Long Reads: their Purpose and Place

Martin O. Pollard*[1,3], Deepti Gurdasani[2], Alexander J. Mentzer[1,4], Tarryn Porter[1,3], Manjinder S. Sandhu[1,3]

1. Global Health and Populations Group - Human Genetics, Wellcome Sanger Institute, Cambridge, UK

2. Sandhu Group - Department of Medicine, University of Cambridge Department of Public Health and Primary Care, Cambridge, UK

3. Big Data Analytics Group - Human Genetics, Wellcome Sanger Institute, Cambridge, UK

4. Wellcome Trust Centre for Human Genetics, Oxford, Oxfordshire, UK

* Corresponding author:

Martin O. Pollard (mp15@sanger.ac.uk)

Global Health and Populations Group - Human Genetics, Morgan Building Wellcome Sanger Institute, Hinxton, Cambridge, CB10 1HH, UK

Tel: +44 (0)1223 834244

Fax: +44 (0)1223 496802

© The Author(s) 2018. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

Abstract

In recent years long read technologies have moved from being a niche and specialist field to a point of relative maturity likely to feature frequently in the genomic landscape. Analogous to next generation sequencing (NGS), the cost of sequencing using long read technologies has materially dropped whilst the instrument throughput continues to increase. Together these changes present the prospect of sequencing large numbers of individuals with the aim of fully characterising genomes at high resolution. In this article, we will endeavour to present an introduction to long read technologies showing: what long reads are; how they are distinct from short reads; why long reads are useful; and how they are being used. We will highlight the recent developments in this field, and the applications and potential of these technologies in medical research, and clinical diagnostics and therapeutics.

When short reads are not enough

DNA is an extraordinarily compact storage medium, so small that developing ways to decode the sequence encoded in these molecules has been a topic of research for many years. The first method developed for sequencing DNA, often known as Sanger sequencing [1], was a low throughput process that detected bases by incorporation into a template strand, sequencing fragments of DNA up to 1000 bp long. The breakthrough allowing sequencing at scale finally came with the advent of NGS technology, which employed massively parallel reactions for high throughput. While these technologies

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

have been able to capture sequence from the majority of the genome and have found utility in the study of disease, their short reads and lack of contextual information has limited their utility in genome assembly and in resolving complex and repetitive regions of the genome.

The incremental improvements in read length that this generation of technology can yield is one of diminishing returns. Thus, to achieve substantial gains in mapping, assembly and phasing one must consider technology that provides an order of magnitude increase in read length [2]. Practically also there are many important problems in genetics where a short read of DNA (< 1000 base pairs) is insufficient (**Panel 1** and **Figure 1**).

Panel 1: Advantages and applications of long read sequencing

Limitations of short read data	Applications and advantages of long read sequencing
• Access to high GC content regions • Resolution of complex regions of the genome (e.g. MHC*) • Repetitive regions where short reads will not map uniquely • Systematic context specific error modes • Structural variation, and large segmental duplications • Paralogous regions of the genome	• De novo assembly from long reads to span the low complexity and repetitive regions, to create accurate assemblies [3]. • Targeted sequencing of complex genomic and paralogous regions and resolution of phase for clinical applications e.g. HLA† typing, ADPKD‡ [4]. • Transcriptomics, allowing full length sequencing of isoforms and examination of

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

• Resolution of phase (read-based phasing)	splicing [5]. • Detection of structural variants (e.g. segmental duplications, gene loss and fusion events) • Single molecule sequencing allows examination of clonal heterogeneity of pathogens, and immunogenic cells • Long-range characterisation of methylation patterns

*MHC: Major Histocompatibility complex

†HLA: Histocompatibility leucocyte antigen

‡ADPKD: Autosomal dominant polycystic kidney disease

Key to achieving high quality results with all long read technologies is the use of high molecular weight (HMW) DNA as a starting material. The utility of these methods depends on a long DNA fragment size, with DNA damage and fragmentation limiting the quality of data obtained. Specific protocols for DNA extraction such as the agarose gel protocol for BioNano are ideal to maximise yield from these methods.

Long Read Technologies Single Molecule RealTime sequencing

The first long read sequencing technology to achieve a widespread deployment is the Single Molecule RealTime (SMRT) sequencing technology from Pacific Biosciences

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

(PacBio). The SMRT system implemented in their Sequel and RS- II platforms uses a massively parallel system of polymerases each bound to a single molecule of target DNA that has been circularised with a pair of hairpin sequencing adaptors (the SMRTbell) (**Figure 2a**). Incorporation of labelled bases by a polymerase on the template strand causes fluorescence. The resulting signal is detected by a CCD camera via a zero-mode waveguide (ZMW) [6, 7], yielding a combination of signal and time series information. Reads produced by this technology typically peak at 100Kbp in length and a typical N50 on recent polymerases is approximately 20Kbp.

One complication of SMRT sequencing is the high error rate of this process relative to short read sequencing, at 11-14% depending on polymerase and chemistry. However, this error mode is stochastic (by contrast with other technologies), and can be mitigated by repeated measurements of the sequence. With PacBio sequencing, this is carried out by repeated forward and reverse sequencing passes over the circularised SMRTbell molecule (**Figure 2a**). Adaptor sequences can be removed from the generated sequence to provide enough subreads to generate a highly accurate consensus of each molecule. This process is known as circular consensus sequencing (CCS) and has been shown to reduce basecalling error substantially [8] whilst also enabling the strand specific calling of base modifications in unamplified DNA [9]. When long DNA fragments are sequenced, these may not be parsed more than once in the SMRTbell; in this case, increasing coverage and then calling a consensus across reads can also achieve a reduced error rate; a method frequently used in polishing assemblies [10].

