COMPUTATIONAL TOOLS

Piercing the dark matter: bioinformatics of long-range sequencing and mapping

*Fritz J. Sedlazeck[1], Hayan Lee[2], Charlotte A. Darby[3] and Michael C. Schatz[3,4]**

Abstract | Several new genomics technologies have become available that offer long-read sequencing or long-range mapping with higher throughput and higher resolution analysis than ever before. These long-range technologies are rapidly advancing the field with improved reference genomes, more comprehensive variant identification and more complete views of transcriptomes and epigenomes. However, they also require new bioinformatics approaches to take full advantage of their unique characteristics while overcoming their complex errors and modalities. Here, we discuss several of the most important applications of the new technologies, focusing on both the currently available bioinformatics tools and opportunities for future research.

Mate pairs
A molecular technique to generate a pair of sequencing reads separated by an approximately known distance. The typical separation distance for mate pairs is a few kilobases, as opposed to paired-end sequencing, which separates the reads by a few hundred bases at most.

Optical mapping
A microscopy technique used to visualize the characteristics of DNA, especially the physical lengths or the position of fluorescent probes.

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

[2]*Department of Genetics, Stanford University, Stanford, CA, USA.*

[3]*Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA.*

[4]*Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA.*

**e-mail: mschatz@cs.jhu.edu*

https://doi.org/10.1038/s41576-018-0003-4

The development of first-generation sequencing technology in the late 1980s and 1990s was crucial for sequencing the first microbial, plant and animal genomes, including the initial sequencing of the human genome. The most important technology of this generation was automated Sanger sequencing instruments that could sequence hundreds of DNA molecules at a time. Several supporting biotechnologies were also developed for these early projects, including mate pairs, bacterial artificial chromosomes (BACs), optical mapping and other assays, to augment the relatively limited sequences that could be produced. In the mid-to-late 2000s, high-throughput second-generation sequencing technology quickly replaced first-generation sequencing owing in large part to the substantially decreased costs needed for whole-genome sequencing[1]. High-throughput short-read sequencing was the major development of this generation and was supplemented by several related biotechnologies, such as paired-end sequencing, pooled fosmids and improved optical mapping technology. These second-generation technologies have enabled the sequencing of many new genomes and widespread resequencing efforts to analyse genomic diversity[2] and pathogenic variants[3], as well as extensive studies of transcription, gene regulation and epigenetics in many species[4,5]. However, although second-generation sequencing has enabled population-scale analysis of many plant and animal species, it also has important limitations, especially poor or ambiguous mapping to repetitive elements, limited ability to span indels or structural variants (SVs) and amplification artefacts during library construction; hence, the limitations of short-read sequencing have left a substantial fraction of most genomes inaccessible and much of their true complexity hidden[6].

Recently, several new genomic sequencing and mapping technologies (TABLE 1) have become available that are increasingly being used to pierce into the remaining genomic dark matter of repetitive sequences, microsatellites and other complex structural variation. These platforms, characterized by long-range single-molecule resolution, have taken on several forms using both new instrumentation and new protocols to improve older technology. One major advance has been the introduction of long-read single-molecule sequencing using single-molecule real time (SMRT) sequencing from Pacific Biosciences (PacBio)[7] or nanopore-based sequencing from Oxford Nanopore Technologies[8]. Unlike second-generation short-read sequencing, which produces reads of a few hundred nucleotides at most, these technologies (sometimes called third-generation sequencing) now routinely produce reads averaging around 10 kb in length, with many over 100 kb and the longest over 1 Mb. Distinct from these true long-read platforms, the Chromium technology from 10X Genomics employs genome partitioning and barcoding to generate linked reads that span tens to hundreds of thousands of bases[9]. As sequencing is ultimately performed on a high-throughput second-generation sequencer (such as Illumina platforms), the linked reads can produce high-quality genomes, including phased genomes, with only a modest cost increase over standard short-read sequencing. Another advance has been to use Hi-C and related chromatin

Table 1 | **Long-range sequencing and mapping platforms**

Platform	General characteristics and costs	Major applications	Bioinformatics challenges
PacBio SMRT sequencing	Single-molecule long reads averaging ~10 kb with some approaching 100 kb; several fold more expensive than short reads	De novo genome assembly, structural variant detection, gene isoform resolution and epigenetic modifications	Raw reads have high error rates dominated by false insertions; requires new alignment and error correction algorithms
Oxford Nanopore sequencing	Single-molecule long reads averaging ~10 kb with some >1 Mb; several fold more expensive than short reads	De novo genome assembly, structural variant detection, gene isoform resolution and epigenetic modifications	Raw reads have high error rates dominated by false deletions and homopolymer errors; requires new alignment and error correction algorithms
10X Genomics Chromium	Linked reads spanning ~100 kb derived from a collection of short-read sequences; moderately more expensive than short reads	De novo genome assembly and scaffolding, phasing, detection of large structural variants (>10 kb) and single-cell gene expression	Sparse sequencing rather than true long reads; more complicated to align, with poorer resolution of locally repetitive sequences
Hi-C-based analysis	Pairs of short reads (<100 bp) formed from crosslinking chromatin interactions; moderately more expensive than short reads	Genome scaffolding and phasing	Sparse sequencing with highly variable genomic distance between pairs (1 kb to 1 Mb or longer)
BioNano Genomics optical mapping	Optical mapping of long DNA molecules (~250 kb or longer) labelled with fluorescent probes; less expensive than short reads	Genome scaffolding and detection of large structural variants (>10 kb)	Limited algorithms to discover high-confidence alignment between an optical map and a sequence assembly

PacBio SMRT, Pacific Biosciences single-molecule real time.

crosslinking protocols to create very long-range mate pair-like data for second-generation sequencing[10–12]. Although the genomic distance between an individual Hi-C read pair is highly variable, the density of pairs spanning a given genomic distance is more predictable. Consequently, these libraries have a remarkable capability for phasing and scaffolding, allowing for nearly entire eukaryotic chromosomes to be resolved from end to end when combined with a high-quality draft assembly[13]. Finally, new optical mapping instruments from BioNano Genomics can rapidly fingerprint megabase segments of a genome, allowing high-quality scaffolding and SV analysis at relatively low cost[14]. See the review by Goodwin et al.[1] for an in-depth discussion of these and related biotechnologies.

As impressive as these biotechnologies are, they cannot by themselves address any questions in genomics without bioinformatics analysis tools tailored for the new data types. Although the analysis needs of these new long-range platforms are similar to those of second-generation sequencing, most tools designed for second-generation sequencing are inadequate for these new data types. Many of the core algorithmic techniques developed for earlier generations remain in use, such as dynamic programming for alignment or sequence graphs for de novo assembly, but an entirely new generation of bioinformatics tools has been created that leverage the unique features and overcome biases of the new sequencing and mapping platforms (TABLE 2). While seemingly mundane, increased read length has been a considerable problem in how aligners, assemblers, variant callers and other bioinformatics tools store or analyse the reads. Even the widely used BAM file format uses a limited number of bits for recording alignment information, and that is insufficient for some of the longest reads now being generated. This has led many programs to crash or corrupt their results, although recent pending proposals have been made on how to fix this within the core libraries (see Related links). Relatedly, software tools whose algorithmic complexity depends on the length of reads or span of the molecule, such as aligning a read or an optical map to a reference genome, have required new algorithmic ideas and compact data structures for estimating or accelerating the analysis. Even more challenging has been the unique error modes and data modalities: the single-molecule sequencing approaches have higher error rates, whereas Hi-C pairs, linked reads and optical mapping only sparsely sample the template molecules and suffer from their own unique biases. Finally, for many bioinformatics applications, the best results have come from combining multiple data types at once, and thus the software packages must be very flexible.

Nevertheless, new bioinformatics tools, combined with new biotechnologies, have already been used to improve our insights into many genomes. One of the most important applications has been to produce highly accurate de novo assemblies of hundreds of microbial, fungal, plant and animal genomes. Indeed, several 'reference' genomes established using older technologies, such as the genomes of maize[15], fruit fly[16], mosquito[12] and many others, have been reassembled with the new technologies to fix errors and improve their resolution. These technologies are also increasingly being applied for resequencing analyses, especially to create detailed maps of structural variation and for phasing variants across essentially entire human chromosomes and other genomes. Notably, the new technologies have been used to fill in many of the gaps in the human reference genome that resisted more than 1 decade of scrutiny[6,17], and studies have used these technologies to find SVs that are hard or impossible to detect using second-generation sequencing[18,19]. Outside of DNA sequencing, these new technologies are increasingly used to study transcriptomes and epigenomes, and thousands of novel isoforms and gene fusions[20] have already been discovered, as well

Indels
A type of DNA sequence variation marked by the insertion or deletion of nucleotides.

Structural variants
(SVs). DNA sequence variants that are 50 bp or larger, including insertions, deletions, inversions, duplications and translocations.

Linked reads
Also known as a read cloud. A set of barcoded short reads derived from the same DNA molecule and therefore highly localized in the genome.

Phased
Grouping together variants located on the same molecule, such as to identify variants from the maternal or the paternal genome in a diploid sample.

Scaffolding
The process of assembling sequences of DNA into a scaffold. A scaffold is similar to a contig but may contain gaps, typically represented as Ns in the sequence.

Table 2 | **Bioinformatics methods for long-range sequencing and mapping**

Bioinformatics analysis	Approach	Selected methods	URL	Refs
De novo genome assembly				
Hybrid error correction	Using short reads or short-read assemblies to correct raw long reads	Nanocorr (I and O)	https://github.com/jgurtowski/nanocorr	33
		MaSuRCA (I and P)	http://www.genome.umd.edu/masurca.html	35
		PBcR (I and P) (also for just long reads)	http://www.cbcb.umd.edu/software/PBcR/	32
		Spades (I, P and O)	http://bioinf.spbau.ru/spades	34
Self error correction	Correct raw long reads using other raw long reads	FALCON-sense (P)	https://github.com/PacificBiosciences/FALCON	30
		pbdagcon (P)	https://github.com/PacificBiosciences/pbdagcon	36
Long-read overlapping	Find pairs of reads that align to each other	MHAP (P and O)	https://github.com/marbl/MHAP	16
		Minimap (P and O)	https://github.com/lh3/minimap	45
		DALIGNER (P and O)	https://github.com/thegenemyers/DALIGNER	43
Contig assembly	Arrange reads that overlap with each other to build a consensus sequence	Canu (P and O)	https://github.com/marbl/canu	29
		FALCON (P)	https://github.com/PacificBiosciences/FALCON	29
		Hinge (P)	https://github.com/HingeAssembler/HINGE	135
		MECAT (P and O)	https://github.com/xiaochuanle/MECAT	136
		Miniasm (O and P)	https://github.com/lh3/miniasm	45
		Spades (I, P and O)	http://cab.spbu.ru/software/spades/	34
		Supernova (G)	https://support.10xgenomics.com/de-novo-assembly/software/overview/welcome	52
		HGAP (P)	https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/HGAP	36
		Flye (P)	https://github.com/fenderglass/Flye	137
		MARVEL (P)	https://github.com/schloi/MARVEL	37
Scaffolding	Order and orient contigs into chromosome sequences	Architect (I)	https://github.com/kuleshov/architect	53
		ARCS (G)	https://github.com/bcgsc/arcs	54
		BioNano Access (B)	https://bionanogenomics.com/support-page/bionano-access/	
		FragScaff (H)	https://sourceforge.net/projects/fragscaff/	55
		LINKS (P and O)	https://github.com/warrenlr/LINKS	138
		npScarf (O)	https://github.com/mdcao/npScarf	139
		RAILS (P and O)	https://github.com/bcgsc/RAILS	59
		SALSA (H)	https://github.com/machinegun/SALSA	56
Gap filling	Localized alignment and assembly to improve an existing assembly	PBJelly (P)	https://sourceforge.net/p/pb-jelly	58
		RAILS and Cobbler (P and O)	https://github.com/bcgsc/RAILS	59
Polishing	Refine the consensus sequence of a de novo assembly by a re-analysis of how the raw reads align	Arrow (A and P)	http://www.pacb.com/	
		Nanopolish (A and O)	https://github.com/jts/nanopolish	22
		Pilon (A and I)	https://github.com/broadinstitute/pilon	47
		Quiver (A and P)	https://github.com/PacificBiosciences/GenomicConsensus	36
		Racon (A, P and O)	https://github.com/isovic/racon	140
Variant detection				
Assembly alignment	Alignment of assembly to a reference or another assembly	LAST (A and P)	http://last.cbrc.jp/	80
		MUMmer (A)	http://mummer.sourceforge.net/	79
Assembly-based SV detection	Scan alignments to find differences relative to a reference genome or another assembly	AsmVar (A)	https://github.com/bioinformatics-centre/AsmVar	66
		Assemblytics (A)	http://assemblytics.com/	81

