference genome, yielding a standard variant call format (VCF) file. Currently, genomic alignment tools and dedicated variant callers (such as MUMmer\(^8\), Assemblytics\(^8\), minimap2 or dipcall\(^9\) and SVIM-asm\(^8\)) are designed to provide a pairwise comparison of two genomes, such as the assembled and a reference genome (TABLE 2). However, in a project with multiple (diploid) genomes, this is clearly not ideal, as a whole-genome alignment-based approach likely suffers from the same biases as a read alignment-based approach.

For example, in the case of novel sequence insertions in samples compared with a single reference genome, these variants will often be more challenging to compare across all samples of the population (FIG. 5a). This issue might be further amplified by gaps in the reference assembly, which potentially reduces the number of regions that can be compared across the population. Although troublesome for comparisons across samples, assembly-based SV calling will more likely correctly represent complex SVs that are longer than the read length and therefore harder to correctly identify with alignment-based methods (FIG. 3b). The likely most comprehensive option would be a compare-all-with-all approach (FIG. 5b), in which unique pairwise comparisons increase quadratically, meaning that with 100 samples there are already 4,950 possible ways to compare samples with each other. Clearly, such an approach would currently not be feasible for most projects, and alternative strategies have to be developed. Most recently, the introduction of progressive Cactus\(^9\), a tool that constructs an ancestral genome when comparing two assemblies based on a guide tree, has enabled comparison across multiple genomes. However, to date this tool has mainly been tested across species and not between individuals of a species.

Another, perhaps even greater, challenge in de novo assembly approaches is the correct representation of ploidy. Many organisms have diploid genomes (for example, humans and many animals) and even higher ploidies exist, such as in some crops. Tools optimized for diploid (that is, haplotype-aware) de novo assembly are available to reconstruct both haplotypes\(^7\). This reconstruction is essential to recover all heterozygous variation, as two different haplotypes may otherwise be collapsed to a single artificial and incorrect representation of the chromosome. However, haplotype-resolved de novo assemblies often require higher coverage and computational cost. The correct genotyping of both heterozygous and homozygous variants is of utmost importance for subsequent population genetic analysis. A recent solution is to first create an unphased assembly, then identify variants and partition reads into haplotypes before creating phased contigs\(^8\),\(^9\).

Even if complete and accurate haplotype-resolved assembly is achieved, then SV calling from assembly-to-assembly comparison might not be straightforward in highly complex regions. For example, the human LPA\(^7\) and SMN1 and SMN2 (REF\(^9\)) loci with their highly repetitive structure lead to problems in genomic alignments. As such, the main challenge may shift to genomic alignments and methods to interpret the detected differences between multiple assemblies.
Table 2 | An overview of software tools for analysing long-read sequencing data