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

Oxford Nanopore Technologies

The next successful single molecule technology to hit the market was that produced by Oxford Nanopore Technologies (ONT) [11]. This technology is based on passing a single strand of DNA through a nanopore with an enzyme attached, and measuring changes in the electrical signal across the pore (**Figure 2b**). The signal is then amplified and measured to determine the bases that passed through. As the pore holds several bases at a time (typically 5-mers), overlapping k-mers that cause changes in raw current must be inferred and used to make base calls, a process which can be error prone. By measuring the shape of the molecule passing through the pore ONT not only reads the sequence of the DNA but like SMRT is also able to detect base modifications [12]. However, unmodelled base modifications and systematic DNA context specific errors [13] currently limit the utility of the technology.

Oxford Nanopore MinION technology heralds the promise of a pocket size sequencer, with reads from ONT that can stretch into the hundreds of kilobases with appropriate DNA preparation, and megabase long reads that have been observed when a large number of flow cells have been used. There appears to be no intrinsic read-length limit for ONT, other than the size of DNA fragments. Recent improvements in technology, library preparation and throughput have allowed the first human line sequenced on the MinION (GM12878) earlier this year [14]. This study generated ultra-long reads (>800Kbp), and suggested that addition of modest coverage with ultra-long read sequencing to existing assemblies may substantially improve resolution of contigs and

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

haplotypes. While the error rate is comparable to SMRT sequencing, a component of the error is systematic and context-specific, limiting the ability to correct this by increasing coverage [13] and requiring polishing with other technologies instead.

ONT has developed a distinct strategy to mitigate stochastic error on their platform, focusing on the way that the template strand passes through the pore. ONT cannot simply circularise the DNA. Instead, both the template and complement strands of the DNA molecule are joined by a hairpin loop during library prep (2D) or tethered in such a way ($1D^2$) to allow sequential forward and reverse strand sequencing. Combining these data greatly enhances accuracy and reduces random error.

The use of nanopores as a nucleic acid sequencing technology is not entirely exclusive to ONT; at least one similar but distinct competing technology is also under development by Roche.

10X Genomics Chromium system

An alternative to the aforementioned single molecule sequencing methods is the 10X Genomics Chromium system. Whilst this is not technically a long read sequencing technology, it is an important member of this ecosystem and can solve similar problems such as mapping, phasing and assembly (**Figure 2c**). Chromium has lower cost compared to ONT and SMRT because of the use of the nearly ubiquitous Illumina short reads in its sequencing process.

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

The basis of this technology [15] is the barcoding of large fragments of DNA (preferably >100Kbp) in an initial digital droplet PCR step. In each droplet, a single fragment is both sheared and then tagged with a semi-unique molecular barcode (**Figure 2c**). The resulting fragments are then sequenced like any other Illumina library. The barcode allows for determination of the relative spatial orientation of the tags, and allows phasing and assembly of contigs by combining information across multiple tags [15, 16]. Additionally, because the data provide spatial orientation across the genome, it is possible to use it to scaffold data from other methods [17].

Allied Technologies

Allied technologies associated with long read sequencing such as: Optical Mapping, HiC and similar have been used to enhance the final results from sequencing. Optical mapping technologies such as BioNano Irys and Saphyr label DNA and then image the labelled DNA to generate genome maps. These genome maps are used to scaffold contigs produced by assembly [18] and also to discover large (>500bp) structural variants and inversions. HiC can be used to assay chromosomal conformation and is particularly useful in assigning assembled sequences to chromosomes [19].

The utility of long-read technology: recent developments

High resolution genome Assemblies

Accurate assemblies of the genomes of organisms are crucial to understanding

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

organismal diversity, speciation, evolution of species, and the impact of genomic diversity on health and disease. The current human genome reference GRCh38 has been assembled from the DNA of multiple donors, and represents a mosaic of haplotypes. However, several studies have suggested that existing human reference genomes may not fully reflect the diversity of global human populations, and may be biased towards diversity in European populations [20, 21, 22]. This has important implications for human basic and medical research. Assembling the human genome has involved extensive curation with clone-based assembly methods and sanger sequencing. Long read technologies provide a high throughput platform for characterisation of genomes through highly contiguous assemblies (**Figure 3**).

The early long read platforms produced reads that were only a few kilobases long with a high per-base cost; however, they quickly carved a niche in the creation and finishing of assemblies. These long reads could close gaps in genomes by spanning the low complexity regions that would otherwise require many costly YAC, BAC and fosmid clones to be created and sequenced. Thus, many of the early tools such as PBJelly were focused on gap closure [23, 24]. The high per-base error rate also required new assembly algorithms, and new tools were created to polish the final assembly with Illumina reads to eliminate basecalling error [12]. Clone based assembly methods were not eliminated entirely either as they provided useful spatial context, but long reads provided a new way to sequence clones in a high throughput manner [25].