Table 2 (cont.) | **Bioinformatics methods for long-range sequencing and mapping**

Bioinformatics analysis	Approach	Selected methods	URL	Refs
Long-read mapping	Seed-and-extend method using clusters of short seeds	BLASR (P)	https://github.com/PacificBiosciences/blasr	75
		BWA-MEM (O and P)	http://bio-bwa.sourceforge.net/	76
		DALIGNER (P)	https://github.com/thegenemyers/DALIGNER	43
		GraphMap (O and P)	https://github.com/isovic/graphmap	141
		Kart (O and P)	https://github.com/hsinnan75/Kart	142
		LASMSA (O and P)	https://github.com/hitbc/LAMSA	143
		LAST (O and P)	http://last.cbrc.jp/	80
		Minimap2 (O and P)	https://github.com/lh3/minimap	77
		NGMLR (O and P)	https://github.com/philres/nextgenmap-lr	73
Linked-read mapping	Two-pass strategy to resolve ambiguous mappings over barcode information	Lariat (G) (also included in LongRanger (G))	https://github.com/10XGenomics/lariat	9
SV calling	Investigating split-read and within-read alignments and coverage	BioNano Access (B)	https://bionanogenomics.com/support-page/bionano-access/	
		GROC-SVs (G)	https://github.com/grocsvs/grocsvs	19
		HiCup (H)	https://www.bioinformatics.babraham.ac.uk/projects/hicup/	74
		LongRanger (G)	https://support.10xgenomics.com/genome-exome/software/pipelines/latest/what-is-long-ranger	
		NAIBR (G)	https://github.com/raphael-group/NAIBR	144
		PBHoney (P)	https://sourceforge.net/projects/pb-jelly/	71
		SMRT-SV (P)	https://github.com/EichlerLab/pacbio_variant_caller	6
		Sniffles (O and P)	https://github.com/fritzsedlazeck/Sniffles	73
SV consolidation	Combines results from multiple callers to increase sensitivity and reduce false-positive calls	MetaSV	https://github.com/bioinform/metasv	82
		SURVIVOR	https://github.com/fritzsedlazeck/SURVIVOR	64
Variant phasing				
Assembly-based	Recognize and partition heterozygous variants during a de novo assembly	FALCON-Unzip (P)	https://github.com/PacificBiosciences/FALCON	30
		Supernova (G)	https://support.10xgenomics.com/de-novo-assembly/software/overview/welcome	52
Mapping-based	Partition the aligned reads into two sets such that the reads within a set strongly agree with each other	HapCut2 (I, P, O and H)	https://github.com/vibansal/HapCUT2	13
		LongRanger (G)	https://support.10xgenomics.com/genome-exome/software/pipelines/latest/what-is-long-ranger	
		WhatsHap (I, O and P)	https://whatshap.readthedocs.io/en/latest/	145
RNA-seq analysis				
Quality control	Detection of artefacts in isoform identification	SQANTI (O and P)	https://bitbucket.org/ConesaLab/sqanti	106
Isoform analysis	Split-read alignments or de novo assembly to find isoforms	TAPIS (P)	https://bitbucket.org/comp_bio/tapis	100
		ToFU (P)	https://github.com/PacificBiosciences/IsoSeq_SA3nUP	105
		BLAT (O)	http://genome.ucsc.edu/FAQ/FAQblat#blat3	146
		Gmap (P)	http://research-pub.gene.com/gmap	147
Single-cell quantification	Counts the number of barcoded reads aligned to each gene	CellRanger (G)	https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger	109
Methylation analysis				
Methylation analysis	Re-analyse raw signal data to find pauses or intensity changes	BaseMods (P)	https://github.com/PacificBiosciences/Bioinformatics-Training/tree/master/basemods	112
		Nanopolish (O)	https://github.com/jts/nanopolish	22
		SignalAlign (O)	https://github.com/ArtRand/signalAlign	21

A, assembly-based; B, BioNano Genomics; G, 10X Genomics; H, Hi-C; I, Illumina; O, Oxford Nanopore; P, PacBio; RNA-seq, RNA sequencing; SV, structural variant.

as more detailed maps of DNA methylation in pathogens and human disease[21,22].

In this article, we discuss the most widely used bioinformatics approaches for analysing these new technologies to address several problems in genomics. We highlight both the strengths and weaknesses of the tools for analysing these new data types and conclude with a discussion of future needs.

De novo genome assembly

Analogous to solving a jigsaw puzzle, a genome is assembled by comparing the sequences of the reads to each other so that overlapping reads can be locked together into contigs[23] (FIG. 1). High-quality assemblies are characterized by high contiguity (typically measured by average or N50 contig size), high completeness (typically measured by the fraction of the genes or fraction of the genome represented) and correctness (typically measured by both base-level accuracy and structural accuracy)[24,25]. A high-quality assembly can be transformative to studying a species: the genome can be annotated to identify genes, regulatory sequences and other important features; it can be used for mapping resequencing data or functional data sets; *cis*-regulation and *trans*-regulation relationships can be more easily identified; topologically associating domains (TADs), synteny blocks and other large chromosome features can be studied; and many other analyses can take place that would be otherwise difficult or impossible.

Producing a high-quality assembly is challenging for many biological and technical reasons, especially repetitive or heterozygous sequences, sequencing errors, chimeric reads, insufficient read length or insufficient or biased coverage[26]. Of these factors, one of the most prominent and challenging is repetitive sequences. Much like how repeated elements in a jigsaw puzzle are the hardest to resolve, such as the blue sky in a landscape, repeats are the hardest sequences to assemble. This is because unresolved repeats confuse how the assembled sequence should be joined together and cause the assembler to end the contig. Consequently, when using reads shorter than the common repeats in the genome, the assembly will be 'shattered' into many small contigs with a small N50 size. In more adverse conditions, repeats can cause misassembly errors where what should be distant regions of the genome are incorrectly assembled together[25]. Consequently, de novo assemblies composed of only second-generation sequencing can lack or misrepresent large portions of the genome, may be fragmented and missing important genes and can lack sufficient robustness to study overall chromosome architecture. In some cases, the assembled sequences have been substantially shorter than the average gene size, rendering the assembled sequence much less useful than earlier reference genomes[27,28].

Assembly algorithm development for long-range sequencing and mapping has been a very active research field. These efforts have begun a resurgence of reference-quality genomes with contig sizes measured in megabases instead of mere kilobases that are common for second-generation attempts[16,29–31]. This development is chiefly due to the new bioinformatics software that

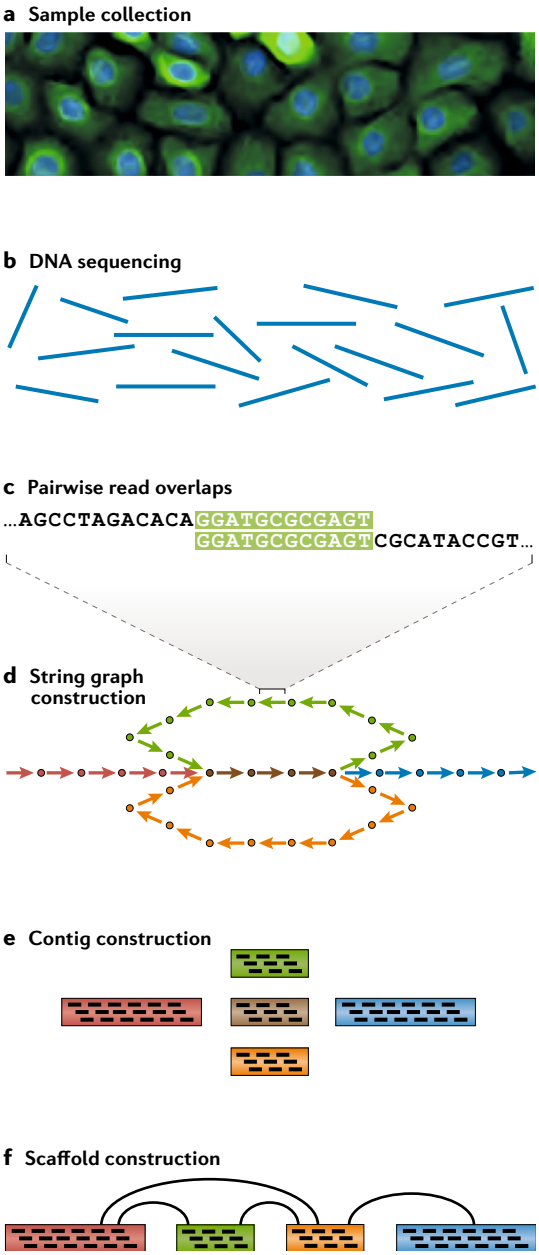

Fig. 1 | **De novo genome assembly. a** | A sample of cells are collected from the individual. **b** | The DNA is extracted and sheared into a library of molecules for sequencing. **c** | The read sequences are compared to each other to find overlaps where the end of one read matches the beginning of another. **d** | An overlap graph or string graph is formed between reads (nodes shown as circles) that overlap (edges shown as arrows between the circles). **e** | The graph is compacted to form the initial contigs ending at repetitive sequences, heterozygous bases or other complicated regions. **f** | The contigs are ordered and oriented using other long-range information (such as linked reads, Hi-C or optical maps) into a scaffold, although some portions may remain as unresolved 'N' sequences. Here, the brown repetitive sequence is left out of the scaffold, leaving gaps between the other contigs. Figure is adapted from REF.[132], BioMed Central, CC-BY-4.0 (https://creativecommons.org/licenses/by/4.0/).

Contigs
Contiguously assembled sequences of DNA.

N50
A weighted average length; specifically, the N50 length is the length such that 50% of the genome has been assembled into contig or scaffold sequences of this length or longer.

***Cis*-regulation**
Any molecular interaction that regulates the transcription of nearby genes on the same DNA molecule, such as the role of a gene promoter.

Trans-regulation
Any molecular interaction that regulates the transcription of genes on a different DNA molecule, such as a transcription factor regulating both alleles of a target gene or genes.

Topologically associating domains
(TADs). Regions of the genome that are enriched for interactions with other elements within the same domain.

Synteny blocks
Genomic regions that are conserved among multiple species.

can effectively use these technologies to span proportionally more of the repeats and heterozygous sequences that are present in a genome. Much of the algorithmic research has been focused around long reads, especially to overcome the high error rates, although linked reads, chromatin-based sequencing data and optical map data can also be assembled, especially to improve the scaffolding of the assembly (see below).

Because of their high frequency of errors and increased read lengths, both PacBio and Oxford Nanopore sequencing require specialized assembly algorithms, including new methods for error correction, overlapping, contig formation and polishing the assembled sequences. Hybrid error correction methods, such as PBcR[32], Nanocorr[33], Spades[34] and MaSuRCA[35], use short-read data to error correct the long reads before assembly and are especially effective when a limited amount of long-read coverage is available (typically below ~30× coverage)[32,33]. Alternatively, self-correction approaches used by HGAP[36], PBcR[32], Canu[28], MARVEL[37] or FALCON[29] error correct the long reads by aligning them to each other and are often beneficial, owing to the more reliable alignments between the long reads if enough coverage is available. With enough coverage available, self-correction often leads to higher accuracy of the error-corrected long reads and less fragmentation of the long-read sequences than when using short reads. This is because the alignment of short reads to (uncorrected) long reads can be unreliable, especially within repetitive sequences or because of biases in short-read coverage (for example, GC content). Deep coverage is also useful because the read-length distribution for long reads is often log normal, so increasing the overall coverage increases the availability of the longest reads. The ultra-long reads are the most useful for resolving repeats or heterozygosity, and their use improves contig sizes and assembly quality the most.

