Towards population-scale long-read sequencing

Wouter De Coster[1,2,5], *Matthias H. Weissensteiner*[3,5] *and Fritz J. Sedlazeck*[4]✉

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.

Genome-wide association studies
(GWAS). Studies involving a statistical approach in genetics to identify variants that correlate with a certain phenotype (for example, a disease).

[1]*Applied and Translational Neurogenomics Group, VIB Center for Molecular Neurology, VIB, Antwerp, Belgium.*

[2]*Applied and Translational Neurogenomics Group, Department of Biomedical Sciences, University of Antwerp, Antwerp, Belgium.*

[3]*Department of Biology, Penn State University, Pennsylvania, PA, USA.*

[4]*Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA.*

[5]*These authors contributed equally: Wouter De Coster, Matthias H. Weissensteiner.*

✉*e-mail: fritz.sedlazeck@bcm.edu*

https://doi.org/10.1038/s41576-021-00367-3

Sequencing the DNA or mRNA of multiple individuals of one or more species (that is, population-scale sequencing) aims to identify genetic variation at a population level to address questions in the fields of evolutionary, agricultural and medical research. Previous population studies, including genome-wide association studies (GWAS), have not been able to exhaustively characterize the genetic factors underlying human traits and diseases[1]. There has been much speculation about the source of this 'missing heritability', often pointing to both structural variants (SVs) and rare variants[2,3]. SVs account for a greater total number of nucleotide changes in human genomes than the far more numerous single-nucleotide variants (SNVs)[4]. To date, such population studies have relied mostly on high-throughput short-read sequencing technologies, which produce reads ranging from 25 bp to 400 bp in length[5]. However, short reads have important limitations in characterizing repetitive regions[6,7]. DNA repeats act as the genomic substrate to facilitate SV formation[8] while also hampering SV discovery owing to read alignment inaccuracies. Even in a non-repetitive genome, variations such as insertions (especially for alleles longer than the read length[7]) or other modifications (for example, methylation) would be missed by an approach relying solely on short reads.

Long-read sequencing has emerged as superior to short-read sequencing and other methods (for example, arrays) for the identification of structural variation, as shown by the Genome in a Bottle (GIAB) and Human Genome Structural Variation (HGSV) consortia, which combined multiple technologies to comprehensively characterize structural variation in human genomes[9,10]. 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 anaemia[11,12]. Analysis of 26 maize genomes[13] 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 phasing[14–16], assessing complex small variants[17] and has been applied to find disease-associated alleles[18–20]. For de novo assemblies, multiple methods have been published over recent years to promote the use of long reads[21–25].

Ongoing advances in sequencing technology and bioinformatics have paved the way to achieving long-read sequencing on a population scale[26]. 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 accurate[27]. The ONT PromethION platform can produce much longer reads (up to 4 Mbp[28]), 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 platforms[29,30] (in-depth technical review and further comparison of long-read sequencing platforms available elsewhere[31]). Within the past 2 years, multiple studies have applied long-read sequencing to answer various questions in

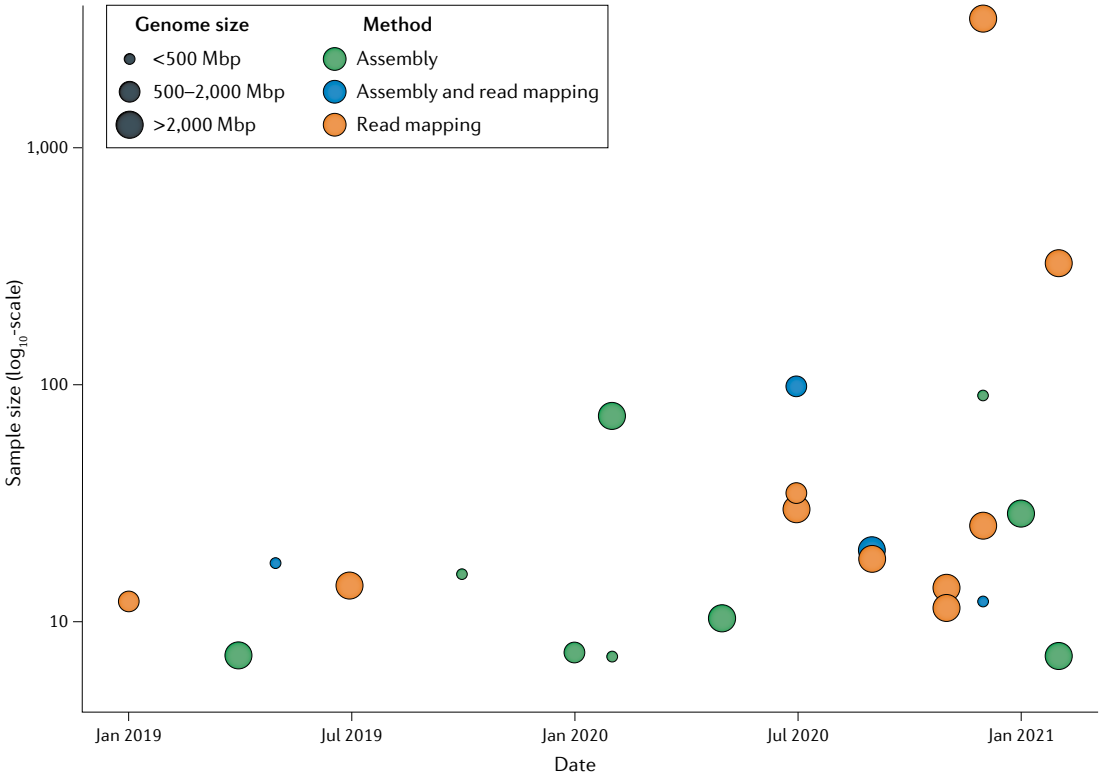


Fig. 1 | **Overview of population-scale studies using long-read sequencing.** Studies published in 2019–2021 in which five or more samples were sequenced are included. Genome size of study organisms is viewed in three different categories (<500 Mbp, 500–2,000 Mbp and >2,000 Mbp), and the methodological 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.

multiple different organisms[32–35] (FIG. 1; TABLE 1). The largest human-focused long-read sequencing study to date investigated the genomic diversity of 3,622 Icelandic genomes[11], 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 project[36]. Aside from human studies, long-read sequencing has been applied on a population scale to discover structural variation associated with phenotypes in crops[32,33], fruitflies[34] and songbirds[35], 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 setting[37,38]. 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)[39], the identification of rare variants (and potentially associated diseases) in a population usually requires sample sizes that are many orders of magnitude higher[40]. 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

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 Mbp.