Long read sequencing methods have contributed to platinum quality reference sequences

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

such as NA12878 [26, 14] and the haploid sequences CHM1 [27] and CHM13 [28], as well filling many gaps in the human reference [29, 18, 25]. Of particular note are the first Chinese [18] and Korean [25] human reference genomes which have been created to answer questions about population-specific sequence. These sequences have resulted in highly contiguous assemblies, closing a high proportion of gaps in the human genome. These have led to discovery of population specific sequences, demonstrating the need for further assemblies from non-European population groups. Recently, higher coverage sequencing (~ 60x) of two haploid genomes has also been used to identify substantial structural variation, the vast majority of which have not been recovered from sequencing using NGS technologies [28]. Characterisation of high resolution population specific reference genomes from initiatives such as the Genome in a Bottle (GIAB) [30] and the Genome Diversity in Africa project (GDAP) [31] (**Figure 3**) will provide important resources for population and medical genetics, and also allow a clearer understanding of the evolutionary demographic history of different populations by better delineation of phase [31].

Most human assemblies have involved a haploid representation of the genome, where information from the two chromosomes is collapsed into a single sequence. Generation of haplotype representations of the genome can reduce error in the final assembly, particularly in the case of segmental duplications [32, 16]. While long read technologies can generate phase information over long contiguous segments, these methods cannot resolve phase over long regions of homozygosity or assembly gaps. Assembly of haploid

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

genomes, therefore, requires additional contextual information, which can be provided by linked-read approaches. More recently, trio based methods (where parents are sequenced using Illumina short reads, with offspring sequenced with long reads) have been used to provide this contextual information by separation of maternal and paternal haplotypes prior to assembly using a father-mother-offspring trio [33]. This method has been applied to yield a highly contiguous diploid assembly of an F1 hybrid of two bovine subspecies with a quality surpassing previous cattle reference genomes [33].

Long reads have been successfully applied to organisms with smaller genomes as well as bacteria and viruses, with the advantage that for some of these the entire genome can be spanned by a single long read [34]. The Tree of Life initiative, a collaboration across multiple centres is in the process of developing high resolution reference sequences for >50 vertebrate species using a combination of long read, short read and linked read approaches. Another leading project is the large bacterial sequencing project NCTC 3000 at the Wellcome Sanger Institute, which is using PacBio sequencing to sequence complete bacterial genomes (https://www.phe-culturecollections.org.uk/collections/nctc-3000-project.aspx). These relatively small genomes (Escherichia coli is for example 4.6Mbases) can often have their chromosomes and plasmids assembled into single contigs. The construction of full and accurate assemblies of these organisms allow fine-scale phylogenies of these organisms to be constructed and is also helpful in the field of epidemiology when tracing the source of an outbreak. A recent example of this was a study where SMRT sequencing was used to identify a reservoir of antibiotic resistant

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

plasmids within hospitals [35].

In addition to DNA sequencing, ONT sequencing has been applied to sequence RNA directly rather than relying on an intermediate cDNA step, allowing direct sequencing of RNA viruses and detection of splice variants and base modifications directly from RNA molecules. An example of this is the recent direct sequencing and assembly of the influenza A virus in a native RNA form without amplification or conversion to DNA [36].

Targeted sequencing

From a clinical point of view targeted sequencing is an area where long reads are likely to have the greatest initial impact. In the diverse, complex and clinically relevant regions such as the HLA [37], KIR [38] and BRCA; and in pharmacologically relevant genes such as CYP2D6 [39, 40], targeted sequencing has allowed clinicians and researchers to characterise areas of the genome which were previously inaccessible using NGS methods. In addition, where diversity is high it has become possible to call and phase variation across the entire gene. This approach has since been used to retype 126 HLA reference samples across 6 loci and is now considered a gold standard for clinical sequencing for stem cell transplants [41].

Typically, when targeting such a region, a long range PCR reaction is used to specifically amplify the genes of interest. However recently there have been studies demonstrating the use of pulldowns and CRISPR to capture the region of interest with little or no

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

amplification [42]. The advantage of these reduced and non-amplification based approaches is the removal of PCR error as a factor, particularly in tandem repeats and GC rich regions [42]. Additionally, in the case of CRISPR methods, capture of raw genomic material allows DNA modification information to be read.

Transcriptomics and RNA

In addition to its many uses with DNA, long read technology also has provided many new insights into the world of transcriptomes and ncRNA by allowing for sequencing of these full length isoforms rather than relying on the assembly of sheared NGS fragments, a method prone to a high rate of false positives and ambiguities [43]. Direct sequencing of isoforms can be particularly useful in complex polyploid genomes such as the coffee plant [44], where construction of a reference transcriptome is otherwise extremely challenging. In addition to its usefulness in reference transcriptomes IsoSeq has been used in functional studies to analyse the expression of various disease-linked proteins such as TP53 in leukaemia [45].

The MinION platform has recently been used to sequence cDNA; applications of this, such as single cell sequencing of immune cells illustrates the power of such methods to examine clonal heterogeneity in gene expression and isoform usage, potentially revolutionising our understanding of the repertoire and functions of immunological cell receptors [46].

Epigenetics

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

SMRT sequencing technology is able to detect base modification, as it records base kinetics of the polymerase, when DNA molecules are sequenced directly without PCR. Similarly, Nanopore technology can also detect base modifications due to variation in ionic currents. However, because amplification of DNA would erase base modifications, these methods requires relatively large amounts of native, unamplified DNA as input material. Recent innovations that combine bi-sulphate conversion with SMRT sequencing have allowed direct high throughput analysis of CpG methylation without requiring large quantities of sample [47], providing an avenue for more accurate assessment of CpG islands, and allele-specific CpG methylation.