| Category                  | Tool name            | Description                                                                 | Ref. |
|---------------------------|----------------------|------------------------------------------------------------------------------|------|
| De novo assembly          | (Hi)Canu             | Versatile de novo assembler                                                  | 23   |
|                           | Flye                 | Fast de novo assembler that can also operate on low coverage data            | 24   |
|                           | Shasta               | Fast ONT assembler                                                           | 25   |
|                           | Falcon Unzip         | PacBio assembler for phased assemblies                                       | 22   |
|                           | Peregrine            | Optimized assembler for HiFi data only                                       | 128  |
|                           | hiiasm               | Optimized assembler for HiFi data only                                       | 139  |
|                           | PGAS                 | Phased assembly including strand seq                                        | 46   |
| Genomic alignment         | LAST                 | Versatile method to align contigs or genomes                                 | 57   |
|                           | MUMmer               | Long-standing genomic aligner                                                | 87   |
|                           | minimap2             | Pairwise alignment method for long reads up to genomes                       | 58   |
|                           | Cactus               | Progressive genomic alignment method allowing integration of more than two genomes at a time | 90   |
|                           | SibeliaZ             | Fast genome aligner of multiple genomes                                      | 140  |
| Read alignment            | minimap2             | Pairwise alignment method for long reads up to genomes                       | 58   |
|                           | NGMLR                | Convex gap cost implementation                                               | 42   |
|                           | Winnowmap            | Improvements for mapping in repetitive regions                               | 59   |
|                           | lra                  | Efficient convex-cost gap penalty sequence and contig aligner                | 60   |
| Graph genome methods      | Giraffe              | Rapid reads to graph aligner                                                 | 45   |
|                           | vg                   | Toolkit to construct and convert graphs with methods to genotype and call variants | 96   |
|                           | minigraph            | A sequence-to-graph mapper and graph constructor based on minimap2          | 97   |
|                           | GraphAligner         | Sequence-to-graph aligner for long reads                                     | 141  |
|                           | GraphTyper2          | Genotyping variants in a graph genome from short reads                      | 100  |
|                           | Paragraph            | Genotyping structural variants in a regional graph genome from short reads   | 101  |
|                           | PanGenie             | k-mer-based genotyping of short reads in a haplotype-resolved graph         | 99   |
| Phasing                   | WhatsHap             | Phasing method for SNVs and smaller indels                                   | 15   |
|                           | HapCut2              | Phasing method for SNVs                                                      | 16   |
| SV calling from alignment | pbsv                 | Joint calling of SVs across samples                                          | 62   |
|                           | Sniffles             | Automatic parameter estimation                                               | 42   |
|                           | CuteSV               | Highly parallelized SV calling                                               | 63   |
|                           | SVIM                 | Uses graph-based clustering of candidates                                    | 61   |
| SV calling from assemblies| dipcall             | Deletion and insertion calling from de novo assembly                         | 89   |
|                           | SVIM-asm             | SV calling from (diploid) de novo assembly                                   | 142  |
|                           | PAV                  | Compares phased assemblies with a reference genome                           | 46   |
| SNV calling               | Clair                | Uses a convolutional neural net                                              | 69   |
|                           | DeepVariant          | Neural network-based SNV caller                                              | 67   |
|                           | Longshot             | Partitioning reads in haplotypes and calling variants in accordance with those haplotypes | 70   |
|                           | Pepper               | Phasing-based SNV calling                                                    | 68   |
| SV merging                | SURVIVOR             | Merging that allows breakpoint inaccuracies                                   | 113  |
|                           | SVAnalyzer           | Assembly based, two samples only                                             | 98   |
|                           | Truvari              | Parameterized stepwise merging including sequence similarity                 | 9    |
|                           | Jasmine              | Merging SV based on sequence similarity                                       | 32   |
| SV genotyping             | cuteSV               | Force-calling of variants from a VCF file                                   | 63   |
|                           | Sniffles             | Uses split reads to identify known SVs over shared breakpoints              | 42   |
|                           | SVJedi               | Compares the alignment of reads against the reference genome and alternative contigs representing the SV to determine the best match | 66   |
|                           | LRcaller             | Genotypes variants of long reads                                            | 11   |
Graph genome methods. Both read alignment and de novo assembly approaches can have systematic issues with complex structural variation, inserted sequences missing from the reference genome, repeat variability and highly polymorphic loci (FIG. 3). Linear reference genomes only represent one allele and thus, do not incorporate polymorphisms and complexity of a population. Reference pan-genome approaches, which combine genomes from multiple individuals within a species, are a better fit to represent genomic diversity. Variant catalogues for pan-genome structures are obtained by ongoing projects using high-quality haplotype-resolved assemblies of diversity panels for the discovery of variants. A reduction of the alignment bias against non-reference alleles is achieved by explicitly taking known population variants into account in the read alignment step. As such, the analysis does not rely on a single reference genome. This goal is realized by graph genome-based tools and their associated data formats, as a way to represent a collection of possible (alternative) sequences. Examples of tools for this purpose include vg, minigraph, the SevenBridges Graph Genome Pipeline, the DRAGEN Graph Mapper and PanGenie. These implementations provide tools to build graphs based on the linear reference genome and a collection of known variants, or alternatively use (haplotype-resolved) assembled contigs. Although a detailed discussion of the methods to construct such pan-genome graphs is beyond the scope of this Review, we note that there are important differences in implementation and data format with regard to compatibility with coordinates on the linear reference genome and storing information of the individual haplotypes that contributed to the included variation. An additional benefit of graph genome methods is that they enable a more correct representation of nested variation, such as smaller variants within inserted sequences.