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 Mbp) but less accurate (average 3–8%) 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.

Table 1 | **An overview of long-read-based population studies**

Study	Organism and category	Technology[a] and analysis approach	Sample size[b]	Genome size (Mbp)	Ref.
Kou et al. (2020)	Rice Agriculture	PacBio Assembly comparison and read mapping	15 (LR); 393 (SR)	430	129
Weissensteiner et al. (2020)	Crow Evolution	PacBio Read mapping	33 (LR); 127 (SR)	1,300	35
Chakraborty et al. (2019)	*Drosophila* Evolution	PacBio Assembly comparison	14 (LR)	180	34
Jiao & Schneeberger (2020)	*Arabidopsis* Evolution	PacBio Assembly comparison	7 (LR)	135	130
Alonge et al. (2020)	Tomato Agriculture	ONT Read mapping	100 (LR)	950	32
Beyter et al. (2020)	Human Human evolution	ONT Read mapping	3622 (LR)	3,200	11
Tusso et al. (2019)	Yeast Evolution	ONT and PacBio Assembly comparison and read mapping	17 (LR); 161 (SR)	12	30
Liu et al. (2020)	Soy bean Agriculture	PacBio Assembly comparison	26 (LR)	1,150	33
Chawla et al. (2020)	Rapeseed Agriculture	ONT and PacBio Read mapping	12 (LR)	1,132	131
Hiatt et al. (2020)	Human Human evolution	PacBio Assembly comparison and read mapping	18 (LR)	3,200	18
Mitsuhashi et al. (2020)	Human Human evolution	ONT and PacBio Read mapping	37 (LR)	3,200	132
Shafin et al. (2020)	Human Human evolution	ONT Assembly comparison	11 (LR)	3,200	25
De Roeck et al. (2020)	Human Human evolution	ONT Read mapping	11 (LR)	3,200	133
Chaisson et al. (2019)	Human Human evolution	ONT and PacBio Assembly comparison	9 (LR)	3,200	10
Morena-Barrio et al. (2020)	Human Human evolution	ONT Read mapping	19 (LR)	3,200	19
Song et al. (2020)	Rapeseed Agriculture	PacBio Assembly comparison	8 (LR)	1,132	134
Sone et al. (2019)	Human Human evolution	ONT and PacBio Read mapping	17 (LR)	3,200	20
Kim et al. (2020)	*Drosophila* Evolution	ONT Assembly comparison	101 (LR)	180	135
Pauper et al. (2020)	Human Human evolution	PacBio Read mapping	15 (LR)	3,200	136
Ebert et al. (2020)	Human Human evolution	PacBio Assembly comparison	64 (LR)	3,200	46
Quan et al. (2020)	Human Human evolution	ONT Read mapping	25 (LR)	3,200	137
Hufford et al. (2021)	Maize Agriculture	PacBio Assembly comparison	26 (LR)	2,200	13
Hu et al. (2021)	Maize Agriculture	PacBio Assembly comparison	6 (LR)	2,200	138
Wu et al. (2021)	Human Human evolution	ONT and PacBio Read mapping	405 (LR)	3,200	12

[a]Two main platforms are used in long-read sequencing projects, Pacific Biosciences (PacBio) high fidelity (HiFi) and Oxford Nanopore Technologies (ONT) PromethION. [b]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[143]. 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[144]. Other efforts have been directed at improving the assembly step. metaFlye[145] 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 longer read length to improve the classification[146,147].

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.[148] 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[149].

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[41]. 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[42]) 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[43], 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[44]. 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[33]. Other studies have used SVCollector[43] 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[25,32]. 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.

High molecular weight DNA (HMW DNA). Extracted DNA containing long DNA molecules (typically ≥50 kbp average molecule size).

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[11]. This strategy has already been applied using diversity panels of human SVs to discover novel expression quantitative trait loci (eQTLs)[45,46] and signatures of evolutionary adaptation[47]. 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[31].

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[48] 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[49], 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[50]. 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[6,7,21]. TABLE 2 lists computational tools that are commonly used in long-read sequencing projects and these are reviewed in-depth elsewhere[6,7]. 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[150] to using the Cas9 system[151] and an in silico sequencer-based selection (for example, Uncalled[152] or Readfish[153]). These approaches typically can target 10–20 kbp 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[151]. 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[150,154]. 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 reads as well as sequencer-based targeting of regions (~30% enrichment on target)[151]. 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[155]. In addition, loop-mediated isothermal amplification (LAMP) and/or capture methods have been very effective in studying the diversity of SARS-CoV-2 isolates[156,157]. 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.[158]).

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[51] and assemblies[52] or in fine-scale resolution of SV breakpoints[11]. 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[53,54]. 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 (FIG. 1; 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[7]. If the reference genome is incomplete, incorrect, fragmented or too divergent from the focal sample, it will lead to biases in the downstream analysis[55,56].