Clinical applications

The advantages of long-read technologies in accessing complex regions of the genome, make these ideal for clinical applications in diagnosis, prognostication and personalised medicine. Early clinical applications have included sequencing of tandem repeats in fragile X syndrome, spinocerebellar ataxia, providing accurate diagnostics, and potential for prognostication in clinical genetics. SMRT sequencing has also been used to resolve structural variants associated with mendelian disease [48].

Long read sequencing technologies are rapidly moving towards the mainstay of high resolution HLA typing for transplant registries in certain regions [37]; with high resolution typing potentially having implications for better matching, and clinical outcomes of patients undergoing transplantation. This is even more important in

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

populations which are poorly represented in current reference sequence databases, limiting disambiguation of clinical types when using standard methods for typing. The HLA diversity in Africa project, which aims to characterise high resolution HLA types across >20 ethno-linguistic in Africa has recently completed sequencing of ~2000 individuals using long-read sequencing, identifying high levels of novelty in class I and class II HLA types [49]. This panel will provide an important resource for clinical HLA typing in populations of African ancestry, as well as a platform for highly accurate imputation of HLA types in medical genetics research.

Using long range PCR amplicons, with barcoding and long read technology also allow better delineation of genes from pseudogenes, such as for sequencing PKD1 for diagnosing autosomal-dominant polycystic kidney disease, for which diagnostic accuracy of NGS technologies has been limited [50]. SMRT sequencing has also been used to tailor treatment in patients with cancer, by identifying low frequency resistant mutations in BCR-ABL1 that affect treatment efficacy in patients with CML [51]. Applications of SMRT sequencing in reproductive medicine, to identify parent of origin effects, and for pre-implantation diagnosis have been previously noted [52].

Full sequencing of several virus genomes in a single contig by long-read sequencing has provided unique avenues for identification of resistant mutations for clinical applications. Proof-of concept studies have generated protocols to examine low frequency (up to 0.25%) associated mutations for HIV and HCV resistance to drugs, through deep sequencing of full length quasispecies [53]. Methylation profiles of pathogens examined

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

using SMRT approaches have also been shown to correlate with pathogenicity, and virulence, potentially providing a new avenue for applications in infectious disease surveillance.

The future

Long read technologies are improving rapidly, and may become the mainstay of sequencing; however, the broader application of long read technologies are currently limited by a lower throughput, higher error rate, and higher cost per base relative to short read sequencing. Wider use of such technologies in the clinical context may rapidly improve our understanding of cancer, pathogen evolution, drug resistance, and genetic diversity in complex regions of the genome that have important implications for clinical care. Parallel development of existing technology to allow high throughput PCR-free sequencing will be important in sequencing difficult regions of the genome [54].

At present, no single long read technology has any clear advantage from a scientific point of view, and thus it seems likely that the future of long read sequencing is more likely to be decided on commercial terms rather than scientific. Whichever technology captures the market, it is clear that as these technologies become more affordable they will continue to shine a light into previously intractable regions of the genome with ever larger sample sizes and longer read lengths, allowing new discovery in these evolving fields.

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

Acknowledgements

Uttara Partap for copyediting. Conflict of Interest statement. None declared.

Funding

This work was supported by the Wellcome Trust [grant number 098051] to MSS; the National Institute for Health Research Cambridge Biomedical Research Centre (UK) to MSS; a Wellcome Trust Fellowship [grant number 106289/Z/14/Z] to AJM; and the Medical Research Council (MRC) (MR/S003711/1) to DG. This work was also partially funded by IAVI with the generous support of USAID, and the Bill & Melinda Gates Foundation; a full list of IAVI donors is available at www.iavi.org. The contents of this manuscript are the responsibility of the authors and do not necessarily reflect the views of USAID or the US Government.

References

- [1] Sanger, F. and Coulson, A. (1975). “A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase”. *J. Mol. Biol.* **94**, pp. 441–448, 10.1016/0022-2836(75)90213- 2.

- [2] Li, W. and Freudenberg, J. (2014). “Mappability and read length”. *Front. Genet.* **5**, p. 381, 10.3389/fgene.2014.00381.

- [3] Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M.,

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

Collins, J. E., Humphray, S., McLaren, K., Matthews, L., et al. (2013). “The zebrafish reference genome sequence and its relationship to the human genome”. *Nature* **496**, pp. 498–503, 10.1038/nature12111.

- [4] Hosomichi, K., Jinam, T. A., Mitsunaga, S., Nakaoka, H., and Inoue, I. (2013). “Phase-defined complete sequencing of the HLA genes by next-generation sequencing”. *BMC Genomics* **14**, pp. 355–355, 10. 1186/1471-2164-14-355.

- [5] Wang, B., Tseng, E., Regulski, M., Clark, T. A., Hon, T., Jiao, Y., Lu, Z., Olson, A., Stein, J. C., and Ware, D. (2016). “Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing”. *Nat. Commun.* **7**, p. 11708, 10.1038/ncomms11708.

- [6] Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G., and Webb, W. W. (2003). “Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations”. *Science* **299**, pp. 682–686, 10.1126/science.1079700.

- [7] Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). “Real-Time DNA Sequencing from Single Polymerase Molecules”. *Science* **323**, pp. 133–138, 10.1126/science.1162986.