One of the most computationally expensive phases of genome assembly is during overlapping, where the sequence of every read must be compared with the sequence of every other read, using an $O(n^2)$ algorithm if performed naively. This quickly becomes a computationally expensive operation for large genomes with millions to billions of sequencing reads, so various k-mer techniques and other seeding techniques have been developed to identify candidate pairs of reads that are likely to overlap with high similarity. However, the techniques developed for Sanger or Illumina sequencing are generally insufficient for overlapping and assembling long reads with high error rates, as the k-mers that can be used must be very short to avoid sequencing errors. Addressing the poor runtime, Canu uses the MHAP overlap algorithm[16] to quickly overlap raw PacBio or Oxford Nanopore reads. This work replaces the exact, but slow, dynamic programming algorithm for computing the alignment between a pair of reads with a much faster approximation algorithm based on locality sensitive hashing[38]. With it, the sequence similarity between a pair of reads is estimated on the basis of the percentage of specifically chosen k-mers shared between the read sequences. This algorithm, originally developed for identifying highly similar webpages across the entire Internet, scales to large numbers of long reads and improved the runtime by several orders of magnitude. See Chu et al.[39] for a more detailed discussion of read overlap algorithms.

Afterwards, within Canu, the error-corrected long reads are assembled, building on the design of the Celera Assembler[40–42], in which the overlaps between the reads are encoded in an overlap graph containing nodes and edges to represent the relationships between the overlapping reads. A series of graph transformations are applied to the graph, including removing contained reads and any transitive overlaps, to identify and reconstruct the sequences of the genome that can be unambiguously assembled. With reads averaging around 10 kb, the initial contig sequences can span millions of nucleotides and typically end because of ambiguity at repetitive sequences, unresolved heterozygosity, lack of coverage or unresolved sequencing error.

The FALCON assembler and the new MARVEL assembler for PacBio reads operate similarly to Canu, although they both use DALIGNER[43] instead of MHAP to compute the overlaps between reads for error correction and assembly, and form a string graph[44] rather than an overlap graph. Like the original Celera Assembler, DALIGNER uses dynamic programming to compute the overlaps, although DALIGNER uses a heavily optimized banded analysis and k-mer screening to accelerate the computation. Unique to FALCON, it also includes a module called FALCON-Unzip that runs after the initial contigs are assembled to create a phased assembly for diploid samples. With it, heterozygous variants contained within the maternal or paternal chromosomes are separated, and the sequences for the homologous chromosomes are individually reported (also see phasing below)[30]. Another active area of research has been to develop methods to assemble a genome from long reads without any error correction. One such assembler, Miniasm and the associated Minimap overlap algorithm[45], is more than one order of magnitude faster than Canu or FALCON, although the resulting sequence accuracy is poor, making a final polishing step essential.

After assembling the contigs from long reads, polishing algorithms correct residual errors in the assembled consensus. These include Quiver[36] or Arrow for PacBio (see Related links), Nanopolish[46] for Oxford Nanopore or Pilon for polishing long-read assemblies with additional Illumina reads[47]. Unlike base calling, which processes one read at a time, or pre-assembly error correction, which uses a rudimentary alignment of reads, post-assembly error correction is much more effective because it can evaluate the alignments and raw signal data of all the reads that confidently align to a given region at the same time, an idea originally developed for Sanger sequencing[48]. This makes it possible to assemble highly contiguous sequences with 99.9% accuracy or greater, even though the initial sequencing reads may have 20% sequencing errors or worse. The remaining residual errors are enriched in homopolymer sequences and other repetitive sequences that are difficult to align[30]. Hybrid polishing using short-read data is effective over much of the genome, although it is limited in repetitive sequences, where short reads cannot be confidently aligned[31].

Additional improvements are also needed to develop polishing methods for diploid or higher ploidy genomes, where true heterozygous variants are obscured by sequencing errors. Finally, research is also needed to develop more accurate base calling software to minimize the errors in the reads before any assembly. Substantial improvements have already been made for Oxford Nanopore base calling in the past few years, and today's leading approaches can achieve >90% average accuracy using recurrent neural net approaches[49] and other statistical learning techniques[50]. Homopolymer sequences and other low-complexity sequences remain challenging to accurately sequence[31], but improved base calling software and further improvements to the reagents and sensors are expected to further improve the raw accuracy rate to 95% or higher within the next few years.

Chromosome scaffolding and gap filling

Scaffolds, like contigs, represent assembled portions of a chromosome. They are formed by ordering and orienting contigs along the chromosomes using genetic markers, optical maps, linked reads or other long-range sequencing information (FIG. 1). Scaffolds typically span much longer distances than contigs, including entire chromosomes in some cases. Unlike contigs, scaffold sequences can contain gap characters (Ns) where certain regions remain unresolved with potentially unknown gap sizes[26]. Returning to the jigsaw puzzle analogy, contigs represent the fully resolved portions of the puzzle, whereas scaffolds can have 'holes' where some of the pieces are missing. The major bioinformatics challenge to accurate scaffolding is to determine the correct order and orientation of all the contigs so that the contigs are congruent with the supporting data (long reads, linked reads, mates and/or optical maps). As with contig assembly, repeats and heterozygous SVs are the most challenging regions to resolve, as the supporting data may suggest multiple possible configurations.

Scaffolding algorithms commonly use either 'greedy' approaches that iteratively join together contigs with the strongest linking support or a 'global optimization' that tries to best satisfy all of the linking information at once. Scaffolding using optical map data is relatively straightforward as the optical data can be compared with an in silico restriction map of the sequence contigs[51]. Scaffolders using linked reads, such as Supernova[52] (which also includes an integrated short-read assembler), Architect[53], ARCS[54] or fragScaff (which is also used for Hi-C data)[55], look for pairs of contigs that are bridged by multiple linked reads, that is, reads labelled with multiple barcodes align to a pair of contigs. One complication of this approach is that a barcode used for labelling a long molecule may be reused to label different molecules, creating conflicts in which contigs should be adjacent to each other. Fortunately, this can often be resolved by requiring multiple linked reads (multiple barcodes) to be shared before scaffolding the contigs together. Scaffolding using Hi-C is the most challenging, as the genomic distance between a given Hi-C-based read pair is highly variable and may span a few kilobases to megabases without any direct indication of the true distance[11,12,56]. Fortunately, the density of pairs linking two regions is more predictable, allowing for neighbouring contigs to be more reliably identified as having the most pairs between them, although inversion errors are common and repetitive sequences can artificially inflate the number of pairs.

A combination of sequencing and mapping data often leads to improved assemblies and is potentially more cost effective than sequencing alone. For example, the goat[57] and human[56] genomes were assembled using a combination of long reads and Hi-C-based data have remarkably high quality with long contigs (contig N50 of 18.7 Mb and 26.8 Mb, respectively), chromosome-length scaffolds (scaffold N50 of 87 Mb and 60.0 Mb) and nearly 100% sequence fidelity. Alternatively, by carefully modelling and leveraging the linked reads data characteristics, the 10X Genomics Supernova assembler has also proved very effective using only linked reads, with scaffold lengths among the best available for any human genome (scaffold N50 sizes of 15 Mb to 18 Mb)[52], although the contig sizes remain relatively short (contig N50 sizes of 106 kb to 123 kb). Supernova further produces a phased genome assembly, often with phase block lengths that are considerably longer than those obtained using FALCON-Unzip with PacBio reads.

A final approach is to use low-coverage long reads for scaffolding or filling gaps within an existing assembly using tools such as PBJelly or Cobbler[58,59]. In many cases, gap filling makes it possible to turn sequence scaffolds (with Ns) into contigs (without any Ns). Although the consensus accuracy from low-coverage long reads will be lower in the gap-filled regions, it is often more useful to have a low-fidelity sequence than totally unresolved N characters. The main bioinformatics challenge is to avoid creating new misassemblies in the gaps, especially as the flanking contigs surrounding a gap will frequently end because of complex repeats.

One of the largest obstacles for chromosome scaffolding is obtaining a sufficiently high-quality contig assembly before scaffolding (for example, at least 50 kb to 500 kb contig N50 depending on the scaffolding technology): for BioNano Genomics, this is required so that each contig has several restriction sites to enable the optical map to be confidently aligned, and for 10X Genomics linked reads or Hi-C data, this is needed to detangle the initial linked read or chromatin mate-pair alignment information, respectively. The success of these technologies is also very sensitive to any biases in the data. BioNano Genomics map data are limited by fragile sites[14], and the Dovetail cHiCago protocol was designed to filter out the biological noise of chromatin domains from the desired technical signal of locality[10]. Hi-C-based approaches also often have high rates of inversion errors, and both Hi-C-based and 10X Genomics will also be biased by the limitations of Illumina sequencing, especially reduced coverage in regions with extreme GC content. Errors in the initial contigs can also limit scaffolding, and the newest approaches, such as SALSA[56] or the 3D-DNA pipeline[12], first try to resolve contig errors before scaffolding. Finally, and most importantly, scaffolding a chromosome has less information than fully sequencing a chromosome, and thus important biological sequences can be missed in the gaps between the contigs and the gap sizes may be poorly estimated.

Fragile sites
Regions of the DNA molecule that are prone to physical shearing, especially when multiple nicking sites targeted by a nicking enzyme are located in close proximity.

Structural variation identification and analysis

SVs account for the largest number of diverged base pairs across human genomes[60–63] and have been shown to have substantial impacts on evolution (such as gene losses and transposon activity), genomic disorders (such as cancer and autism), gene regulation (such as gene duplications or rearrangements of transcription factors) and other phenotypes in many species (such as mating and intrinsic reproductive isolation)[64]. SVs are characterized as 50 bp or larger events that fall into one of five major categories: deletions, duplications, insertions, inversions and translocations. Copy number variants (CNVs) are an important subtype of SVs formed by genomic deletions and duplications that have been implicated in several human diseases[65]. SVs can also be nested or chained together into larger regions of variation, especially in cases of chromothripsis or chromoplexy in cancer[66].

Our understanding of SVs has been limited by the technology at hand. The earliest reports, based on cytogenetics and other low-resolution approaches, suggested that SVs were rare and played a minor role in human variation[67]. However, in 2004, Sebat et al. used microarrays to discover that CNVs between human genomes are surprisingly frequent and contribute a large number of variable bases within a person[68]. Today, SV analysis is most commonly based on short reads and routinely discovers copy number alterations in individuals whether healthy or with disease. Nevertheless, detecting SVs from short reads often suffers from low sensitivity (30–70%) and up to 85% false discovery[17,69–72]. Thus, many projects elect to focus on certain subtypes of SVs (for example, just CNVs or just deletions) or opt to not call SVs at all. This is done to reduce the false discovery rate but also risks missing important variations.

Most recently, the new sequencing and mapping technologies have been shown to improve sensitivity and reduce false discovery while inferring all SV types. Using the newest approaches, several studies have reported around 20,000 SVs per human genome, most of which could not be detected using short-read sequencing[6,17,73]. This development has started to widen our knowledge of the emergence, complexity and impact of SVs owing to higher resolution and long-range information. Importantly, the longer reads and longer spans allow detection of SVs within repetitive elements and segmental duplications, which are otherwise inaccessible using short reads. For example, these technologies have enabled the closing of many of the gaps in the human reference genome, and it was found that the majority of gaps (78%) carried long runs of degenerate short tandem repeats embedded within (G + C)-rich genomic regions that are difficult to sequence using short-read sequencing[6]. Similar to de novo assembly, the main bioinformatics challenges centre around accurate alignment of the data, especially to overcome errors and to avoid misalignments induced by repetitive sequences.