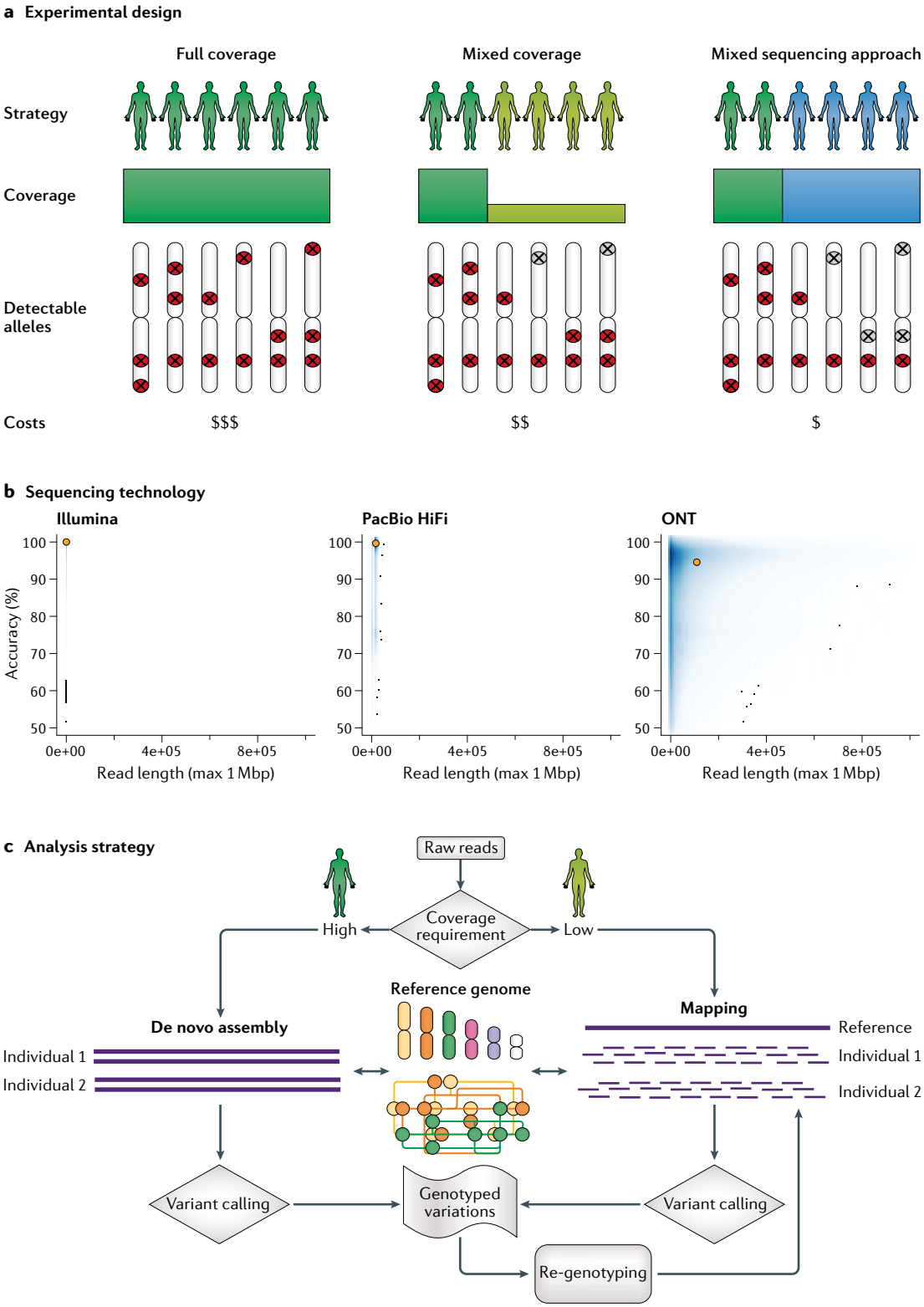
The software for long-read sequence data analysis is under constant development, and alignment methods in particular have become much faster in recent years (TABLE 2). The NGMLR[42] and LAST[57] methods speed up the alignment process and improve the accuracy of long-read alignment. The minimap2 aligner is considerably faster than its competitors while often delivering similar results, and thus it is currently the most popular, widely accepted long-read aligner[58]. Two noteworthy recent innovations are Winnowmap, which improves alignments (specifically in repetitive regions)[59], and lra, which improves the alignment in the presence of SVs[60].

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[42], SVIM[61], PBHoney[62], CuteSV[63] 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[7,64]). Recently developed tools are removing the need for high sequencing coverage by enabling SV calling[42,65] and genotyping[42,66] 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 SNV 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[67] Pepper[68], Clair[69] (both using deep learning) and LongShot[70] (which explicitly requires alleles to be concordant with the haplotype structure), which also outperforms Illumina-based SNV calling[71]. 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[72]. 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[73]. Some tools have been developed specifically for this purpose[74], such as tandem-genotypes[75] and TRiCoLOR[76]. 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[77], 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.[78]) 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[35].



Population-scale de novo assemblies. Many reference genomes based on short-read sequencing are incomplete or highly fragmented with many gaps[80]. Furthermore, hundreds of megabases of population- and individual-specific sequences are absent from the human reference genome[81]. 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[82]. Therefore, creating and comparing de novo assemblies is desirable (FIG. 1).

The increased availability and affordability of long-read sequencing data have led to an explosion of faster and more accurate genome assembly tools (TABLE 2), of

◀ Fig. 2 | **Overview of long-read population study design. a** | The experimental design of three different approaches is outlined. In the first strategy (left), all samples are sequenced at medium to high coverage by long-read sequencing. In the second approach (middle), a proportion of the samples are sequenced with medium to high coverage and the remainder using low coverage by long-read sequencing (similar to the initial 1000 Genomes project). In the third approach (right), a proportion of the samples are sequenced at medium to high coverage by long-read sequencing and the remainder by short-read sequencing. The decision of which approach to take will affect the ability to detect common (red symbols) or rare (grey symbols) events in the population. The decision also depends on the available budget, existing data and the sample DNA availability. **b** | Overview of current established sequencing technologies based on CHM13 sequencing data[79]: Illumina, Pacific Biosciences (PacBio) High Fidelity (HiFi) reads or ultra-long reads from Oxford Nanopore Technologies (ONT). The N50 read length and average read accuracy are highlighted in orange. Although each technology has advantages and disadvantages, HiFi and ONT are the most promising for future applications. **c** | Overview of analysis strategies. Although multiple approaches are available, the main decision is whether to use an alignment-based approach or a de novo assembly-based approach, which has implications for sequencing requirements and the approaches, resolution and comprehensiveness of downstream computational analysis.