- [8] Travers, K. J., Chin, C.-S., Rank, D. R., Eid, J. S., and Turner, S. W. (2010). “A flexible and efficient template format for circular consensus sequencing and SNP detection”. *Nucleic Acids Res.* **38**, e159–e159, 10.1093/nar/gkq543.

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

- [9] Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., Korlach, J., and Turner, S. W. (2010). "Direct detection of DNA methylation during single-molecule, real-time sequencing". *Nat. Methods* **7**, pp. 461–465.

- [10] Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., et al. (2013). "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data". *Nat. Methods* **10**, pp. 563–569, 10.1038/nmeth.2474.

- [11] Deamer, D., Akeson, M., and Branton, D. (2016). "Three decades of nanopore sequencing". *Nat. Biotechnology* **34**, pp. 518–524, 10.1038/nbt. 3423.

- [12] Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J., and Timp, W. (2017). "Detecting DNA cytosine methylation using nanopore sequencing". *Nat. Methods* **14**, pp. 407–410.

- [13] Krishnakumar, R., Sinha, A., Bird, S. W., Jayamohan, H., Edwards, H. S., Schoeniger, J. S., Patel, K. D., Branda, S. S., and Bartsch, M. S. (2018). "Systematic and stochastic influences on the performance of the MinION nanopore sequencer across a range of nucleotide bias". *Scientific Reports* **8**, p. 3159, 10.1038/s41598-018-21484-w.

- [14] Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., et al. (2018). "Nanopore sequencing and assembly of a human genome with ultra-long reads". *Nat. Biotechnology* **36**,

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

pp. 338–345, 10.1038/nbt.4060.

- [15] Zheng, G. X. Y., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D. A., Merrill, L., Terry, J. M., et al. (2016). "Haplotyping germline and cancer genomes with high-throughput linked-read sequencing". *Nat. Biotechnology* **34**, pp. 303–311, 10.1038/nbt.3432.

- [16] Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., and Jaffe, D. B. (2017). "Direct determination of diploid genome sequences". *Genome Res.* **27**, pp. 757–767, 10.1101/gr.214874.116.

- [17] Yeo, S., Coombe, L., Warren, R. L., Chu, J., and Birol, I. (2018). "ARCS: scaffolding genome drafts with linked reads". *Bioinformatics* **34**, pp. 725–731, 10.1093/bioinformatics/btx675.

- [18] Shi, L., Guo, Y., Dong, C., Huddleston, J., Yang, H., Han, X., Fu, A., Li, Q., Li, N., Gong, S., et al. (2016). "Long-read sequencing and de novo assembly of a Chinese genome". *Nat. Commun.* **7**, p. 12065, 10. 1038/ncomms12065.

- [19] Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., Lander, E. S., Aiden, A. P., et al. (2017). "De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds". *Science* **356**, pp. 92–95, 10.1126/science.aal3327.

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

- [20] Brandt, D. Y. C., Aguiar, V. R. C., Bitarello, B. D., Nunes, K., Goudet, J., and Meyer, D. (2015). "Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data". *G3: Genes, Genomes, Genetics* **5**, pp. 931–941, 10.1534/g3.114.015784.

- [21] Lunter, G. and Goodson, M. (2011). "Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads". *Genome Res.* **21**, pp. 936–939, 10.1101/gr.111120.110.

- [22] Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., and Pritchard, J. K. (2009). "Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data". *Bioinformatics* **25**, pp. 3207–3212, 10.1093/bioinformatics/btp579.

- [23] English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D. M., Reid, J. G., Worley, K. C., et al. (2012). "Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology". *PLOS ONE* **7**, pp. 1–12, 10.1371/journal.pone.0047768.

- [24] Worley, K. C., English, A. C., Richards, S., Ross-Ibarra, J., Han, Y., Hughes, D., Deiros, D. R., Vee, V., Wang, M., Boerwinkle, E., et al. (2014). "Improving Genomes Using Long Reads and PBJelly 2". Presented at: *Plant & Animal Genome XXII*.

- [25] Seo, J.-S., Rhie, A., Kim, J., Lee, S., Sohn, M.-H., Kim, C.-U., Hastie, A., Cao,

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

H., Yun, J.-Y., Kim, J., et al. (2016). “De novo assembly and phasing of a Korean human genome”. *Nature* **538**, p. 243, 10.1038/nature20098.

- [26] Pendleton, M., Sebra, R., Pang, A. W. C., Ummat, A., Franzen, O., Rausch, T., Stu¨tz, A. M., Stedman, W., Anantharaman, T., Hastie, A., et al. (2015). “Assembly and diploid architecture of an individual human genome via single-molecule technologies”. *Nat. Methods* **12**, pp. 780–786, 10.1038/nmeth.3454.

- [27] Chaisson, M. J., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., et al. (2015). “Resolving the complexity of the human genome using single-molecule sequencing”. *Nature* **517**, pp. 608–611, 10.1038/nature13907.

- [28] Huddleston, J., Chaisson, M. J., Steinberg, K. M., Warren, W., Hoekzema, K., Gordon, D., Graves-Lindsay, T. A., Munson, K. M., Kronenberg, Z. N., Vives, L., et al. (2017). “Discovery and genotyping of structural variation from long-read haploid genome sequence data”. *Genome Res.* **27**, pp. 677–685, 10.1101/gr.214007.116.

- [29] Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., et al. (2017). “Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly”. *Genome Res.* **27**, pp. 849–864, 10.1101/gr.213611.116.