Two main approaches have been used for SV discovery: a mapping-based approach or a de novo assembly-based approach. In the first approach, SVs are detected on the basis of the direct mapping of the reads or other data from a sample to a reference genome using methods such as LongRanger (10X Genomics), GROC-SVs (10X Genomics)[19], PBHoney (PacBio)[71], SMRT-SV (PacBio)[6], Sniffles (PacBio and Oxford Nanopore)[73] or HiCup (Hi-C)[74] (FIG. 2). The most successful methods use multiple sources of information to predict the location and type of SV, including the following: within-alignment analysis, to detect indels up to a few kilobases in size; split-read alignments (where two or more regions of a read are aligned to different regions) to detect larger indels and other types of SVs; paired-end or linked-read alignments to indicate SVs based on the abnormal distance or orientation of the segments or read pairs; and alterations in alignment coverage, which can identify CNVs but not rearrangements. Several technology-specific algorithms have been introduced to address the specific error models and characteristics of the data, including BLASR (PacBio)[75], BWA-MEM (PacBio and Oxford Nanopore)[76], Minimap2 (PacBio and Oxford Nanopore)[77], NGMLR (PacBio and Oxford Nanopore)[73] or Lariat (10X Genomics)[9], which is itself based on the Random Field Aligner (RFA)[78] and uses BWA-MEM. Another major challenge is that novel sequences in the sample will not map well or at all in the reference, making it difficult to resolve large insertions. Nevertheless, the main strengths of a mapping approach are that it requires the least amount of coverage (minimally only ~15×), is able to identify heterozygous SVs and is more robust to genomic amplifications, such as highly amplified oncogenes, which tend to assemble poorly[73].

The second main approach for detecting SVs is de novo assembly followed by whole-genome alignment between the samples or the reference genome. The advantage of this approach is that in principle it can reveal the full genome of each sample, including any sample-specific sequences or large insertions that can be harder to resolve by read mapping. De novo assembled contigs also provide strong evidence that no SVs, especially homozygous SVs, have gone undetected. Commonly used methods to perform the whole-genome alignment include MUMmer[79] or LAST[80], followed by methods such as AsmVar[66] or Assemblytics[81] to scan the alignments for SVs.

The accuracy of detecting SVs within an assembly strongly depends on the quality of the de novo assembly, the quality of the reference and the quality of the alignments. Repetitive sequences are the largest obstacle and can mask or confuse where SVs occur, although longer contigs can generally be more robustly aligned than short contigs. Another important consideration is the increased costs required for SV detection based on de novo assembly relative to the direct mapping of reads to a reference, both in the increased sequence coverage needed and the more computationally demanding methods that are applied. Furthermore, although new methods are starting to become available for diploid genome assembly, detecting heterozygous variants or analysis of polyploid regions (for example, in plant or cancer genomes) remains challenging. Heterozygous variants will often be left out of an assembly or will be represented only as alternative contigs. This is often based on the interpretation of the de novo assembler and can lead to artefacts

Nested
Two or more adjacent or even overlapping variants in the same region of the genome, such as a deletion within the middle of a larger inverted sequence.

Chromothripsis
A phenomenon by which many chromosomal rearrangements occur in a single event in a localized region of the genome. Also called chromosome shattering.

Chromoplexy
A complex mutation where genetic material from multiple chromosomes is broken and ligated to each other in a new configuration, especially in cancer.

Polyploid
Cells and organisms that contain more than two paired (homologous) sets of chromosomes.

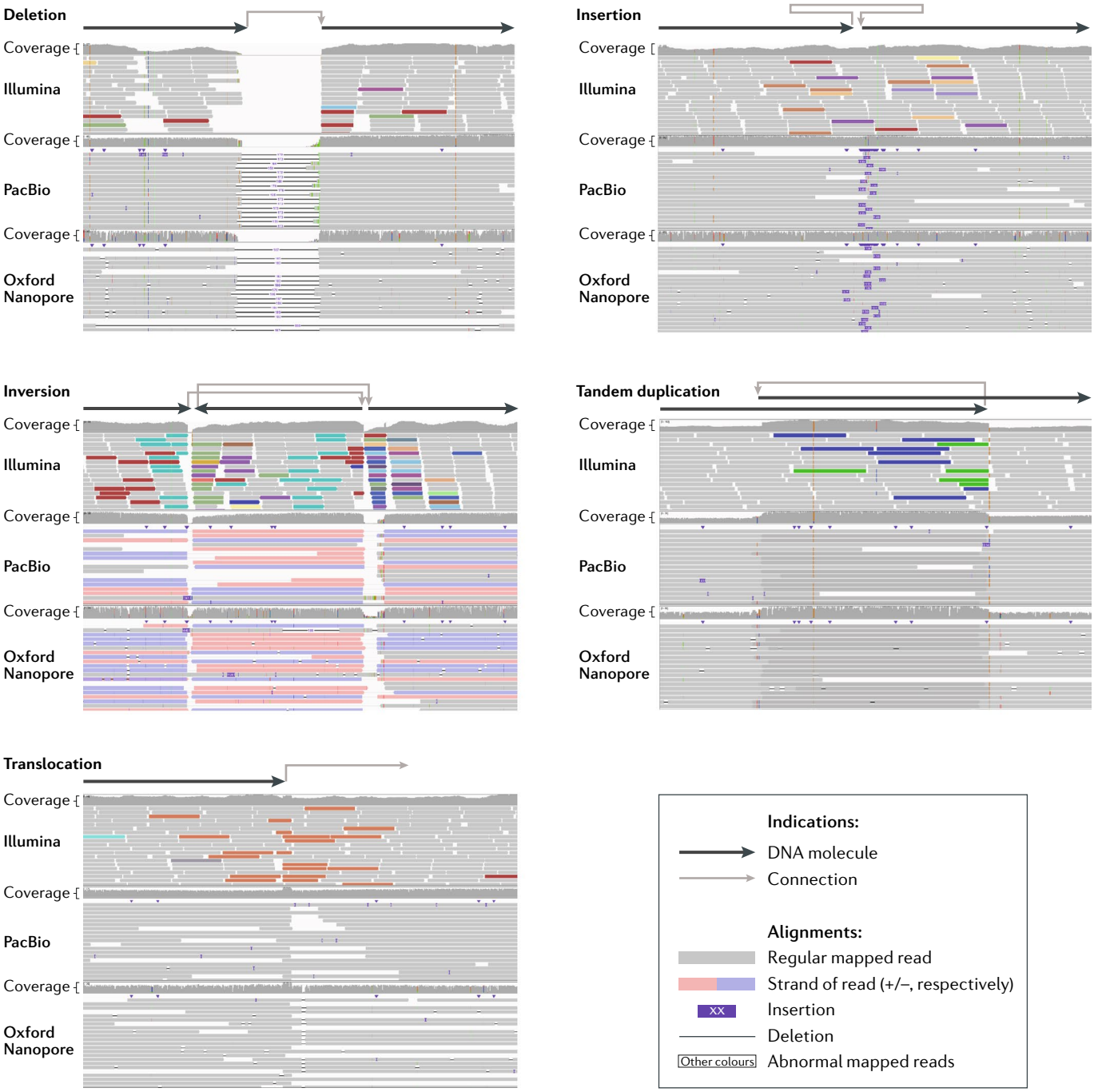

Fig. 2 | **Structural variant detection with long-read sequencing.** Structural variant (SV) types as illustrated in the Integrative Genomics Viewer (IGV)[83] based on Illumina reads (using BWA-MEM), Pacific Biosciences reads (using NGMLR) and Oxford Nanopore reads (using NGMLR) for the NA12878 human genome[73]. The black DNA molecule arrows show the canonical reference genome, and the grey connection arrows show the breakpoints of the SVs. The coloured regions of the reads highlight the discordant pairs or split-read alignments that are informative for identifying the SVs. Deletions are the easiest to identify as regions with reduced alignment coverage. Insertions can be recognized using Illumina data based on paired-end alignments, especially where one read in the pair is not mapped (shown by coloured boxes). Insertions with long reads are more straightforward to identify, as the alignment can extend through the inserted sequence (indicated as a small purple box). Inversions, duplications and translocations are recognized as pairs that have incorrect orientations and/or split-read alignments to different locations in the genome or to different strands (shown as different coloured boxes).

when the algorithm fails to detect the heterozygosity because of complicated repeat structure or insufficient read coverage.

A specialized form of de novo assembly-based SV detection is based on optical mapping data[51]. The patterns of restriction sites identified along each molecule can be de novo assembled into more complete optical maps, analogous to a de novo sequence assembly. SVs in the sample can then be detected by comparing the assembled map of the sample with an in silico restriction map of the reference genome. This can be a relatively inexpensive way to detect SVs across the genome, but

the approach is constrained by the spacing of the restriction sites and a general inability to determine the specific sequence of the SVs. The shorter the distance between the markers is, the higher the precision and resolution of SVs. For example, optical maps produced by BioNano Genomics using a 7 bp recognition site are typically capable of detecting SVs that are 10 kb or larger[14]. However, as the optical maps provide a sparse representation of the genome, currently available algorithms, such as the BioNano Access algorithm, have limited accuracy when two or more SVs are close to each other (both are located within a given pair of digest sites), as the signals will get mixed together so that only their net effect can be measured.

It is important to highlight that each strategy for detecting SVs has unique advantages and disadvantages. There are two main points to consider. First is coverage versus costs: de novo assembly-based methods typically require much higher coverage (~50× or more), whereas mapping-based methods can detect most of the SVs in a sample using lower coverage (~15× coverage), including heterozygous SVs. Consequently, for the same sequencing cost, about three times as many samples can be analysed using a mapping approach than a de novo assembly approach. The second consideration is the complexity of the sample: although haploid genomes are relatively straightforward to analyse with de novo assembly or optical mapping, these approaches are more challenging in regions of high copy number. Conversely, if there isn't a high-quality reference of a closely related species available or the genome has been highly rearranged or mutated compared with the reference, de novo assembly may be necessary to capture the sequences that are specific to the new sample. Finally, recent reports have shown that SVs in the human genome are enriched in segmental duplications and other repetitive elements[6,73]. As these are the most difficult sequences to assemble or align, it is essential that the bioinformatics tools carefully evaluate the evidence to make accurate calls.

There are multiple challenges left to be addressed for the high-quality detection and analysis of SVs. As the different technologies and approaches have different strengths and weaknesses, it is important to filter and combine results across data types and samples. One strategy is to use a consensus approach to improve sensitivity (for example, MetaSV[82]) or limit the false discovery rate (for example, SURVIVOR[64]). It often helps to visualize the SVs to better understand the limitations and accuracy of the approach. The widely used Integrative Genomics Viewer (IGV)[83] program was recently improved to better support long or linked reads, and Ribbon[84] can visualize long reads or assemblies mapped to multiple chromosomes. Another active line of research is to develop hybrid strategies that use read mapping to localize the data and then use de novo assembly to resolve the specific sequence of the SV[85]. Overall, more-sensitive methods are needed to tackle remaining problems arising from sequencing or PCR artefacts, de novo assembly or alignment artefacts or regions that are simply hard to assess owing to their diversity. Finally, unlike single nucleotide polymorphisms (SNPs), which have databases of hundreds of millions of known variants, SVs do not have extensive databases available. This makes it difficult to determine whether an observed SV is common in the general population, which has proved to be one of the most powerful signals for determining whether a variant is pathogenic or carries some other important function[86].

Haplotype phasing and allele-specific analysis

Many eukaryotic genomes, including most higher plant and animal species, have more than one copy of each chromosome. For example, a typical human genome varies from the reference genome at about 4.1 to 5.0 million sites, and most variants are heterozygous, with the density of heterozygous sites across the genome depending on the relatedness and ethnic background of the parents[2]. Distinguishing the maternal versus paternal haplotypes allows the recognition of compound heterozygous versus hemizygous mutations, the analysis of allele-specific expression and occupancy by DNA-binding proteins, the determination of the parent of origin for de novo mutations, the detection of subclones in tumours and other evolutionary and medical applications[87,88].