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[83]. 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[84,85]. 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[86] (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[25,46], 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[25].

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[87], Assemblytics[88], minimap2 or dipcall[89] and SVIM-asm[61]) 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. 3a). 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. 3a), 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[90], 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[22]. 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[86,91].

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*[77] and *SMN1* and *SMN2* (REF.[92]) 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.

Variant call format
(VCF). A tabular file consisting of a header and entries that hold information about each variant detected.

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
	hifiasm	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

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

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

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

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[93,94] (FIG. 3c). 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[46]. 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[95]. Examples of tools for this purpose include vg[96], minigraph[97], the SevenBridges Graph Genome Pipeline[98], the DRAGEN Graph Mapper and PanGenie[99]. 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[97]. 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[94].

A major benefit of graph genomes is the genotyping of SVs using short reads. Multiple tools, such as GraphTyper2[100], Paragraph[101] and tools from the vg package[45,96], 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[96,101] (FIG. 4a). Graph genotyping methods enable the assessment of variants that remain undetected by the current state-of-the-art short-read SV discovery methods[46]. 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[98,102]. Graph genome methods are reviewed in greater depth elsewhere[94,95,103].

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[104]; however, for complex SVs that contain highly repetitive regions, other, non-sequencing-based methods such as optical mapping might be more suitable[46]. 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[105].

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 F_{ST}) rely on obtaining accurate allele frequencies of variants[106]. In particular, variants in repetitive regions are more readily genotyped using long reads than using short reads (FIG. 4b). For SNVs, sophisticated genotyping approaches have been developed that consider

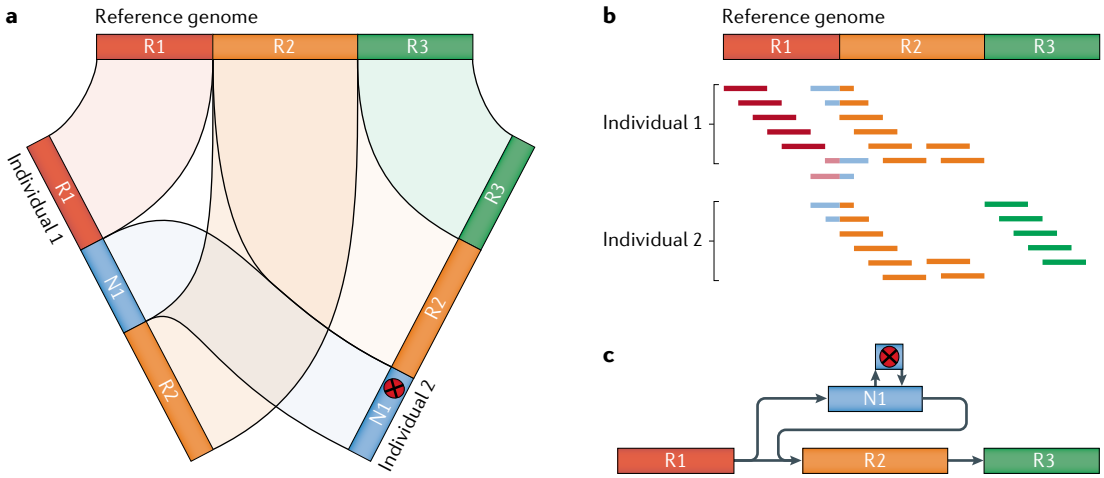


Fig. 3 | **Potential problems for different genome comparison approaches. a** | Schematic depiction of a potential problem in a de novo assembly-based approach. The presence of a novel segment N1 in two de novo assemblies, at different locations and, even more so, a sequence variant (red x), poses a 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 t he 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 allelic 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[107]. The concept of a genomic variant call format (gVCF) has been implemented in applications such as freebayes[108] and GATK[109], 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, ANGSD[110]). Merging SNVs is typically done with tools such as bcftools[111] and RTGTools[112].

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 merge[113]). 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, Truvari[9], SVanalyzer and Jasmine[32]). However, at present, no universal standards are available for the thresholds. Thus, approaches rely on arbitrary thresholds of breakpoint distances 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 difficult[114]. 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 alignments[66].

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 ($4N_e$, where N_e = 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)[115]. Any variants exhibiting polymorphic genotypes 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 jackdaws[35].

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[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[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[114]. For functional annotation and pathogenicity prediction, approaches using joint linear models[118], supervised learning[119] and existing databases[120] have been developed, and there are promising examples demonstrating that SVs are indeed associated with important traits of interest[118,119].