A major benefit of graph genomes is the genotyping of SVs using short reads. Multiple tools, such as GraphTyper2, Paragraph and tools from the vg package, have been developed specifically for alignment of short-read sequencing data to graph genome structures. SNVs, small indels or SVs within a sample are genotyped as reads following a certain path (‘walk’) through the pan-genome graph (FIG. 4). Graph genotyping methods enable the assessment of variants that remain undetected by the current state-of-the-art short-read SV discovery methods. In the next step, variants that were not yet explicitly encoded in the graph can be identified, with the option to incrementally augment the graph structure with the newfound variation to further improve accuracy. Graph genome methods are reviewed in greater depth elsewhere.

With such graph-based approaches, the often discussed dichotomy of either using an existing reference genome for alignment or constructing a novel reference genome through de novo assembly can potentially be avoided for population studies, as downstream of this step all sequences have to be compared with a single (reference) assembly or a backbone of a pangenomic graph, for identification of variation, annotation and statistical evaluation. However, these approaches are less straightforward in practice than the use of a linear reference genome and are not entirely mature, with competing implementations and data formats. Although graph genome methods are good candidates to solve biases when assessing (structural) genomic diversity, it remains unclear whether these methods will become mainstream in clinical or diagnostic applications, in which a single reference genome is an attractive simplification.

Variant validation and genotyping. To determine whether any given variant constitutes the biological reality and is not just an artefact, it is important to perform validation. Ideally, this is done using orthogonal approaches, to capitalize on the strengths of different technologies. Traditionally, PCR validation of variants has been the method of choice; however, for complex SVs that contain highly repetitive regions, other, non-sequencing-based methods such as optical mapping might be more suitable. Visual inspection of alignments and subsequent manual curation of variant sets are arguably a very accurate validation approach but certainly not feasible for more than a few hundred variants. A semi-automated pipeline, SV-plaudit, has been developed to enable rapid, streamlined and efficient curation of thousands of SVs.

Of similar importance is variant genotyping, which we define as determining the presence and zygosity of a variant. Although the initial discovery of variation is relatively straightforward, obtaining reliable genotypes for a given variant across a population is usually much more difficult. However, knowing the alleles (that is, the genotypes) of variants for a given sample is particularly important in population genetic and evolutionary studies, in which population size estimation and measures of genetic differentiation (such as the fixation index) rely on obtaining accurate allele frequencies of variants. In particular, variants in repetitive regions are more readily genotyped using long reads than using short reads. For SNVs, sophisticated genotyping approaches have been developed that consider

| Category | Tool name | Description | Ref. |
|----------|-----------|-------------|-----|
| Other    | TRiCoLOR  | Detects and genotypes repeat lengths separated by phase | 76  |
|          | Iris      | Local assembly of insertions | 32  |
|          | SVCollector | Optimized sample selection | 43  |
|          | NanoComp  | Comparison of sequencing data | 53  |

HfI, high fidelity; indel, insertions–deletions; ONT, Oxford Nanopore Technologies; PacBio, Pacific Biosciences; SNV, single-nucleotide variant; SV, structural variant; VCF, variant call format.

Table 2 (cont.) An overview of software tools for analysing long-read sequencing data

NATURE REVIEWS | GENETICS
VOLUME 22 | SEPTEMBER 2021 | 581
Fig. 3 | Potential problems for different genome comparison approaches. a | Schematic depiction of a potential problem in a de novo assembly-based approach. The presence of a novel segment N1 in two de novo assemblies, at different locations and, even more so, a sequence variant (red x), poses a challenge to correct reporting by current state-of-the-art methods and variation formats. b | Similar representation of the N1 problem in an alignment-based approach, where the coordinates of N1 are shared, but remain challenging for the identification of the single-nucleotide variant (SNV) or the entire N1 sequence. c | A graph-based representation of N1, which enables a clearer comparison of the variant across the samples, illustrating the potential benefits of graph genomes. R1–R3 represents the backbone of the graph genome and N1, and its SNV represents novel sequencing for a given sample set.