Most variant callers[89–91] developed for second-generation sequencing report unphased variants (also called genotypes), although a few approaches are available for phasing variants into haplotypes[88]. The most reliable strategy is to sequence or genotype the individual's parents or a larger pedigree and directly determine the parental origin of each variant, except at those positions where both parents share the same variant. However, this potentially increases the cost of a study, and the parents may not be available. Another popular method is to use statistical inference from large collections of genomes to impute the haplotypes[87]. This is very robust for common variants but is less reliable for rare variants and offers no resolution for private mutations or somatic mutations.

Consequently, there is great interest to derive haplotypes using only sequencing reads from one individual. The key concept is that heterozygous variants can be phased when reads span across them (FIG. 3a). Thus, the methods are limited by read length, sequencing errors and fluctuation in coverage that can cause false variants to be introduced or true heterozygous variants to be missed. Minimum error correction (MEC) is a widely used formalism for computing the optimal phasing on the basis of how the reads are aligned to a reference genome[92]. Conceptually, the haplotypes within a diploid chromosome are determined by finding a partitioning of the reads into two sets, one for each haplotype, such that the reads within each partition have a minimal number of errors with respect to a consensus haplotype. Exactly solving this problem is generally computationally intractable as it is NP-hard[82], although WhatsHap[83] provides a dynamic programming exact solution to MEC as well as its weighted extension but can efficiently process only at most 15× coverage. Other heuristic approaches, such as HapCut2 (REF.[13]) and LongRanger[9], adjust which reads are in which partition by optimizing a maximum-likelihood objective. Concomitant with de novo assembly, FALCON-Unzip[30] uses a greedy approach, whereby

Polymorphisms
Variants observed in the genome that are present to some appreciable degree within a population (for example, >1 %).

Compound heterozygous
Two different mutant alleles at a particular gene locus, one on each chromosome of a pair.

Hemizygous mutations
Two or more heterozygous mutations, especially loss-of-function mutations, occurring on the same chromosome so that they disrupt one copy of the gene but leave one functional copy.

Private mutations
Rare variants observed only within a single person or family.

NP-hard
In computational complexity theory, this is a class of problems in which no fast solutions are known to exist.

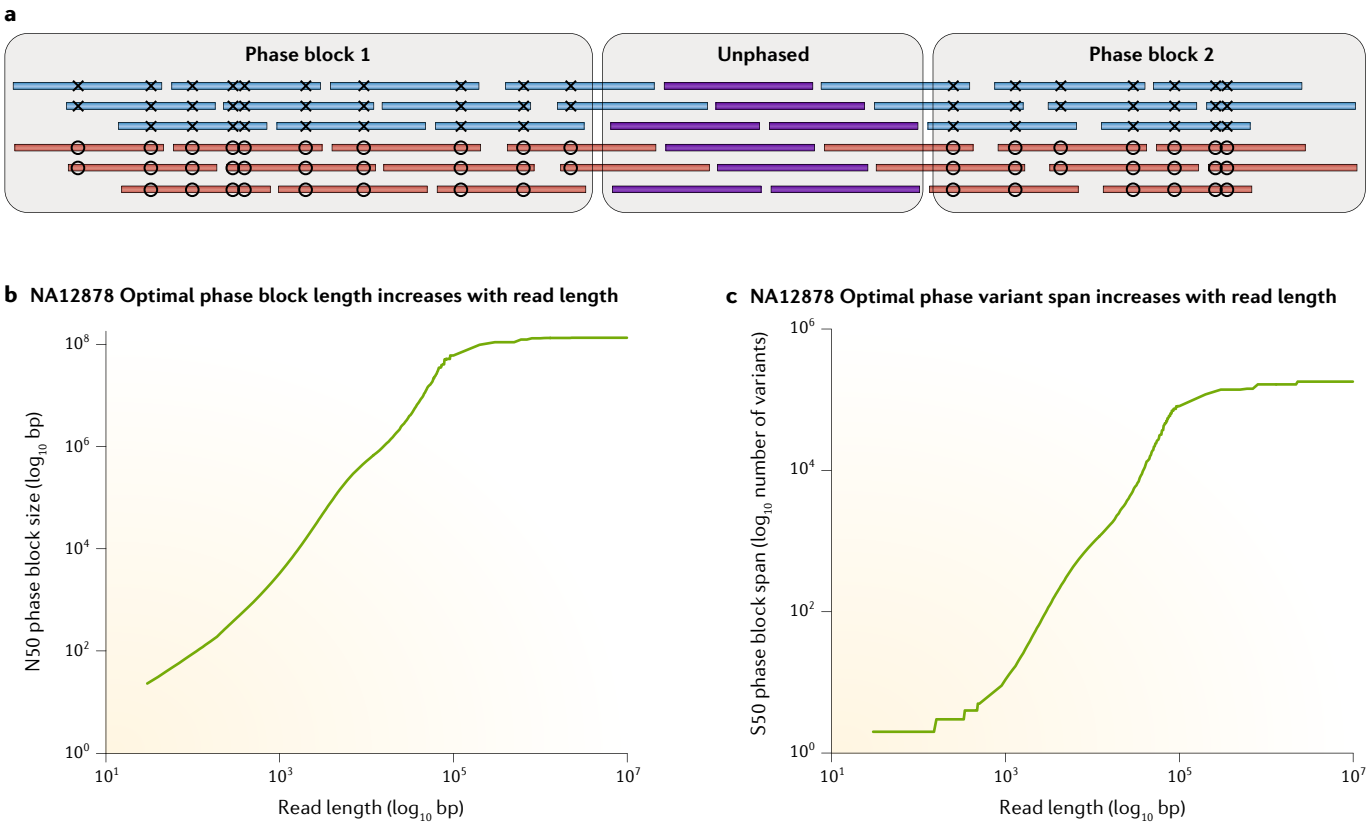

Fig. 3 | **Phasing concepts and requirements. a** | An illustration of using the reads to phase heterozygous variants, with coloured rectangles representing reads, and X and O representing heterozygous alleles. Reads spanning heterozygous variants can be used to phase the reads between the two haplotypes, shown here as red and blue reads. If the heterozygous variants are separated by a distance greater than the read length or span, it will not be possible to phase those variants, creating two phase blocks and an unphased region in this example (purple reads). **b** | Idealized analysis of the maximum phase block N50 length possible by phasing heterozygous single nucleotide polymorphisms (SNPs) in the NA12878 human genome using reads of different lengths. This shows that substantial improvements in phase block length are possible over short-read sequencing using long reads (~10 kb), linked-read sequencing (~100 kb) or Hi-C-based mate pairs (up to 1 Mb or longer). Note: real data containing errors and variable read lengths may not achieve the lengths shown here. **c** | Similar to part **b** but showing the number of variants phased in each phased block as a function of the read length. SNP data are available from REF.[133].

reads are iteratively added to the partition with the most similar consensus haplotype and then used to update the haplotype, resulting in the separation of the individual diploid chromosomes. This process works best when there are several heterozygous SNPs or SVs spanned by a single long read, but the high error rate of PacBio sequencing makes it less reliable for phasing individual pairs of SNPs.

To compare the quality of different haplotype assemblies or haplotype phase blocks, the phase block N50 length statistic is widely used. However, this alone can be misleading, as overlapping long blocks that each contain only a few variants would have a high N50. The metrics S50 and AN50 have been proposed to reflect the contiguity of haplotype assemblies[93]: S50 measures the number of SNPs contained in the block instead of the length in base pairs, and AN50 (adjusted N50) combines the length and the fraction of included heterozygous variants to reflect the phasing quality.

With sufficiently deep coverage available, the phase block size is directly related to the span of each read or data type (FIG. 3b,c). For example, in the human genome, where heterozygous variants occur on average every 1 kb to 1.5 kb, but also includes large spans that are void of heterozygosity, the phase block N50 size derived from short or paired-end reads is typically around 1 kb; with PacBio or Oxford Nanopore long reads, the phase block N50 size improves to around 100 kb to 500 kb; and with 10X Genomics linked reads with an average fragment size of 100 kb, the phase block N50 can extend to 10 Mb or beyond[13]. Hi-C-based approaches for phasing can, in principle, phase over even larger distances, including up to the entire length of the chromosome. However, the main challenge for Hi-C data is that the connections between variants will be sparse, so the genome may be phased into many regions containing few variants. A powerful technique is to combine multiple technologies, such as 10X Genomics or long reads, to establish the initial phase blocks and then combine those phase blocks using Hi-C data into nearly complete phased chromosomes[13].

After creating a phased VCF file of variants, a variety of downstream processing tools are available, including the AlleleSeq pipeline for studying allele-specific expression or binding[94]. Multiple challenges remain, including phasing higher ploidy sequences (for example, plant or

cancer genomes), phasing SVs and providing additional functional analysis of phased genomes.

Isoform resolution and gene quantification

Transcription is one of the most important molecular processes, as it is a major determinant of the repertoire and abundances of RNAs and proteins within each cell. Alternative splicing is a widely used mechanism in eukaryotic organisms to increase the variety of proteins. For example, in *Drosophila melanogaster*, alternative splicing is used to determine sex-specific forms of the gene *dsx*, which is one of the major genes of the sex determination system. Whereas in males, exons 1–3, 5 and 6 are spliced together (that is, skipping of exon 4), females use just exons 1–4 to obtain a protein product that is important for female development[95]. Alternative splicing also plays a major role in human genetics, where it is estimated that 95% of multi-exon genes are alternatively spliced[96]. The complexity of the human transcriptome is astounding, as the average human gene consists of 12 exons and has an average length of about 2,100 bp (REF.[97]). Thus, there are thousands of potential isoforms for an average gene, although the number of annotated isoforms is typically in the dozens to hundreds per gene.

The development of RNA sequencing (RNA-seq) using short reads greatly improved the quantification of gene expression compared with older microarray approaches[98]. Various gene quantification and differential expression methods are now available for use in many species and disease conditions[99]. However, one of the major limitations of these approaches is that owing to their short read lengths, they are fundamentally unable to resolve the structures of the most complex genes or gene families containing many similar isoforms. PacBio[100] (specifically, Iso-Seq cDNA transcript sequencing) and Oxford Nanopore[101,102] long-read sequencing have the potential to dramatically enhance the analysis of alternative splicing. As the read length of these technologies is now commonly over 10 kb, cDNA sequencing or direct RNA sequencing can capture entire transcripts within single reads and thus directly determine the underlying exon combinations. Thus, maybe not surprisingly, current studies using long reads often find thousands of new isoforms but only a few novel genes in the human genome and other well annotated genomes[20,100,103,104] (FIG. 4). For example, a recent maize multi-tissue analysis using PacBio long reads revealed over 100,000 transcripts, most of them novel and tissue-specific[104].

The main bioinformatics methods for studying gene isoforms relying exclusively on long reads are the TAPIS[100] pipeline over multiple rounds of mapping, as well as the ToFU PacBio pipeline[105]. These pipelines include mechanisms to control for different artefacts, such as sequencing errors, that would otherwise obscure the true exon boundaries. For example, ToFU first clusters the RNA-seq reads belonging to the same isoform to obtain an error-corrected isoform assembly that is then aligned to the reference genome. Nevertheless, there have been large discrepancies observed in the number of potential isoforms between different pipelines[106]. Differences in sample preparation are likely to contribute to the discrepancies, although the underlying biases of using long reads for transcriptome analysis are not fully understood. These conflicts can be addressed only with the further development of novel quality control (QC) methods and benchmarking projects. Furthermore, the lower yield of the long-read technologies limits the quantification as well as the assessment of low-expressed isoforms, motivating the development of hybrid approaches, such as isoform detection and prediction (IDP), that use long reads to determine isoform structure and short reads for quantification[107,108]. Finally, although not currently used for isoform resolution, the 10X Genomics Chromium platform[109] can be used for high-throughput gene expression analysis of single cells, and their CellRanger bioinformatics software can be used for analysing and clustering of up to 1 million cells.