Conclusions

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

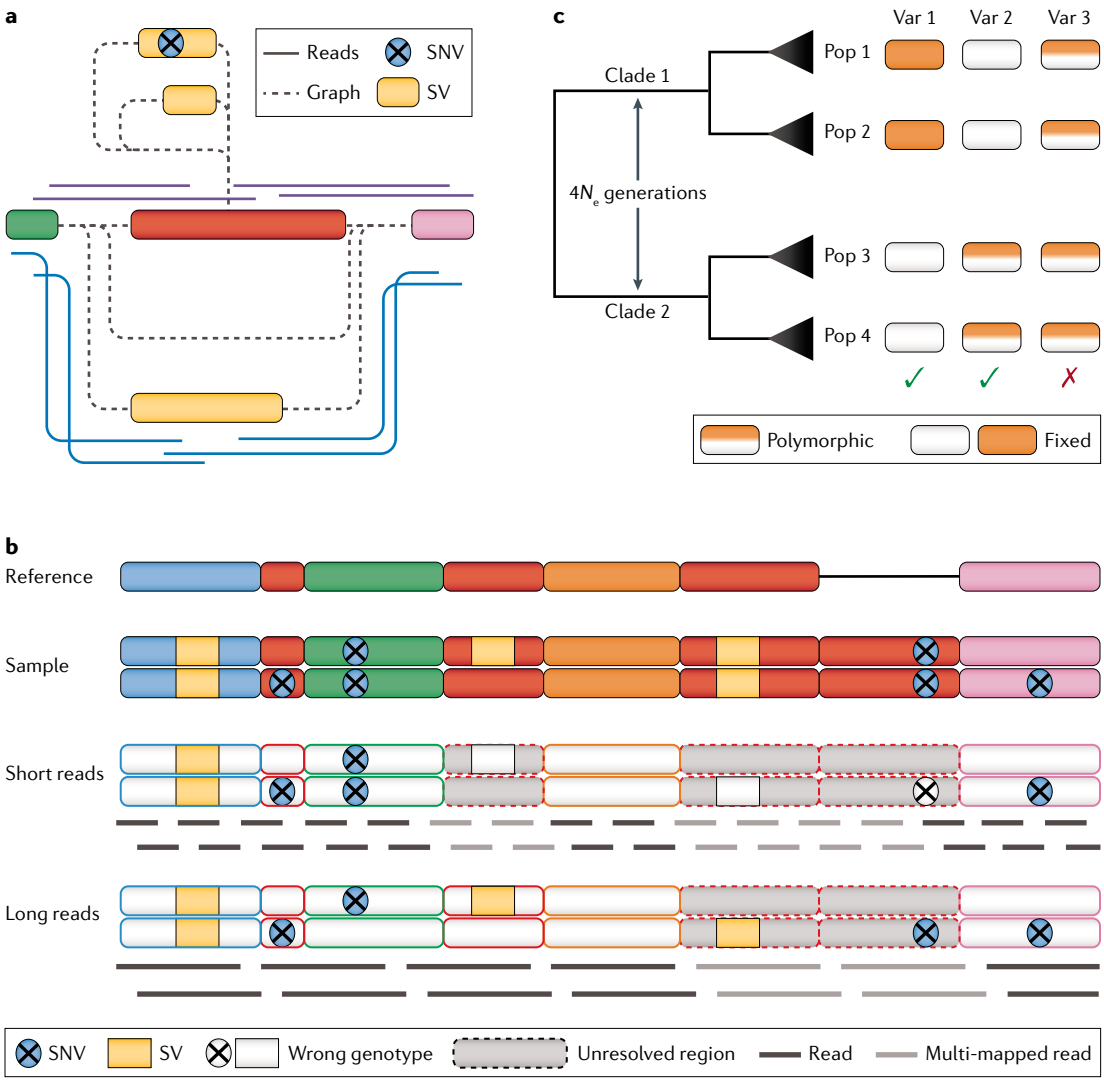


Fig. 4 | **Genotyping of SVs and SNVs across a population set. a** | Graph genome-based genotyping of a region with multiple alleles between two genome segments (green and pink). Insertions of different sizes (yellow) can be genotyped at the same locus using spanning reads (blue and purple) to identify 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[31]. 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[28]. 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[121]. 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[122]. 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 phasing of modifications in parental haplotypes. Detection of nucleotide modifications further opens up a wealth of opportunities for specialized assays such as chromatin accessibility profiling[123] and replication fork detection[124].

Complementary to DNA-based population sequencing, long-read sequencing of mRNA and complementary DNA (cDNA) also enable the identification of isoform diversity[125]. 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[125], 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[126] and single-cell transcriptomics[127]. Although these applications are likely to lead to biologically fascinating insights, the implications for population studies remain unclear[127].

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[24,86,128]. However, some conceptual challenges remain, such as the representation of nested and highly complex variation[97]. Recent advances, such as pan-genome graphs, have the potential to address this challenge[97]. 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[96,100,101]. 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 SNV, 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[27,68]. 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