Genomic variant call format (gVCF). Includes not only alternative alleles (standard VCF) but also information about reference allele position that enables merging and full genotyping of variants.

Single-nucleotide polymorphisms (SNPs). Genomic alterations of 1–50 bp that are present in 1% or more of the population.

Important parameters such as mutation dynamics (for example, transition to transversion ratios) and information about non-variant sites to improve genotype accuracy. The concept of a genomic variant call format (gVCF) has been implemented in applications such as freebayes108 and GATK109, which has improved the efficiency of the comparison and made multiple rounds of genotyping obsolete. Another approach is to completely abandon genotype calling and instead calculate posterior probabilities of genotypes to directly incorporate uncertainty in the downstream analysis (for example, ANGSĐ110). Merging SNVs is typically done with tools such as bcftools111 and RTGTools112.

For SVs, the situation is much more complicated, as establishing homology of variants between samples is not straightforward. One of the first approaches to be developed is based on 50% reciprocal overlap, which allows two SVs to be merged if they overlap substantially. Although this works well for large copy number variation events, there may be some limitations for smaller SVs (for example, 50 bp to 1 kbp) with more localized breakpoints. Another approach is to require breakpoints from each individual to be approximately in agreement to establish that a variant in two samples is indeed homologous (for example, SURVIVOR merge113). In some cases, such as when two insertions are homologous, but their sequence slightly deviates, an approach based on breakpoints may be too conservative, and some tools have been used to attempt to address this issue (for example, Truvar114, SVanalyzer and Jasmine115). However, at present, no universal standards are available for the thresholds. Thus, approaches rely on arbitrary thresholds of breakpoint distance and sequence similarity. Deletions are arguably the most straightforward type of variation to genotype, but calling heterozygotes for even this seemingly simple type of SV can be difficult116.

Tools such as Sniffles and SVJedi are capable of genotyping SVs based on a candidate VCF file, following an initial step of SV discovery based on the long-read alignments117.

Another potentially very powerful approach to improve SV genotypes is to harness the information contained in a sampling scheme consisting of phylogenetically distant populations (FIG. 4c). In this approach, basic population genetic assumptions are made to reduce the number of false positives for genotyped SVs. After a sufficient number of generations (4Ne, where Ne = effective population size), variation is likely fully sorted and no polymorphisms should occur across lineages any more, assuming that there are no repeated mutations at the same locus (that is, the infinite sites model)118. Any variants exhibiting polymorphic genomes across the divergent lineages are excluded. Although this approach neglects the fact that certain types of SV have much higher mutation rates and thus indeed have the potential for repeated mutations (for example, variation within tandem repeat arrays), it provides a first step towards more reliable SV genotyping. This approach has recently been successfully applied in the corvids crows and jackdaws119.

Prediction of functional impact. The mathematical framework for the analysis of (small) genetic variants predates the advent of high-throughput sequencing by almost a century and is therefore well established. Large-scale single-nucleotide polymorphism (SNP)-array-based GWAS projects enabled the interrogation of thousands of variants and haplotypes for their association with disease. Although quality assessment steps such as principal component analysis and testing for Hardy–Weinberg equilibrium still hold for indel variants (that is, >50 bp), these models do not necessarily
cover all types of SV, for example, in the case of a continuous spectrum of repeat lengths\textsuperscript{[116].} A solution, albeit with loss of resolution, would be to binarize the distribution into ‘reference’ and ‘expanded’ alleles, but historically it has been difficult to unambiguously establish a cut-off length. Association testing of the role of partially overlapping variants for a certain trait requires an approach conceptually similar to that used for burden analysis in rare variant association studies.