Direct sequencing of epigenetic modifications

Long-read sequencing has also improved the analysis of epigenetic modifications, especially the direct detection of methylated nucleotides. Methylation signals are a key concept of genetics[110]. For example, the lack of methylation is widely used by bacteria to detect and cleave invasive phage DNA. In eukaryotes, the role of methylated nucleotides varies across species, cell type and sequence composition, with major roles in the repression of gene expression, regulation of embryo development and the determination of chromatin structure of cells, to name a few. The most commonly studied forms of methylation are 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC) and 6-methyladenine (6 mA), although several additional forms are known. The current standard to study 5mC modifications is bisulfite sequencing using short reads[110]. Despite being a major improvement over older protocols, this approach suffers from multiple biases. Most notably, bisulfite treatment can introduce various coverage and sequencing artefacts, and the short reads become even more difficult to map correctly because most of their sequences consist of a three-letter alphabet as non-methylated cytosines are converted to thymines following bisulfite treatment[111].

Both PacBio and Oxford Nanopore sequencing allow for direct identification of methylation while sequencing native DNA (FIG. 5). The major bioinformatics challenge for these approaches is the relatively subtle shift in signal induced by methylation, requiring powerful statistical techniques to detect real methylation events from technical noise. With PacBio's BaseMods software, methylation is detected by analysing the rate at which the polymerase incorporates nucleotide along the template DNA — the interpulse duration (IPD)[112]. If the current nucleotide is methylated, the polymerase momentarily pauses before incorporating the next nucleotide, leading to a detectable shift in IPD. Analysis of the IPD has been a powerful technique to discover new methyltransferases in microbial genomes[113] and even to detect the presence of 6 mA methylation in *Caenorhabditis elegans*[114]. Nevertheless, the detection of these signals requires deep coverage to robustly identify methylated nucleotides, which has limited the application primarily to microbial genomes. For example, PacBio recommends over 500×

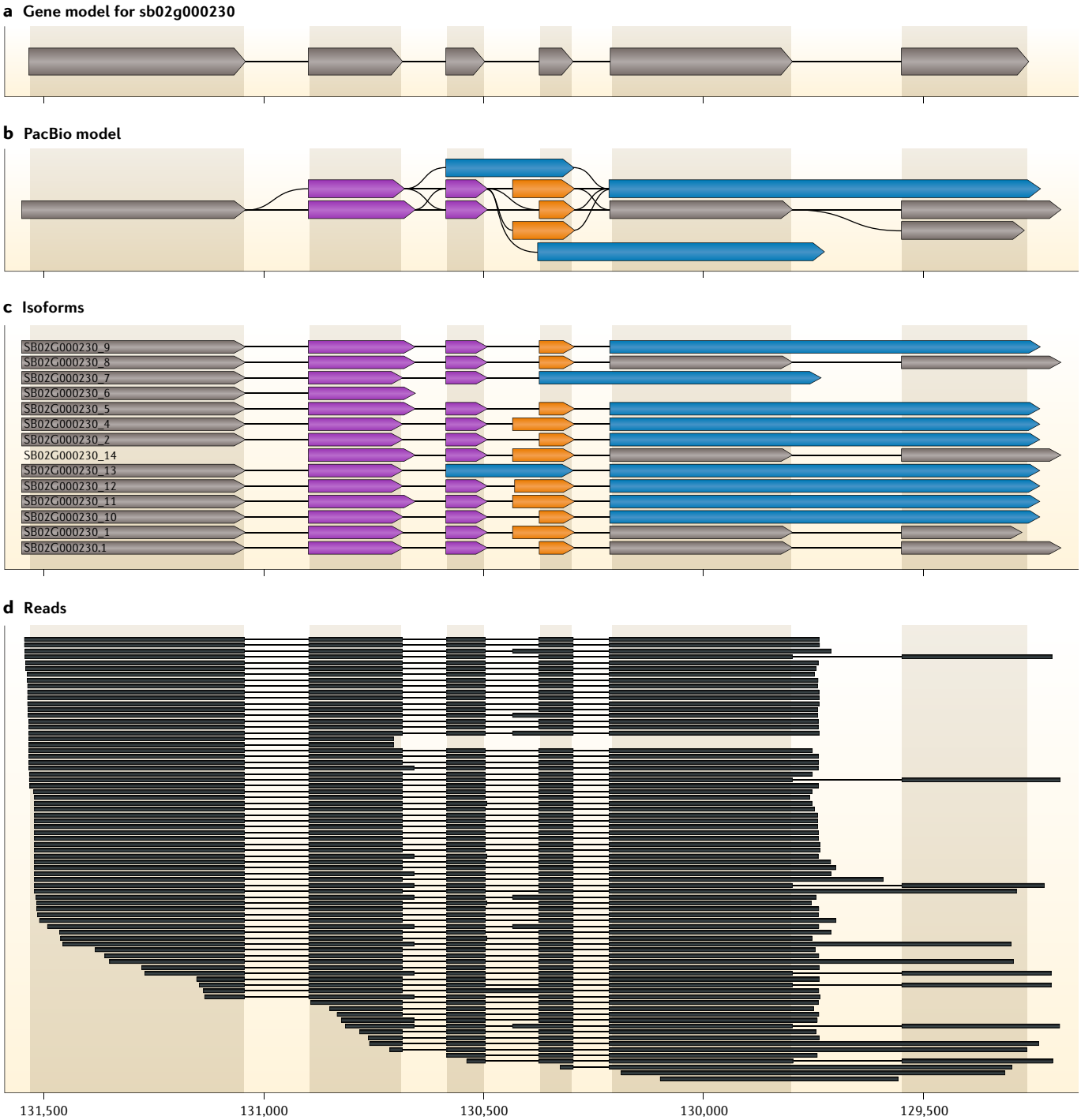

Fig. 4 | **Example of a novel isoforms discovered using long-read sequencing.** An example of a gene sb02g000230 in the cereal *Sorghum bicolor* L. Moench. The overlapping exons and alternative splicing make this gene particularly difficult to resolve with short reads; hence, the previous gene models based on short reads (part **a**) contain a single splice isoform for this gene. By contrast, the Pacific Biosciences (PacBio) long reads allow for 13 novel splice isoforms to be confidently determined. The PacBio-based gene model (part **b**) and individual isoforms (part **c**) are shown, as well as the individual aligned reads (part **d**). The grey arrows represent previously annotated exons, newly identified exons are shown in blue and exons with differing exon boundaries are shown in purple and orange. Figure is adapted from REF.[100], Macmillan Publishers Limited, CC-BY-4.0 (https://creativecommons.org/licenses/by/4.0/).

coverage to detect common 5mC methylation, although other modifications do not require as extreme levels.

More recently, two groups introduced methods, Nanopolish[22] and SignalAlign[21], to analyse methylated bases using Oxford Nanopore sequencing. Similar to PacBio, the methylation is detected in native DNA, although these approaches analyse the electric current of methylated and non-methylated nucleotides, which causes a minute, but detectable, shift in current of a few picoamps. A hidden Markov model (HMM) is then used to distinguish these patterns and can distinguish three cytosine variations

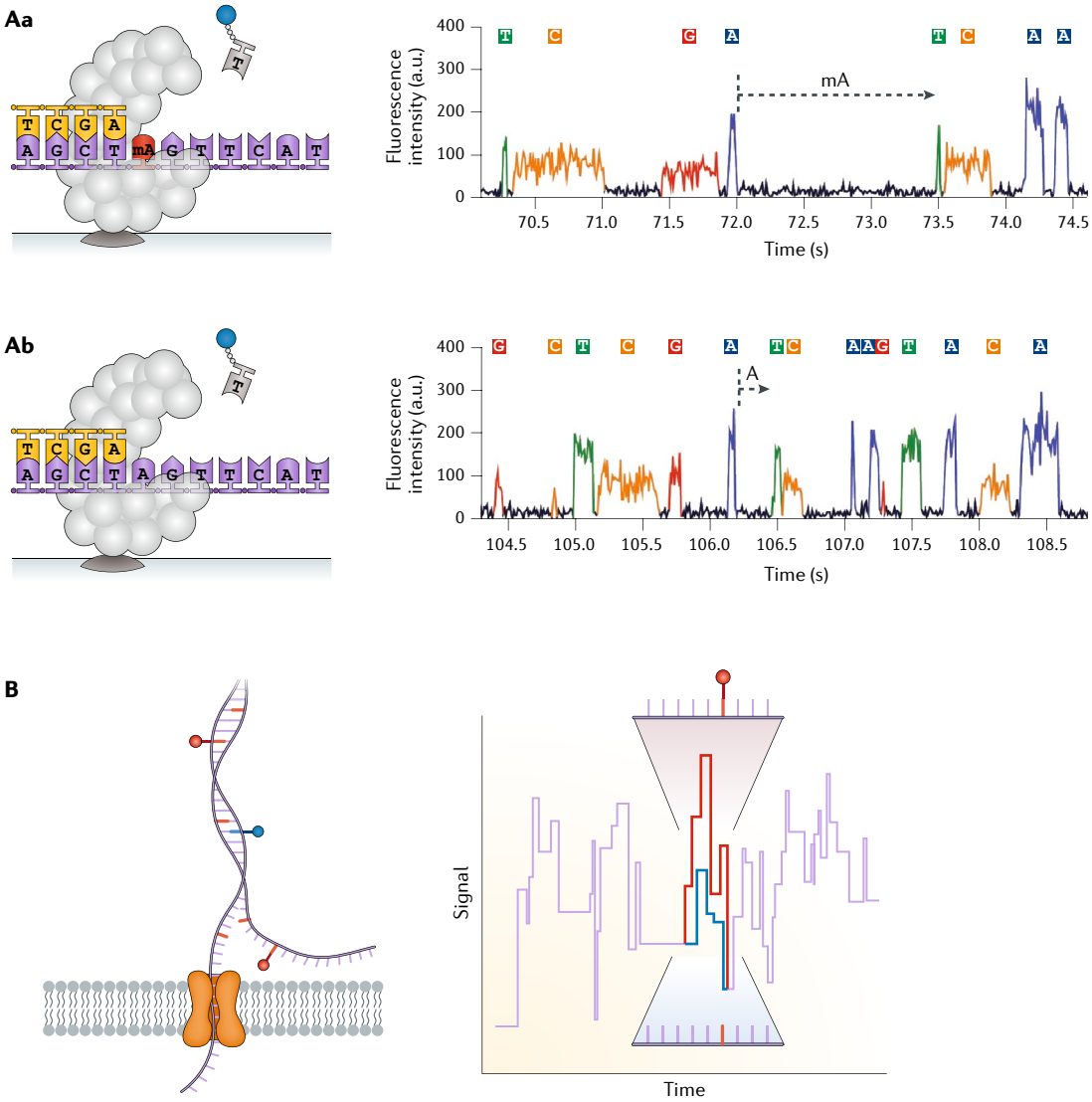

Fig. 5 | **Detecting methylated nucleotides using single-molecule sequencing. A** | Example of Pacific Biosciences (PacBio) sequencing to detect a methylated adenine base (mA), leading to a significantly longer polymerase pause on that base (part **Aa**) compared with a non-methylated adenine (A) (part **Ab**). **B** | Example of Oxford Nanopore sequencing to detect a methylated base (red) by detecting a shift in the raw signal data. Part **A** is adapted from REF.[112], Macmillan Publishers Limited. Part **B** is adapted from REF.[134], Macmillan Publishers Limited.

(C, 5mC and 5hmC) and two adenine variations (A and 6 mA). Although these are clear accomplishments, multiple issues remain, such as collecting accurate training data for the HMM for extending the model to study additional types of methylation. Furthermore, both groups reported the variability in output of the sequencer to be the major confounder of errors, including those caused by confounding environmental conditions, such as the temperature of the room[22].

Overall, both PacBio and Oxford Nanopore can provide novel insights and advantages compared with the current standard bisulfite sequencing. Very recent work has even shown that Oxford Nanopore sequencing can directly read and detect modifications within RNA molecules[115], something that is not possible to study using second-generation sequencing without using complex protocols. However, the field is currently limited in QC methods and scalable analysis methods that will enable routine whole-genome analysis of eukaryotic genomes.