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

- [30] Zook, J. M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., and Salit, M. (2014). “Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls”. *Nat. Biotechnology* **32**, pp. 246–251, 10.1038/nbt.2835.

- [31] Gurdasani, D., Martinez, J. ., Pollard, M. ., Carstensen, T., Pomilla, C., and GDAP Investigators (2016). “The Genome Diversity in Africa Project: A deep catalogue of genetic diversity across Africa”. Presented at the *66th Annual Meeting of The American Society of Human Genetics*, Vancouver, Canada.

- [32] Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., Dunn, C., O’Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., et al. (2016). “Phased diploid genome assembly with single-molecule real-time sequencing”. *Nat. Methods* **13**, pp. 1050–1054, 10.1038/nmeth. 4035.

- [33] Koren, S., Rhie, A., Walenz, B. P., Dilthey, A. T., Bickhart, D. M., Kingan, S. B., Hiendleder, S., Williams, J. L., Smith, T. P., and Phillippy, A. (2018). “Complete assembly of parental haplotypes with trio binning”. *bioRxiv*, 10.1101/271486.

- [34] Koren, S. and Phillippy, A. M. (2015). “One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly”. *Curr. Opin. Microbiol.* **23**, pp. 110–120, https://doi.org/ 10.1016/j.mib.2014.11.014.

- [35] Weingarten, R. A., Johnson, R. C., Conlan, S., Ramsburg, A. M., Dekker, J. P., Lau, A. F., Khil, P., Odom, R. T., Deming, C., Park, M., et al. (2018). “Genomic

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

Analysis of Hospital Plumbing Reveals Diverse Reservoir of Bacterial Plasmids Conferring Carbapenem Resistance". *mBio* **9**, 10.1128/mBio.02011-17.

- [36] Keller, M. W., Rambo-Martin, B. L., Wilson, M. M., Ridenour, C. A., Shepard, S. S., Stark, T. J., Neuhaus, E. B., Dugan, V. G., Wentworth, D. E., and Barnes, J. R. (2018). "Direct RNA Sequencing of the Complete Influenza A Virus Genome". *bioRxiv*, 10.1101/300384.

- [37] Mayor, N. P., Robinson, J., McWhinnie, A. J. M., Ranade, S., Eng, K., Midwinter, W., Bultitude, W. P., Chin, C.-S., Bowman, B., Marks, P., et al. (2015). "HLA Typing for the Next Generation". *PLOS ONE* **10**, pp. 1–12, 10.1371/journal.pone.0127153.

- [38] Roe, D., Vierra-Green, C., Pyo, C.-W., Eng, K., Hall, R., Kuang, R., Spellman, S., Ranade, S., Geraghty, D. E., and Maiers, M. (2017). "Revealing complete complex KIR haplotypes phased by long-read sequencing technology". Genes And Immunity 18, pp. 127–134, 10.1038/gene.2017.10.

- [39] Buermans, H. P., Vossen, R. H., Anvar, S. Y., Allard, W. G., Guchelaar, H.-J., White, S. J., Dunnen, J. T. den, Swen, J. J., and Straaten, T. van der (2017). "Flexible and Scalable Full-Length CYP2D6 Long Amplicon PacBio Sequencing". *Hum. Mutat.* **38**, pp. 310–316, 10.1002/humu.23166.

- [40] Yang, Y., Botton, M. R., Scott, E. R., and Scott, S. A. (2017). "Sequencing the CYP2D6 gene: from variant allele discovery to clinical pharmacogenetic testing".

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

Pharmacogenomics **18**, pp. 673–685, 10.2217/pgs-2017- 0033.

- [41] Turner, T. R., Hayhurst, J. D., Hayward, D. R., Bultitude, W. P., Barker, D. J., Robinson, J., Madrigal, J. A., Mayor, N. P., and Marsh, S. G. E. (2018). "Single molecule real-time DNA sequencing of HLA genes at ultra-high resolution from 126 International HLA and Immunogenetics Workshop cell lines". *HLA* **91**, pp. 88–101, 10.1111/tan.13184.

- [42] Tsai, Y.-C., Greenberg, D., Powell, J., Hoijer, I., Ameur, A., Strahl, M., Ellis, E., Jonasson, I., Mouro Pinto, R., Wheeler, V., et al. (2017). "Amplification-free, CRISPR-Cas9 Targeted Enrichment and SMRT Sequencing of Repeat-Expansion Disease Causative Genomic Regions". *bioRxiv*, 10.1101/203919.

- [43] Steijger, T., Abril, J. F., Engstr¨om, P. G., Kokocinski, F., Consortium, T. R., Hubbard, T. J., Guig´o, R., Harrow, J., and Bertone, P. (2013). "Assessment of transcript reconstruction methods for RNA-seq". *Nat. Methods* **10**, pp. 1177–1184, 10.1038/nmeth.2714.

- [44] Cheng, B., Furtado, A., and Henry, R. J. (2017). "Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts". *GigaScience* **6**, pp. 1–13, 10.1093/gigascience/gix086.

- [45] Lod´e, L., Ameur, A., Coste, T., M´enard, A., Richebourg, S., Gaillard, J.-B., Le Bris, Y., B´en´e, M.-C., Lavabre-Bertrand, T., and Soussi, T. (2018). "Single-molecule DNA sequencing of acute myeloid leukemia and myelodysplastic

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

syndromes with multiple TP53 alterations". *Haematologica* **103**, e13–e16, 10.3324/haematol.2017.176719.