Whereas classification of the functional impact of small variants on protein function for synonymous, missense and loss-of-function variants is relatively mature with tools such as the Ensembl VEP\textsuperscript{[117]}, it is less straightforward to judge the impact of SVs on the expression of nearby genes. This is mainly because it is unclear how the length of an SV impacts the surrounding genomic region and it is often hard to obtain robust allele frequencies for SVs\textsuperscript{[114].} For functional annotation and pathogenicity prediction, approaches using joint linear models\textsuperscript{[118]}, supervised learning\textsuperscript{[119]} and existing databases\textsuperscript{[120]} have been developed, and there are promising examples demonstrating that SVs are indeed associated with important traits of interest\textsuperscript{[118,119].}

**Conclusions**

Ongoing significant technological improvements have paved the way to apply long-read sequencing to population-scale sequencing projects and demonstrate

---

**Fig. 4 | Genotyping of SVs and SNVs across a population set.**

- **a** | Graph genome-based genotyping of a region with multiple alleles between two genome segments (green and pink). Insertions of different sizes (yellow) can be genotyped at the same locus using spanning reads (blue and purple) to identify the presence of two different alleles.
- **b** | An example of structural variants (SVs) and single-nucleotide variants (SNVs) across different unique and repeat regions being correctly or incorrectly genotyped based on read length.
- **c** | A phylogenetically informed filtering approach for SVs. Assuming that after a sufficiently long time ($4N_e$ generations, where $e$ = effective population size) most or all genetic variation should be fully sorted between two clades; variants that do not adhere to this assumption and are polymorphic across clades (for example, variant 3) can be removed. Although this approach is certainly very conservative and ignores the fact that some types of variation exhibit repeated mutations on the same locus, it can be considered a first step towards more reliable genotyping of SVs.
that this sequencing approach is here to stay. This process already started with the first larger data sets generated by targeted sequencing of certain genes (BOX 2) and continues with an increasing number of projects that leverage long reads at scale (FIG. 1; TABLE 1). The analysis of population-scale long-read sequencing data sets remains challenging, with the read alignment-based approach currently being the most feasible. Nevertheless, we anticipate this to change to alignment of either haplotype-resolved de novo assemblies or individual sequencing reads to graph genome structures. This development will have a profound impact on the field and holds the promise of improved variant representation and complexity of the underlying biology, but would require a paradigm shift from a linear to a more complex version of the reference genome.

PacBio and ONT lead the current development of long-read sequencing for multiple applications. However, other companies (for example, Base4, Quantapore and Omniome) are developing novel long-read technologies, the viability of which remains to be evaluated in the coming years. Although not discussed here, improved DNA extraction, conservation and library preparation is also adding to the rapid growth of long-read sequencing population studies. Among the biggest achievements in recent years is the generation of sequence reads of 4 Mbp and longer; although this is not yet routinely possible without compromising yield. Once sequencing reads routinely approach chromosome length, the process of de novo assembly seems obsolete; however, whether such reads can be directly used in a framework that is based on de novo assemblies instead of read alignment remains to be seen.

**Future directions**

The future of long-read population-scale sequencing holds many opportunities for multiple types of omics assays. For example, both the PacBio and ONT platforms are able to simultaneously detect the nucleotide sequence and modifications of DNA such as 5-methylcytosine. The identification of such modifications has unprecedented implications for epigenetics and the analysis of DNA damage. More recent versions of the ONT base callers are trained to detect common nucleotide modifications, which together with the plateauing accuracy potentially alleviates the need to store raw data. Several studies have shown excellent reproducibility and correlation with bisulfite sequencing, suggesting that nanopore sequencing could become the gold standard for detecting methylation patterns. Although methods tailored to short-read bisulfite sequencing exist, there is a lack of statistical methods for differential methylation assessment that leverages the unique features of large distance phased of modifications in parental haplotypes. Detection of nucleotide modifications further opens up a wealth of opportunities for specialized assays such as chromatin accessibility profiling and replication fork detection.

Complementary to DNA-based population sequencing, long-read sequencing of mRNA and complementary DNA (cDNA) also enable the identification of isoform diversity. Multiple pipelines have been developed to investigate known and novel isoforms, but the field is far from mature. A survey of multiple tissues has already been undertaken, and an extension of this to the population scale, such as in the short-read GTEx project, is highly likely to yield valuable information about transcript structure and the influence of regulatory (structural) variation. Long-read sequencing approaches have also been extended to the direct sequencing of proteins and single-cell transcriptomics. Although these applications are likely to lead to biologically fascinating insights, the implications for population studies remain unclear.