Conclusions and future directions

Emerging long-range sequencing and mapping technologies, coupled with new bioinformatics software, are starting to produce genomes, transcriptomes and epigenomes of remarkable quality. Even for large mammalian or plant genomes, great gains have been made, with results approaching or exceeding those from older, more expensive BAC-by-BAC or fosmid-based assemblies. This is most easily measured by the contig and scaffold sizes that are hundreds to thousands of times longer than corresponding second-generation sequencing assemblies or the multitude of variants or isoforms that are detectable only using the new technologies. These results, with full

chromosomal resolution, are truly reference quality and enable improved analysis of nearly every aspect of a genome: more complete and accurate representations of genes; better determination of clinically relevant variants (BOX 1); improved mapping of regulatory regions and other important genomic elements; and improved phasing of variants for allele-specific analysis, as well as better resolution of the overall chromosome organization. It is also important to highlight that several of these applications, especially de novo assembly, SV detection and phasing, are highly interrelated to each other and must be addressed at the same time to produce the best results.

With this resurgence of quality, the remaining bioinformatics research focuses on cost, accuracy, computational performance, the complexity of the sample and the scale of the analysis. The highest-quality genome assemblies have been achieved with the longest possible reads, aided by the longest possible mapping information[12,15,57], such as a combination of PacBio or Oxford Nanopore sequencing along with 10X, Hi-C or BioNano Genomics data for scaffolding. Interestingly, thanks to advanced bioinformatics approaches, the per nucleotide sequencing error rate of the reads has had relatively little effect on the per nucleotide assembled sequence accuracy, as they can effectively reduce even 30% per nucleotide error to below 1% with sufficient coverage (~30× or greater coverage). However, as long-read sequencing is currently more expensive than short-read sequencing, further research into hybrid methods that can more effectively combine data types is necessary if these methods will ever be able to scale to large population studies. This includes both using a combination of data types for sequencing a single sample, such as inexpensive short-read sequencing to augment more expensive long-read sequencing, and leveraging the relatively few high-quality genomes that have been sequenced using the long-range technologies to improve the analysis of the much larger numbers of genomes sequenced with only short-read technologies. Finally, certain combinations of data, such as using Hi-C data for scaffolding or variant detection, are relatively unexplored compared with more mature topics, such as short-read and long-read error correction and assembly.

The computational costs of using these data are not insignificant, with some recent long-read analyses requiring nearly 1 million central processing unit (CPU) hours for the wheat genome[116] and tens of thousands of CPU hours for the human genome[29]. Additional algorithmic and systems research is clearly needed to

Box 1 | **Clinical applications for long-range sequencing and mapping**

Long-read sequencing and long-range mapping are currently used primarily for research applications, although their superior resolution is appealing for clinical applications as well. One important clinical application has been to improve the characterization of the major histocompatibility complex (MHC) locus among patients and research subjects[123]. This locus encodes many important genes for the immune system, making it a major target of translational research. It is also essential to precisely characterize the sequence between transplant patients to ensure compatibility of the human leukocyte antigen (HLA) genes. However, because of its variability and complexity, second-generation sequencing is generally not sufficient to fully resolve the sequence in a given patient, although several reports have shown that the new biotechnologies allow for complete resolution[124–126].

Another important clinical application has been to detect complex structural variants (SVs) associated with disease that were missed or difficult to detect by earlier technologies. In one example, Pacific Biosciences (PacBio) long-read sequencing was used to detect a 2,184 bp heterozygous deletion of the *PRKAR1A* gene in a patient displaying multiple neoplasia and cardiac myxomata. This SV was determined to be the causal pathogenic variant[18] and could be robustly found only with 10× coverage long-read sequencing, whereas it was more challenging to detect with >30× coverage short-read sequencing. In another study, researchers used PacBio long-read genome and transcriptome sequencing of the SKBR3 breast cancer cell line to discover tens of thousands of SVs that had been missed using short-read approaches[127]. This approach also allowed for a precise characterization of the amplification of the important *ERBB2* (also known as *HER2*) oncogene, revealing a complex series of nested translocations and duplications between chromosomes 17 and 8. A final example has been how the hand-held, low-cost Oxford Nanopore MinION instrument was used for sequencing Ebola[128] and Zika[129] virus isolates in West Africa and South America. These examples showed that rapid genomic characterization was possible in the field and hints towards the possibility of widespread deployment to essentially any hospital, office or school in the world[130].

Despite their many advantages, these technologies have multiple drawbacks for clinical care and are therefore not yet widely used outside of research. One major concern is the requirement for large quantities of high molecular weight DNA to exploit the new technologies. PacBio and Oxford Nanopore are especially challenging, requiring 1,000 times more DNA than second-generation sequencing (~10 micrograms instead of ~10 nanograms) and special handling protocols to limit shearing. The shearing requirement is also a major obstacle for 10X Genomics, BioNano Genomics or Hi-C approaches, especially in solid tissues or solid tumours, where high molecular weight DNA can be challenging to extract.

Furthermore, as highlighted several times throughout this Review, these technologies have their own biases and limitations, but there are relatively few methods for quality control that are necessary to inform clinical decisions. The challenges are compounded by the lack of database systems for interpreting complicated SVs, but these can be developed only through widespread deployment to many patients over many years. Finally, as in a research setting, the increased costs are a major barrier, although a few reports are starting to highlight cases where the increased costs were offset by the improved diagnostic power[18]. These technologies decrease in price every year and have now reached similar levels to other widely used diagnostic equipment (for example, computed tomography (CT) scans). One pragmatic cost-saving approach is the use of capture-based technologies to focus the analysis on the most relevant regions first[126], analogous to how exome capture or targeted panels with second-generation sequencing are commonly used today[131].

Metagenomes
The genomes of all the species present in a sample, studied without culturing or otherwise isolating any individual.

make these analyses faster and more practical. Accuracy improvements have been made possible through advances in base calling and polishing, although new machine-learning techniques, such as advanced graphical models or deep-learning technologies, could be used to further improve sequence accuracy or improve the detection of genomic variants or epigenetic modifications[117]. The new technologies have led to some improvements for assembling metagenomes[118–120], although virtually no bioinformatics tools are currently available for assembling polyploid or aneuploid genomes. Algorithms for phasing polyploid or aneuploid genomes are also in their early stages, despite being very common for plant genomes or in human diseases such as cancer. We also recommend that researchers carefully monitor the developments in the field, as these technologies are all rapidly evolving, and new technologies are already under development.

Finally, as these technologies mature, it is likely that many projects will begin with fully assembled genomes instead of variant lists, opening new opportunities for studying genetic variation across large populations, especially SVs that are difficult to analyse without these technologies. As such, a very active area of research is developing methods to assemble and analyse the pan-genome for a species where the genomes of multiple individuals are represented in one unified graph structure[121,122]. Achieving this will take years of effort to retool and rethink analyses that are now performed with a single linear reference, including downstream aligners, variant callers, epigenetic modification detection tools, visualization tools and related software. Nevertheless, we encourage researchers to focus on these methods as such representations will offer many advantages for studying population genetics for research or clinical needs.

Published online: 29 March 2018

1. 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). **This is a comprehensive Review of all major sequencing and mapping platforms, including a detailed discussion of their relative strengths and weaknesses.**
2. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
3. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
4. The Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
5. Celniker, S. E. et al. Unlocking the secrets of the genome. *Nature* **459**, 927–930 (2009).
6. Chaisson, M. J. et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015). **This is the first major publication describing how PacBio long reads could be used for human genetics, showing that over 20,000 SVs are present in a typical human genome.**
7. Roberts, R. J., Carneiro, M. O. & Schatz, M. C. The advantages of SMRT sequencing. *Genome Biol.* **14**, 405 (2013).
8. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 239 (2016).
9. Zheng, G. X. et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* **34**, 303–311 (2016).
10. Putnam, N. H. et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
11. Burton, J. N. et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
12. Dudchenko, O. et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
13. Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* **27**, 801–812 (2017). **This paper describes the very flexible HapCUT2 phasing algorithm for use with short, long or linked reads, as well as Hi-C-based mate pairs.**
14. Cao, H. et al. Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *Gigascience* **3**, 34 (2014).
15. Jiao, Y. et al. Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524–527 (2017).
16. Berlin, K. et al. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33**, 623–630 (2015).
17. Pendleton, M. et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).
18. Merker, J. D. et al. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet. Med.* https://doi.org/10.1038/gim.2017.86 (2017).
19. Spies, N. et al. Genome-wide reconstruction of complex structural variants using read clouds. *Nat. Methods* **9**, 915–920 (2017).
20. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* **31**, 1009–1014 (2013). **This is one of the first reports describing how long-read sequencing can be used to detect novel isoforms in the human transcriptome.**
21. Rand, A. C. et al. Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods* **14**, 411–413 (2017). **This paper presents one of the first methods able to detect methylation changes directly from Oxford Nanopore long-read sequencing. It can detect three cytosine variants and two adenine variants.**
22. Simpson, J. T. et al. Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017). **This paper presents one of the first methods able to detect 5mC methylation changes directly from Oxford Nanopore long-read sequencing.**
23. Phillippy, A. M. New advances in sequence assembly. *Genome Res* **27**, xi–xiii (2017).
24. Bradnam, K. R. et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2**, 10 (2013).
25. Phillippy, A. M., Schatz, M. C. & Pop, M. Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol.* **9**, R55 (2008).
26. Nagarajan, N. & Pop, M. Sequence assembly demystified. *Nat. Rev. Genet.* **14**, 157–167 (2013).
27. Ling, H. Q. et al. Draft genome of the wheat A-genome progenitor Triticum urartu. *Nature* **496**, 87–90 (2013).
28. Li, R. et al. The sequence and de novo assembly of the giant panda genome. *Nature* **463**, 311–317 (2010).
29. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017). **This study describes Canu, one of the most commonly used long-read assemblers supporting both PacBio and Oxford Nanopore data.**
30. Chin, C. S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016). **This study describes FALCON-Unzip, the first long-read-based assembler reporting phased diploid contigs.**
31. Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* https://doi.org/10.1038/nbt.4060 (2018).
32. Koren, S. et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30**, 693–700 (2012).
33. Goodwin, S. et al. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res.* **25**, 1750–1756 (2015).
34. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
35. Zimin, A. V. et al. The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).
36. Chin, C. S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013). **This study describes HGAP, the first non-hybrid long-read de novo assembler.**
37. Nowoshilow, S. et al. The axolotl genome and the evolution of key tissue formation regulators. *Nature* **554**, 50–55 (2018).
38. Broder, A. in *SEQUENCES '97 Proceedings of the Compression and Complexity of Sequences.* 21 (Washington, DC, 1997).
39. Chu, J., Mohamadi, H., Warren, R. L., Yang, C. & Birol, I. Innovations and challenges in detecting long read overlaps: an evaluation of the state-of-the-art. *Bioinformatics* **33**, 1261–1270 (2017).
40. Myers, E. W. et al. A whole-genome assembly of *Drosophila*. *Science* **287**, 2196–2204 (2000).
41. Venter, J. C. et al. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
42. Miller, J. R. et al. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**, 2818–2824 (2008).
43. Myers, G. Efficient local alignment discovery amongst noisy long reads. *Lect. Notes Bioinf.* **8701**, 52–67 (2014).
44. Myers, E. W. The fragment assembly string graph. *Bioinformatics* **21** (Suppl. 2), ii79–ii85 (2005).
45. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).
46. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12**, 733–735 (2015).
47. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
48. Gajer, P., Schatz, M. & Salzberg, S. L. Automated correction of genome sequence errors. *Nucleic Acids Res.* **32**, 562–569 (2004).
49. Boza, V., Brejova, B. & Vinar, T. DeepNano: deep recurrent neural networks for base calling in MinION nanopore reads. *PLoS ONE* **12**, e0178751 (2017).
50. Teng, H. et al. Chiron: Translating nanopore raw signal directly into nucleotide sequence using deep learning. Preprint at *bioRxiv* https://doi.org/10.1101/179531 (2017).
51. Mendelowitz, L. & Pop, M. Computational methods for optical mapping. *Gigascience* **3**, 33 (2014).
52. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res.* **27**, 757–767 (2017).