- [46] Byrne, A., Beaudin, A. E., Olsen, H. E., Jain, M., Cole, C., Palmer, T., DuBois, R. M., Forsberg, E. C., Akeson, M., and Vollmers, C. (2017). "Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells". *Nat. Commun*. **8**, p. 16027.

- [47] Yang, Y., Sebra, R., Pullman, B. S., Qiao, W., Peter, I., Desnick, R. J., Geyer, C. R., DeCoteau, J. F., and Scott, S. A. (2015). "Quantitative and multiplexed DNA methylation analysis using long-read single-molecule real-time bisulfite sequencing (SMRT-BS)". *BMC Genomics* **16**, p. 350, 10.1186/s12864-015-1572-7.

- [48] Merker, J. D., Wenger, A. M., Sneddon, T., Grove, M., Zappala, Z., Fresard, L., Waggott, D., Utiramerur, S., Hou, Y., Smith, K. S., et al. (2017). "Long-read genome sequencing identifies causal structural variation in a Mendelian disease". *Gen. Med.* **20**, pp. 159–163.

- [49] Pollard, M. ., Tommy, C., Cristina, P., Gurdasani, D., and Investigators, G. (2017). "The MHC Diversity in Africa Resource: A roadmap to understanding HLA diversity in Africa". Presented at the *67th Annual Meeting of The American Society of Human Genetics*, Orlando, Florida, USA.

- [50] Borr`as, D. M., Vossen, R. H. A. M., Liem, M., Buermans, H. P. J., Dauwerse,

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

H., Heusden, D., Gansevoort, R. T., Dunnen, J. T., Janssen, B., Peters, D. J. M., et al. (2017). “Detecting PKD1 variants in polycystic kidney disease patients by single-molecule long-read sequencing”. *Hum. Mutat.* **38**, pp. 870–879, 10.1002/humu.23223.

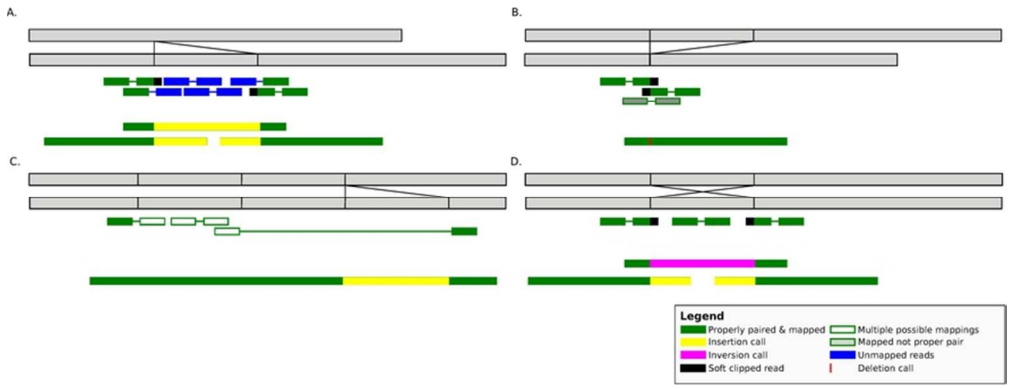
- [51] Cavelier, L., Ameur, A., H¨aggqvist, S., H¨oijer, I., Cahill, N., Olsson-Str¨omberg, U., and Hermanson, M. (2015). “Clonal distribution of BCR-ABL1 mutations and splice isoforms by single-molecule long-read RNA sequencing”. *BMC Cancer* **15**, p. 45, 10.1186/s12885-015-1046- y.

- [52] Wilbe, M., Gudmundsson, S., Johansson, J., Ameur, A., Stattin, E., Anner´en, G., Malmgren, H., Frykholm, C., and Bondeson, M. (2017). “A novel approach using long-read sequencing and ddPCR to investigate gonadal mosaicism and estimate recurrence risk in two families with developmental disorders”. *Prenatal Diagnosis* **37**, pp. 1146–1154, 10.1002/pd.5156.

- [53] Bull, R. A., Eltahla, A. A., Rodrigo, C., Koekkoek, S. M., Walker, M., Pirozyan, M. R., Betz-Stablein, B., Toepfer, A., Laird, M., Oh, S., et al. (2016). “A method for near full-length amplification and sequencing for six hepatitis C virus genotypes”. *BMC Genomics* **17**, p. 247, 10.1186/ s12864-016-2575-8.

- [54] Ardui, S., Ameur, A., Vermeesch, J. R., and Hestand, M. S. (2018). “Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics”. *Nuc. Acids Res.* **46**, pp. 2159–2168, 10.1093/nar/gky066.

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018



Caption : Fig 1 | Behavior of reads around genomic events
┼ A) Large Insertion: Short reads at the edge of the variant are be soft-clipped. Reads within the insertion will either be unmapped or mapped incorrectly. Large reads with either span insertion or have enough context to be marked as inserted sequence.
┼ B) Large Deletion: Short reads spanning the deletion may be mismapped or only have one of the reads marked as mapped because the reference measured length indicates the insert size deviates from the expected distribution. Long reads will span the gap but most will have enough context to call the deletion.┼ C) Copy number variation: Where the read length exceeds the length of the CNV region reads will map correctly. Shorter reads may be collapsed and show up as increased depth in a pileup or be marked as mapping poorly.
D) Inversion: Reads will either be represented as a primary alignment with an inverted supplementary or manifest as soft clipping around the edge of the inversion with a reduction in depth where reads span the edge of the inversion.