1. Patron, J., Serra-Cayuela, A., Han, B., Li, C. & Wishart, D. S. Assessing the performance of genome-wide association studies for predicting disease risk. *PLoS ONE* **14**, e0220215 (2019).
2. Hartman, K. A., Rashkin, S. R., Witte, J. S. & Hernandez, R. D. Imputed genomic data reveals a moderate effect of low frequency variants to the heritability of complex human traits. *bioRxiv* https://doi.org/10.1101/2019.12.18.879916 (2019).
3. Halvorsen, M. et al. Increased burden of ultra-rare structural variants localizing to boundaries of topologically associated domains in schizophrenia. *Nat. Commun.* **11**, 1842 (2020).
4. Huddleston, J. et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–685 (2017).
5. 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).
6. Ho, S. S., Urban, A. E. & Mills, R. E. Structural variation in the sequencing era. *Nat. Rev. Genet.* **21**, 171–189 (2020).
7. Mahmoud, M. et al. Structural variant calling: the long and the short of it. *Genome Biol.* **20**, 246 (2019).
The review articles by Ho et al. and Mahmoud et al. provide an overview of structural variation calling and why long reads are important.
8. Weckselblatt, B. & Rudd, M. K. Human structural variation: mechanisms of chromosome rearrangements. *Trends Genet.* **31**, 587–599 (2015).
9. Zook, J. M. et al. A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38**, 1347–1355 (2020).
10. Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
11. Beyter, D. et al. Long read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *bioRxiv* https://doi.org/10.1101/848366 (2020).
A large population-scale sequencing study involving 3,622 individuals, using the ONT PromethION platform to identify diversity and correlate it with disease phenotypes in an Icelandic population.
12. Wu, Z. et al. Structural variants in Chinese population and their impact on phenotypes, diseases and population adaptation. *bioRxiv* https://doi.org/10.1101/2021.02.09.430378 (2021).
13. Hufford, M. B. et al. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *bioRxiv*, https://doi.org/10.1101/2021.01.14.426684 (2021).
14. Majidian, S. & Sedlazeck, F. J. PhaseME: automatic rapid assessment of phasing quality and phasing improvement. *Gigascience* **2020**, giaa078 (2020).
15. Martin, M. et al. WhatsHap: fast and accurate read-based phasing. *bioRxiv* https://doi.org/10.1101/085050 (2016).
16. Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* **27**, 801–812 (2017).
17. Wagner, J. et al. Benchmarking challenging small variants with linked and long reads. *bioRxiv* https://doi.org/10.1101/2020.07.24.212712 (2020).
18. Hiatt, S. M. et al. Long-read genome sequencing for the diagnosis of neurodevelopmental disorders. *bioRxiv* https://doi.org/10.1101/2020.07.02.185447 (2020).
19. de la Morena-Barrio, B. et al. Long-read sequencing resolves structural variants in SERPINC1 causing antithrombin deficiency and identifies a complex rearrangement and a retrotransposon insertion not characterized by routine diagnostic methods. *bioRxiv* https://doi.org/10.1101/2020.08.28.271932 (2020).
20. Sone, J. et al. Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. *Nat. Genet.* **51**, 1215–1221 (2019).
21. 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).
22. Chin, C.-S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
23. 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).
24. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
25. 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).
This study reports a method to rapidly assemble and polish genomes, showcasing the throughput on ONT data by establishing 11 human genomes in 9 days.
26. Brenner, S. Life sentences: Detective Rummage investigates. *Genome Biol.* **3**, comment1013.1 (2002).
27. 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).
28. Payne, A., Holmes, N., Rakyan, V. & Loose, M. BulkVis: a graphical viewer for Oxford Nanopore bulk FAST5 files. *Bioinformatics* **35**, 2193–2198 (2018).
29. Fatima, N., Petri, A., Gyllensten, U., Feuk, L. & Ameur, A. Evaluation of single-molecule sequencing technologies for structural variant detection in two swedish human genomes. *Genes* **11**, 1444 (2020).
30. Tusso, S. et al. Ancestral admixture is the main determinant of global biodiversity in fission yeast. *Mol. Biol. Evol.* **36**, 1975–1989 (2019).
31. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
This review provides key insights into the long-read sequencing machines.
32. Alonge, M. et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**, 145–161.e23 (2020).
This study reports the population-scale sequencing for a plant (tomato) and details the impact of the detected variation on phenotypes.
33. Liu, Y. et al. Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162–176.e13 (2020).
34. Chakraborty, M., Emerson, J. J., Macdonald, S. J. & Long, A. D. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat. Commun.* **10**, 4872 (2019).
35. Weissensteiner, M. H. et al. Discovery and population genomics of structural variation in a songbird genus. *Nat. Commun.* **11**, 3403 (2020).
A large-scale sequencing study in crows highlights segregation of structural variation in natural populations.
36. National Human Genome Research Institute. Advancing the reference sequence of the human genome. *Genome.gov* https://www.genome.gov/news/news-release/NIH-funds-centers-for-advancing-sequence-of-human-genome-reference (2019).
37. Levy-Sakin, M. et al. Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nat. Commun.* **10**, 1–14 (2019).
38. Lutgen, D. et al. Linked-read sequencing enables haplotype-resolved resequencing at population scale. *Mol. Ecol. Resour.* **20**, 1311–1322 (2020).
39. Willing, E.-M., Dreyer, C. & van Oosterhout, C. Estimates of genetic differentiation measured by FST do not necessarily require large sample sizes when using many SNP markers. *PLoS ONE* **7**, e42649 (2012).
40. Audano, P. A. et al. Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663–675.e19 (2019).
41. 1000 Genomes Project Consortium A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
42. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
43. Ranallo-Benavidez, T. R. et al. Optimized sample selection for cost-efficient long-read population sequencing. *Genome Res.* https://doi.org/10.1101/gr.264879.120 (2021).
This article describes a method for optimized sample selection given an existing variation catalogue.
44. Thibodeau, M. L. et al. Improved structural variant interpretation for hereditary cancer susceptibility using long-read sequencing. *Genet. Med.* **22**, 1892–1897 (2020).
45. Sirén, J. et al. Genotyping common, large structural variations in 5,202 genomes using pangenomes, the Giraffe mapper, and the vg toolkit. *bioRxiv* https://doi.org/10.1101/2020.12.04.412486 (2020).
46. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
47. Yan, S. M. et al. Local adaptation and archaic introgression shape global diversity at human structural variant loci. *bioRxiv* https://doi.org/10.1101/2021.01.26.428314 (2021).
48. Kingan, S. B. et al. A high-quality genome assembly from a single mosquito using PacBio sequencing. *Genes* **10**, 62 (2019).
49. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 129 (2019).
50. Chandak, S., Tatwawadi, T., Sridhar, S. & Weissman, T. Impact of lossy compression of nanopore raw signal data on basecalling and consensus accuracy. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btaa1017 (2020).
51. Holley, G. et al. Ratatosk: hybrid error correction of long reads enables accurate variant calling and assembly. *Genome Biol.* **22**, 28 (2021).
52. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
53. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
54. Lanfear, R., Schalamun, M., Kainer, D., Wang, W. & Schwessinger, B. MinIONQC: fast and simple quality control for MinION sequencing data. *Bioinformatics* **35**, 523–525 (2019).
55. Peona, V., Weissensteiner, M. H. & Suh, A. How complete are 'complete' genome assemblies? An avian perspective. *Mol. Ecol. Resour.* **18**, 1188–1195 (2018).
56. Günther, T. & Nettelblad, C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genet.* **15**, e1008302 (2019).
57. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493 (2011).
58. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
59. Jain, C., Rhie, A., Hansen, N., Koren, S. & Phillippy, A. M. A long read mapping method for highly repetitive reference sequences. *bioRxiv* https://doi.org/10.1101/2020.11.01.363887 (2020).
60. Ren, J. & Chaisson, M. J. P. lra: the long read aligner for sequences and contigs. *bioRxiv* https://doi.org/10.1101/2020.11.15.383273 (2020).
61. Heller, D. & Vingron, M. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35**, 2907–2915 (2019).
62. English, A. C., Salerno, W. J. & Reid, J. G. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics* **15**, 180 (2014).
63. Jiang, T. et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189 (2020).
64. De Coster, W. & Van Broeckhoven, C. Newest methods for detecting structural variations. *Trends Biotechnol.* **37**, 973–982 (2019).
65. Tham, C. Y. et al. NanoVar: accurate characterization of patients' genomic structural variants using low-depth nanopore sequencing. *Genome Biol.* **21**, 56 (2020).
66. Lecompte, L., Peterlongo, P., Lavenier, D. & Lemaitre, C. SVJedi: genotyping structural variations with long reads. *Bioinformatics* **36**, 4568–4575 (2020).
67. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
68. Shafin, K. et al. Haplotype-aware variant calling enables high accuracy in nanopore long-reads using deep neural networks. *bioRxiv* https://doi.org/10.1101/2021.03.04.433952 (2021).
69. Luo, R. et al. Exploring the limit of using a deep neural network on pileup data for germline variant calling. *Nat. Mach. Intell.* **2**, 220–227 (2020).
70. Edge, P. & Bansal, V. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nat. Commun.* **10**, 4660 (2019).
71. Olson, N. D. et al. precisionFDA Truth Challenge V2: calling variants from short- and long-reads in difficult-to-map regions. *bioRxiv* https://doi.org/10.1101/2020.11.13.380741 (2021).