This study describes the Supernova assembler for 10X Genomics linked reads, which reports phased diploid genomes.
53. Kuleshov, V., Snyder, M. P. & Batzoglou, S. Genome assembly from synthetic long read clouds. *Bioinformatics* **32**, i216–i224 (2016).
54. Yeo, S., Coombe, L., Chu, J., Warren, R. L. & Birol, I. ARCS: scaffolding genome drafts with linked reads. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btx675 (2017).
55. Adey, A. et al. In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Res.* **24**, 2041–2049 (2014).
56. Ghurye, J., Pop, M., Koren, S., Bickhart, D. & Chin, C. S. Scaffolding of long read assemblies using long range contact information. *BMC Genomics* **18**, 527 (2017).
57. Bickhart, D. M. et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* **49**, 643–650 (2017).
58. English, A. C. et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE* **7**, e47768 (2012).
59. Warren, R. L. RAILS and Cobbler: scaffolding and automated finishing of draft genomes using long DNA sequences. *J. Open Source Software* **1**, 116 (2016).
60. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).
61. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
62. Lupski, J. R. Structural variation mutagenesis of the human genome: Impact on disease and evolution. *Environ. Mol. Mutag.* **56**, 419–436 (2015).
63. Chiang, C. et al. The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).
64. 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).
65. Carvalho, C. M. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* **17**, 224–238 (2016).
66. Moncunill, V. et al. Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat. Biotechnol.* **32**, 1106–1112 (2014).
67. Trask, B. J. Human cytogenetics: 46 chromosomes, 46 years and counting. *Nat. Rev. Genet.* **3**, 769–778 (2002).
68. Sebat, J. et al. Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
69. Huddleston, J. et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–685 (2017).
70. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
71. 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).
72. English, A. C. et al. Assessing structural variation in a personal genome-towards a human reference diploid genome. *BMC Genomics* **16**, 286 (2015).
73. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single molecule sequencing. Preprint at *bioRxiv* https://doi.org/10.1101/169557 (2017).
This study introduces an improved long-read mapping algorithm NGMLR and a comprehensive structural variation detection pipeline Sniffles.
74. Harewood, L. et al. Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome Biol.* **18**, 125 (2017).
75. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
76. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at *arXiv* arXiv:1303.3997 (2013).
77. Li, H. Minimap2: fast pairwise alignment for long nucleotide sequences. Preprint at *arXiv* arXiv:1708.01492 (2017).
This paper introduces the very fast Minimap2 long-read aligner for both PacBio and Oxford Nanopore sequencing.
78. Bishara, A. et al. Read clouds uncover variation in complex regions of the human genome. *Genome Res.* **25**, 1570–1580 (2015).
79. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
80. Kielbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493 (2011).
81. Nattestad, M. & Schatz, M. C. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**, 3021–3023 (2016).
82. Mohiyuddin, M. et al. MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics* **31**, 2741–2744 (2015).
83. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
84. Nattestad, M., Chin, C. S. & Schatz, M. C. Ribbon: visualizing complex genome alignments and structural variation. Preprint at *bioRxiv* https://doi.org/10.1101/082123 (2016).
85. Narzisi, G. et al. Accurate de novo and transmitted indel detection in exome-capture data using microassembly. *Nat. Methods* **11**, 1033–1036 (2014).
86. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
87. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12**, 703–714 (2011).
88. Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J. & Schork, N. J. The importance of phase information for human genomics. *Nat. Rev. Genet.* **12**, 215–223 (2011).
89. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
90. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
91. Luo, R., Schatz, M. C. & Salzberg, S. L. 16GT: a fast and sensitive variant caller using a 16-genotype probabilistic model. *Gigascience* **6**, 1–4 (2017).
92. Cilibrasi, R., Iersel, L. v., Kelk, S. & Tromp, J. The complexity of the single individual SNP haplotyping problem. *Algorithmica* **49**, 13–36 (2007).
93. Lo, C., Bashir, A., Bansal, V. & Bafna, V. Strobe sequence design for haplotype assembly. *BMC Bioinformatics* **12**, S24 (2011).
94. Rozowsky, J. et al. AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* **7**, 522 (2011).
95. Lynch, K. W. & Maniatis, T. Assembly of specific SR protein complexes on distinct regulatory elements of the *Drosophila* doublesex splicing enhancer. *Genes Dev.* **10**, 2089–2101 (1996).
96. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–1415 (2008).
97. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
98. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
99. Conesa, A. et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).
100. Abdel-Ghany, S. E. et al. A survey of the sorghum transcriptome using single-molecule long reads. *Nat. Commun.* **7**, 11706 (2016).
101. Byrne, A. et al. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* **8**, 16027 (2017).
102. Garalde, D. R. et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* https://doi.org/10.1038/nmeth.4577 (2018).
This is the first demonstration of direct RNA sequencing on an Oxford Nanopore MinION sequencer.
103. Tilgner, H. et al. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.* **33**, 736–742 (2015).
104. Wang, B. et al. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* **7**, 11708 (2016).
105. Gordon, S. P. et al. Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *PLoS ONE* **10**, e0132628 (2015).
This paper describes the ToFU algorithm for studying alternative splicing and isoform diversity using long-read sequencing.
106. Tardaguila, M. et al. SQANTI: extensive characterization of long read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* https://doi.org/10.1101/gr.222976.117 (2018).
107. Au, K. F. et al. Characterization of the human ESC transcriptome by hybrid sequencing. *Proc. Natl Acad. Sci. USA* **110**, E4821–E4830 (2013).
108. Deonovic, B., Wang, Y., Weirather, J., Wang, X. J. & Au, K. F. IDP-ASE: haplotyping and quantifying allele-specific expression at the gene and gene isoform level by hybrid sequencing. *Nucleic Acids Res.* **45**, e32 (2017).
109. Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
110. Lister, R. & Ecker, J. R. Finding the fifth base: genome-wide sequencing of cytosine methylation. *Genome Res.* **19**, 959–966 (2009).
111. Dinh, H. Q. et al. Advanced methylome analysis after bisulfite deep sequencing: an example in *Arabidopsis*. *PLoS ONE* **7**, e41528 (2012).
112. Flusberg, B. A. et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* **7**, 461–465 (2010).
This is one of the first demonstrations of the ability to directly detect methylated bases using PacBio long-read sequencing.
113. Fang, G. et al. Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. *Nat. Biotechnol.* **30**, 1232–1239 (2012).
114. Greer, E. L. et al. DNA methylation on N6-adenine in *C. elegans*. *Cell* **161**, 868–878 (2015).
115. Graralde, D. R. et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* https://doi.org/10.1038/nmeth.4577 (2018).
116. Zimin, A. V. et al. The first near-complete assembly of the hexaploid bread wheat genome, Triticum aestivum. *Gigascience* **6**, 1–7 (2017).
117. Poplin, R. et al. Creating a universal SNP and small indel variant caller with deep neural networks. Preprint at *bioRxiv* https://doi.org/10.1101/092890 (2016).
118. Danko, C. D., Meleshko, D., Bezcan, D., Mason, C. E. & Hajirasouliha, I. Minerva: an alignment and reference free approach to deconvolve linked-reads for metagenomics. Preprint at *bioRxiv* https://doi.org/10.1101/217869 (2017).
119. Tsai, Y. C. et al. Resolving the complexity of human skin metagenomes using single-molecule sequencing. *MBio* **7**, e01948–01915 (2016).
120. Burton, J. N., Liachko, I., Dunham, M. J. & Shendure, J. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3* **4**, 1339–1346 (2014).
121. Novak, A. M. et al. Genome graphs. *bioRxiv* https://doi.org/10.1101/101378 (2017).
122. Church, D. M. et al. Extending reference assembly models. *Genome Biol.* **16**, 13 (2015).
123. Matzaraki, V., Kumar, V., Wijmenga, C. & Zhernakova, A. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol.* **18**, 76 (2017).
124. Mayor, N. P. et al. HLA typing for the next generation. *PLoS ONE* **10**, e0127153 (2015).
125. Hayward, D. R., Bultitude, W. P., Mayor, N. P., Madrigal, J. A. & Marsh, S. G. The novel HLA-B*44 allele, HLA-B*44:220, identified by single molecule real-time DNA sequencing in a British caucasoid male. *Tissue Antigens* **86**, 61–63 (2015).
126. Wang, M. et al. PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations. *BMC Genomics* **16**, 214 (2015).
127. Nattestad, M. et al. Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. Preprint at *bioRxiv* https://doi.org/10.1101/174938 (2017).
128. Quick, J. et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232 (2016).

129. Faria, N. R. et al. Mobile real-time surveillance of Zika virus in Brazil. *Genome Med.* **8**, 97 (2016).
130. Schatz, M. C. & Phillippy, A. M. The rise of a digital immune system. *Gigascience* **1**, 4 (2012).
131. Biesecker, L. G. & Green, R. C. Diagnostic clinical genome and exome sequencing. *N. Engl. J. Med.* **370**, 2418–2425 (2014).
132. Schatz, M. C., Witkowski, J. & McCombie, W. R. Current challenges in de novo plant genome sequencing and assembly. *Genome Biol.* **13**, 243 (2012).
133. Eberle, M. A. et al. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res.* **27**, 157–164 (2017).
134. Schatz, M. C. Nanopore sequencing meets epigenetics. *Nat. Methods* **14**, 347–348 (2017).
135. Kamath, G. M., Shomorony, I., Xia, F., Courtade, T. A. & Tse, D. N. HINGE: long-read assembly achieves optimal repeat resolution. *Genome Res.* **27**, 747–756 (2017).
136. Xiao, C. L. et al. MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat. Methods* **14**, 1072–1074 (2017).
137. Lin, Y. et al. Assembly of long error-prone reads using de Bruijn graphs. *Proc. Natl Acad. Sci. USA* **113**, E8396–E8405 (2016).
138. Warren, R. L. et al. LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *Gigascience* **4**, 35 (2015).
139. Cao, M. D. et al. Scaffolding and completing genome assemblies in real-time with nanopore sequencing. *Nat. Commun.* **8**, 14515 (2017).
140. Vaser, R., Sovic, I., Nagarajan, N. & Sikic, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
141. Sovic, I. et al. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat. Commun.* **7**, 11307 (2016).
142. Lin, H. N. & Hsu, W. L. Kart: a divide-and-conquer algorithm for NGS read alignment. *Bioinformatics* **33**, 2281–2287 (2017).
143. Liu, B., Gao, Y. & Wang, Y. LAMSA: fast split read alignment with long approximate matches. *Bioinformatics* **33**, 192–201 (2017).
144. Elyanow, R., Wu, H. T. & Raphael, B. J. Identifying structural variants using linked-read sequencing data. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btx712 (2017).
145. Patterson, M. et al. WhatsHap: weighted haplotype assembly for future-generation sequencing reads. *J. Comput. Biol.* **22**, 498–509 (2015).
This study describes WhatsHap, a widely used and very fast phasing algorithm for long reads.
146. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
147. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).

Acknowledgements
The authors thank A. Phillippy, W. Timp, W. R. McCombie, S. Goodwin and R. Gibbs for helpful discussions. This work was supported, in part, by awards from the National Science Foundation (DBI-1350041) and from the National Institutes of Health (R01-HG006677 and UM1-HG008898). Also, this work was completed in part while H.L. was visiting the Simons Institute for the Theory of Computing, University of California, Berkeley, USA.

Author contributions
All authors contributed to all aspects of this manuscript, including researching data, discussing content and writing, reviewing and editing the manuscript before submission.

Competing interests
M.C.S. and F.J.S. have participated in Pacific Biosciences (PacBio) sponsored meetings over the past few years and have received travel reimbursement and honoraria for presenting at these events. PacBio had no role in decisions relating to the study and/or work to be published, data collection and analysis or the decision to publish.

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

Reviewer information
Nature Reviews thanks Heng Li, Renê Warren and the other anonymous reviewer(s) for their contribution to the peer review of this work.

RELATED LINKS
10X Genomics: https://www.10xgenomics.com/
Bionano Genomics: https://bionanogenomics.com/
Illumina: https://www.illumina.com/
Oxford Nanopore Technologies: https://nanoporetech.com/
Pacific Biosciences: http://www.pacb.com/
Proposed solutions to corruption of long-read sequencing files: https://github.com/samtools/hts-specs/issues/40