Alongside the technological improvements in long-read sequencing, computational analysis has also improved, which is key to enabling population-scale projects. Analyses that took weeks to months to accomplish a year ago can now be completed within a day to a week and at a lower cost. Recent advances, such as pan-genome graphs, have the potential to address this challenge. Furthermore, the use of pan-genome graphs could indeed improve the analysis itself, as they overcome the problem of a linear reference bias by including different alleles. Another related computational challenge is the accurate and rapid genotyping of complex alleles. Here, graph genomes have already shown significant benefits, although the process of obtaining a fully genotyped population-level VCF is still far from trivial. This is due to the lack of a gVCF for SV representation, to represent information not only about the alternative alleles (that is, SV) but also about reference alleles. For SV, this allows the easy comparison of samples and is a requirement for future SV studies.

Despite significant advances in long-read sequencing, several challenges remain to be addressed. The frequently discussed issue regarding the lack of precision and lack of sensitivity in identifying SNVs and small indels, especially involving homopolymers, is likely to be resolved by advancements in sequencing accuracy. However, difficulties remain in assessing variation in complex regions such as segmental duplications, ribosomal DNA (rDNA) tandem arrays, telomeres or centromeres. Spurred by the efforts led by the T2T consortium, which aims to provide the full linear nucleotide sequence of all human chromosomes, new software tools are being developed that specifically aim to resolve these large tandem arrays and also to assess the allelic variation within them. However, whether this solves the problem completely remains to be determined, as at the time of writing even the T2T reference genome has a few gaps remaining and only represents one ethnicity.

In this Review, we provide a snapshot of the present state of large-scale long-read sequencing and discuss the exciting developments in biotechnology and bioinformatics. Despite its challenges, we argue that long-read sequencing has contributed immensely to the advancement of genomics in humans, model organisms and beyond, and that this is the way forward for population-scale studies.

Published online 28 May 2021
Hiatt, S. M. et al. Long-read genome sequencing. Goodwin, S., McPherson, J. D. & McCombie, W. R. et al. Structural Chaisson, M. J. P. et al. Multi-platform discovery of neuronal intranuclear inclusion disease. repeat expansions in NOTCH2NLC associated with Nat. Rev. Genet. Piercing the dark matter: bioinformatics of long-range topologically associated domains in schizophrenia. structural variants localizing to boundaries of genome-wide association studies for predicting disease risk. PLoS ONE 14, e0220125 (2019). Hulteberg, K. et al. Long-read discovery and genotyping of structural variation from long-read haplotype sequence data. Genome Res. 27, 677–685 (2017). Goodwin, S. J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17, 535–551 (2016). Ho, S. S., Urban, A. E. & Mills, R. E. Structural genome in human genomes in natural populations. Nat. Rev. Genet. 17, 1189–1197 (2016). Luo, R. et al. Exploring the limit of using a deep neural network-based phasing. bioRxiv doi.org/10.1101/2021.01.14.426684 (2021). Gunther, T. & Nettelblad, C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. PLoS Genet. 15, e1008302 (2019). Kellbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. Genome Res. 20, 487–493 (2010). Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 5086–5100 (2018). Jain, C., Rhee, A., Hansen, N., Koren, S. & Philippy, A. M. Long-read mapping method for highly repetitive reference sequences. bioRxiv https://doi.org/10.1101/2021.03.04.433952 (2021). Ren, J. & Chaisson, M. J. P. Ira the long read aligner for sequences and contigs. bioRxiv https://doi.org/10.1101/2020.11.15.382373 (2020). Keller, D. & Vingrong, M. SVM: structural variant identification using mapped long reads. Bioinformatics 35, 2907–2915 (2019). Engelsen, C. A. Salerno, W. J. & Reid, J. C. PBhoney: identifying genomic variants via long-read discordance and interrupted mapping. BMC Bioinformatics 15, 180 (2014). Jiang, T. et al. Long-read-based human genomic structural variation detection with cutesV Genome Biol. 21, 189 (2020). De Coster, W. & Van Broeckhoven, C. Newest methods for detecting structural variations. Trends Biotechnol. 37, 972–982 (2019). Tham, C. V. et al. NanVar: accurate characterization of patients’ genomic structural variants using long-depth nanopore sequencing. Genome Biol. 21, 56 (2020). Lecompte, L., Peterlongo, P., Lavener, D. & Lemaître, C. SVDee: genotyping structural variations with long reads. Bioinformatics 36, 4568–4575 (2020). Pipkin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. Bioinformatics 36, 985–987 (2020). Shafin, K. et al. Haplotype-aware variant calling enables high accuracy in nanopore long-reads using deep neural networks. bioRxiv https://doi.org/10.1101/2021.03.03.435952 (2021). Luo, L. et al. Exploring the limit of using a deep neural network on pileup data for single nucleotide variant calling. Nat. Mach. Intell. 2, 220–227 (2020). Edge, P. & Bansal, V. Longshot enables accurate variant calling in diploids and haplotypes using single-molecule long read sequencing. Nat. Commun. 10, 4660 (2019). Ong, C. et al. precisionFDA Truth Challenge V2: calling variants from short- and long-reads in difficult-to-map regions. bioRxiv https://doi.org/10.1101/2020.09.25.382471 (2020).
149. Haro-Moreno, J. M., López-Pérez, M. & Rodríguez-Valera, F. Long read metagenomics, the next step? bioRxiv https://doi.org/10.1101/2020.11.11.378109 (2020).