72. Garg, P. et al. A survey of rare epigenetic variation in 23,116 human genomes identifies disease-relevant epivariations and CGG expansions. *Am. J. Hum. Genet.* **107**, 654–669 (2020).
73. Mirkin, S. M. Expandable DNA repeats and human disease. *Nature* **447**, 932 (2007).
74. Chiara, M., Zambelli, F., Picardi, E., Horner, D. S. & Pesole, G. Critical assessment of bioinformatics methods for the characterization of pathological repeat expansions with single-molecule sequencing data. *Brief. Bioinform.* **21**, 1971–1986 (2019).
75. Mitsuhashi, S. et al. Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. *Genome Biol.* **20**, 58 (2019).
76. Bolognini, D., Magi, A., Benes, V., Korbel, J. O. & Rausch, T. TRiCoLOR: tandem repeat profiling using whole-genome long-read sequencing data. *GigaScience* **9**, giaa101 (2020).
77. McLean, J. W. et al. cDNA sequence of human apolipoprotein(a) is homologous to plasminogen. *Nature* **330**, 132–137 (1987).
78. Logsdon, G. A. et al. The structure, function and evolution of a complete human chromosome 8. *Nature* https://doi.org/10.1038/s41586-021-03420-7 (2021).
79. Miga, K. H. et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
This study reports the first assembly of a human chromosome resolved from end to end by leveraging long reads.
80. Schmid, M. et al. Pushing the limits of de novo genome assembly for complex prokaryotic genomes harboring very long, near identical repeats. *Nucleic Acids Res.* **46**, 8953–8965 (2018).
81. Sherman, R. M. et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* **51**, 30–35 (2018).
82. Ameur, A. et al. De novo assembly of two swedish genomes reveals missing segments from the human GRCh38 reference and improves variant calling of population-scale sequencing data. *Genes* **9**, 486 (2018).
83. Asalone, K. C. et al. Regional sequence expansion or collapse in heterozygous genome assemblies. *PLoS Comput. Biol.* **16**, e1008104 (2020).
84. Vollger, M. R. et al. Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019).
85. Heller, D., Vingron, M., Church, G., Li, H. & Garg, S. SDip: a novel graph-based approach to haplotype-aware assembly based structural variant calling in targeted segmental duplications sequencing. *bioRxiv* https://doi.org/10.1101/2020.02.25.964445 (2020).
86. Garg, S. et al. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat. Biotechnol.* **39**, 309–312 (2021).
87. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
88. Nattestad, M. & Schatz, M. C. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**, 3021–3023 (2016).
89. Li, H. et al. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat. Methods* **15**, 595–597 (2018).
90. Armstrong, J. et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020).
91. 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).
92. Chen, X. et al. Spinal muscular atrophy diagnosis and carrier screening from genome sequencing data. *Genet. Med.* **22**, 945–953 (2020).
93. Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Brief. Bioinform.* **19**, 118–135 (2018).
94. Sherman, R. M. & Salzberg, S. L. Pan-genomics in the human genome era. *Nat. Rev. Genet.* **21**, 243–254 (2020).
95. Paten, B., Novak, A. M., Eizenga, J. M. & Garrison, E. Genome graphs and the evolution of genome inference. *Genome Res.* **27**, 665–676 (2017).
96. Hickey, G. et al. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* **21**, 35 (2020).
97. Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* **21**, 265 (2020).
98. Rakocevic, G. et al. Fast and accurate genomic analyses using genome graphs. *Nat. Genet.* **51**, 354–362 (2019).
99. Ebler, J. et al. Pangenome-based genome inference. *bioRxiv* https://doi.org/10.1101/2020.11.11.378133 (2020).
100. Eggertsson, H. P. et al. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* **10**, 5402 (2019).
101. Chen, S. et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol.* **20**, 291 (2019).
102. Garrison, E. et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
103. Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J. & Edwards, D. Plant pan-genomes are the new reference. *Nat. Plants* **6**, 914–920 (2020).
104. Korbel, J. O. et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
105. Belyeu, J. R. et al. SV-plaudit: a cloud-based framework for manually curating thousands of structural variants. *Gigascience* **7**, giy064 (2018).
106. Charlesworth, B. Measures of divergence between populations and the effect of forces that reduce variability. *Mol. Biol. Evol.* **15**, 538–543 (1998).
107. McKenna, A. et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
108. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv*, doi:arxiv.org/abs/1207.3907 (2012).
109. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43**, 11.10.1–11.10.33 (2013).
110. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).
111. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
112. Cleary, J. G. et al. Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. *bioRxiv* https://doi.org/10.1101/023754 (2015).
113. 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).
114. Chander, V., Gibbs, R. A. & Sedlazeck, F. J. Evaluation of computational genotyping of structural variation for clinical diagnoses. *Gigascience* **8**, giz110 (2019).
115. Motoo Kimura, T. O. The average number of generations until fixation of a mutant gene in a finite population. *Genetics* **61**, 763 (1969).
116. Chen, B., Cole, J. W. & Grond-Ginsbach, C. Departure from Hardy Weinberg equilibrium and genotyping error. *Front. Genet.* **8**, 167 (2017).
117. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
118. Han, L. et al. Functional annotation of rare structural variation in the human brain. *bioRxiv* https://doi.org/10.1101/711754 (2019).
119. Sharo, A. G., Hu, Z. & Brenner, S. E. StrVCTVRE: a supervised learning method to predict the pathogenicity of human structural variants. *bioRxiv* https://doi.org/10.1101/2020.05.15.097048 (2020).
120. Geoffroy, V. et al. AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics* **34**, 3572–3574 (2018).
121. Gouil, Q. & Keniry, A. Latest techniques to study DNA methylation. *Essays Biochem.* **63**, 639–648 (2019).
122. Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
123. Lee, I. et al. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nat. Methods* **17**, 1191–1199 (2020).
124. Müller, C. A. et al. Capturing the dynamics of genome replication on individual ultra-long nanopore sequence reads. *Nat. Methods* **16**, 429–436 (2019).
125. Glinos, D. A. et al. Transcriptome variation in human tissues revealed by long-read sequencing. *bioRxiv* https://doi.org/10.1101/2021.01.22.427687 (2021).
126. Asandei, A. et al. Nanopore-based protein sequencing using biopores: current achievements and open challenges. *Small Methods* **4**, 1900595 (2020).
127. Tian, L. et al. Comprehensive characterization of single cell full-length isoforms in human and mouse with long-read sequencing. *bioRxiv* 10.1101/2020.08.10.243543 (2020).
128. Chin, C.-S. & Khalak, A. Human genome assembly in 100 minutes. *bioRxiv* 10.1101/705616 (2019).
129. Kou, Y. et al. Evolutionary genomics of structural variation in Asian rice (*Oryza sativa*) domestication. *Mol. Biol. Evol.* **37**, 3507–3524 (2020).
130. Jiao, W.-B. & Schneeberger, K. Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat. Commun.* **11**, 989 (2020).
131. Chawla, H. S. et al. Long-read sequencing reveals widespread intragenic structural variants in a recent allopolyploid crop plant. *Plant. Biotechnol. J.* **19**, 240–250 (2021).
132. Mitsuhashi, S., Ohori, S., Katoh, K., Frith, M. C. & Matsumoto, N. A pipeline for complete characterization of complex germline rearrangements from long DNA reads. *Genome Med.* **12**, 67 (2020).
133. De Roeck, A. et al. NanoSatellite: accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on PromethION. *Genome Biol.* **20**, 239 (2019).
Nanopore sequencing of patients with Alzheimer disease to investigate an associated variable number of tandem repeats expansion.
134. Song, J.-M. et al. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. *Nat. Plants* **6**, 34–45 (2020).
135. Kim, B. Y. et al. Highly contiguous assemblies of 101 drosophilid genomes. *bioRxiv* https://doi.org/10.1101/2020.12.14.422775 (2020).
136. Pauper, M. et al. Correction: Long-read trio sequencing of individuals with unsolved intellectual disability. *Eur. J. Hum. Genet.* **29**, 637–648 (2021).
137. Quan, C. et al. Characterization of structural variation in Tibetans reveals new evidence of high-altitude adaptation and introgression. *bioRxiv* https://doi.org/10.1101/2020.12.01.401174 (2020).
138. Hu, Y. et al. Genome assembly and population genomic analysis provide insights into the evolution of modern sweet corn. *Nat. Commun.* **12**, 1227 (2021).
139. 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).
140. Minkin, I. & Medvedev, P. Scalable multiple whole-genome alignment and locally collinear block construction with SibeliaZ. *Nat. Commun.* **11**, 1–11 (2020).
141. Rautiainen, M. & Marschall, T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol.* **21**, 253 (2020).
142. Heller, D. & Vingron, M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* **36**, 5519–5521 (2020).
143. Sevim, V. et al. Shotgun metagenome data of a defined mock community using Oxford Nanopore, PacBio and Illumina technologies. *Sci. Data* **6**, 285 (2019).
144. Maghini, D. G., Moss, E. L., Vance, S. E. & Bhatt, A. S. Improved high-molecular-weight DNA extraction, nanopore sequencing and metagenomic assembly from the human gut microbiome. *Nat. Protoc.* **16**, 458–471 (2020).
145. Kolmogorov, M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).
146. Johnson, J. S. et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* **10**, 5029 (2019).
147. Pootakham, W. et al. High resolution profiling of coral-associated bacterial communities using full-length 16S rRNA sequence data from PacBio SMRT sequencing system. *Sci. Rep.* **7**, 2774 (2017).
148. Overholt, W. A. et al. Inclusion of Oxford Nanopore long reads improves all microbial and viral metagenome-assembled genomes from a complex aquifer system. *Environ. Microbiol.* **22**, 4000–4013 (2020).

149. Haro-Moreno, J. M., López-Pérez, M. & Rodrīguez-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 MinION. *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/10.1101/2020.07.27.223495 (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/10.1101/2020.09.18.20195370 (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. Halldorsson, 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/card
pbsv: https://github.com/PacificBiosciences/pbsv
SVanalyzer: https://github.com/nhansen/SVanalyzer