150. Leija-Salazar, M. et al. Evaluation of the detection of GBA missense mutations and other variants using the Oxford Nanopore MiniION. Mol. Genet. Genom. Med. 7, e564 (2019).

151. Gilpatrick, T. et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. Nat. Biotechnol. 38, 433–438 (2020).

152. Kovaka, S., Fan, Y., Ni, B., Timp, W. & Schatz, M. C. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. Nat. Biotechnol. 39, 431–441 (2020).

153. Payne, A. et al. Readfish enables targeted nanopore sequencing of gigabase-sized genomes. Nat. Biotechnol. 39, 442–450 (2020).

154. Miller, D. E. et al. Targeted long-read sequencing resolves complex structural variants and identifies missing disease-causing variants. bioRxiv https://doi.org/10.1101/2020.11.03.365395 (2020).

155. Tyson, J. R. et al. Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. bioRxiv https://doi.org/10.1101/2020.09.04.283077 (2020).

156. Doddapaneni, H. et al. Oligonucleotide capture sequencing of the SARS-CoV-2 genome and subgenomic fragments from COVID-19 individuals. bioRxiv https://doi.org/10.1101/2020.07.27.223595 (2020).

157. Butler, D. et al. Shotgun transcriptome, spatial omics, and isothermal profiling of SARS-CoV-2 infection reveals unique host responses, viral diversification, and drug interactions. Nat. Commun. 12, 1660 (2021).

158. Peto, L. et al. Diagnosis of SARS-CoV-2 infection with LamPORE, a high-throughput platform combining loop-mediated isothermal amplification and nanopore sequencing. medRxiv https://doi.org/10.1101/2020.09.18.20195570 (2020).

Acknowledgements
The authors thank A. Wenger, P. Rescheneder and Anonymous Giraffe for helpful discussions and feedback. This work was supported in part by awards from the US National Institutes of Health (UM1-HG008898) and a postdoctoral fellowship of the Research Foundation – Flanders (FWO).

Author contributions
The authors contributed equally to all aspects of the manuscript.

Competing interests
W.D.C. and F.J.S. have received sponsored travel from PacBio and/or Oxford Nanopore. M.H.W. declares no competing interests.

Peer review information
Nature Reviews Genetics thanks B. V. Halldursson, A. Ameur, C. Lemaitre, 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.

Supplementary information
The online version contains supplementary material available at https://doi.org/10.1038/s41576-021-00367-3.

RELATED LINKS
NIH All of Us: https://allofus.nih.gov/
NIH Center for Alzheimer’s and Related Dementias: https://www.nia.nih.gov/research/center
Phv: https://github.com/PacificBiosciences/phv
SVanalyzer: https://github.com/nhansen/SVanalyzer

© Springer Nature Limited 2021