74x28mm (300 x 300 DPI)

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

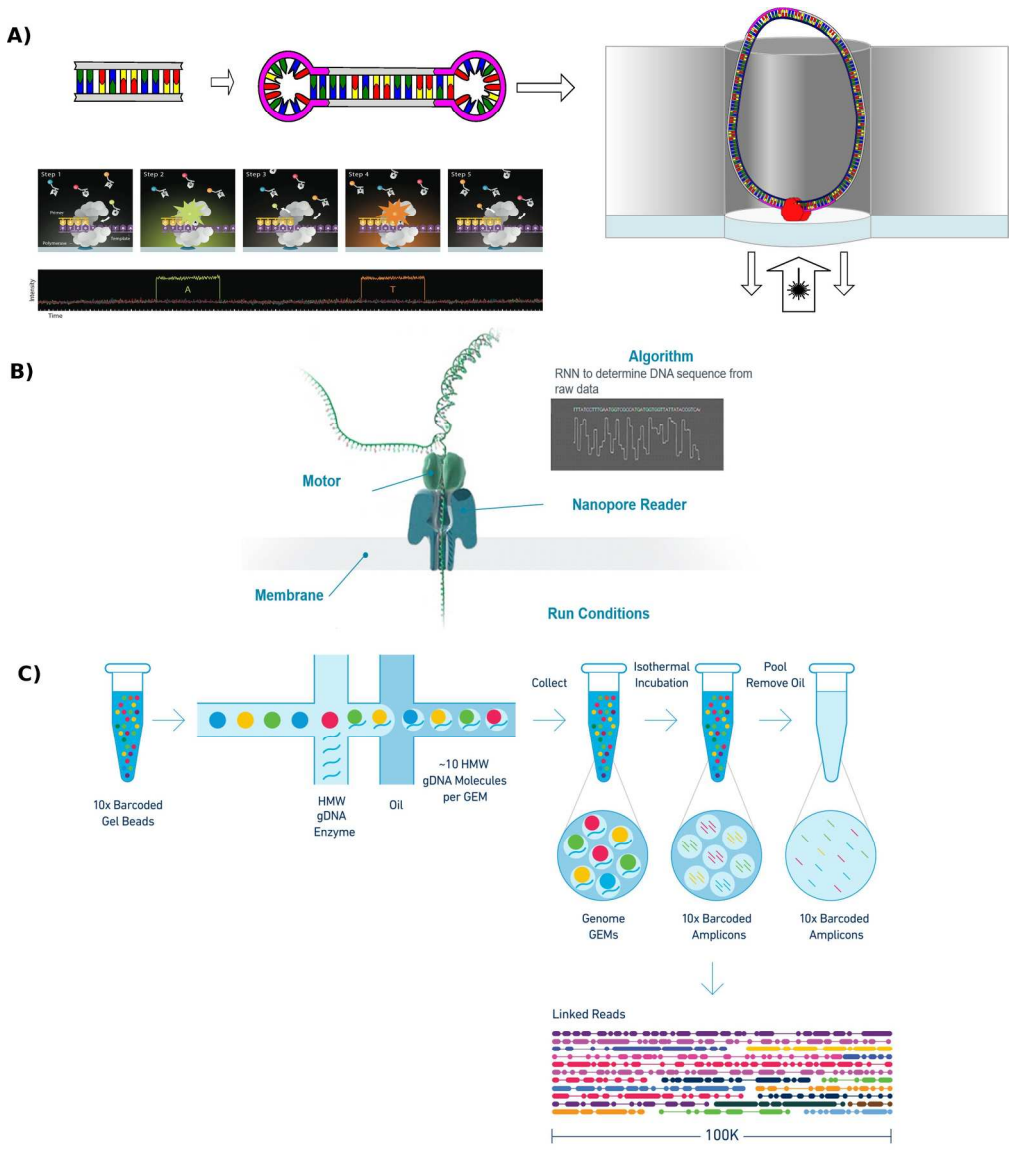


Fig 2.| Long read sequencing technologies

A) PacBio SMRT sequencing. Double stranded DNA is first sheared and size selected to the desired length and then sequencing adaptors are annealed. The adaptors are bound to a sequencing primer and strand displacing polymerase which adheres to the bottom of a well containing a zero mode wave guide. Following a pre-extension period where the polymerase reaction is run in the dark, the fragment is illuminated with a laser and as each base in the sequencing solution is incorporated, the fluorophore is detected and the polymerase reaction displaces it, giving a time and intensity signal which is converted into a base call.

B) Oxford Nanopore technology passes the DNA molecule through a nanopore attached the flow cell surface membrane. As each base of the DNA molecule passes through the pore changes to the current passing through the pore are detected and converted into a signal. The signal detected is passed to a recurrent neural network (RNN) which converts it into base calls.

C) 10X Genomics Chromium technology works by means of an emulsion droplet technology, where gel beads are mixed with high molecular weight genomic DNA and an enzyme. Within each gel bead DNA is sheared and barcoded, creating fragments which can then be sequenced with Illumina sequencing. The presence of the Chromium barcode then provides a mapper or assembler with linked reads, allowing the

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

relative spatial position of the fragments to be estimated

Components of figure reproduced with permission from Pacific Biosciences, Oxford Nanopore Technologies and 10X Genomics.

198x227mm (300 x 300 DPI)

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018

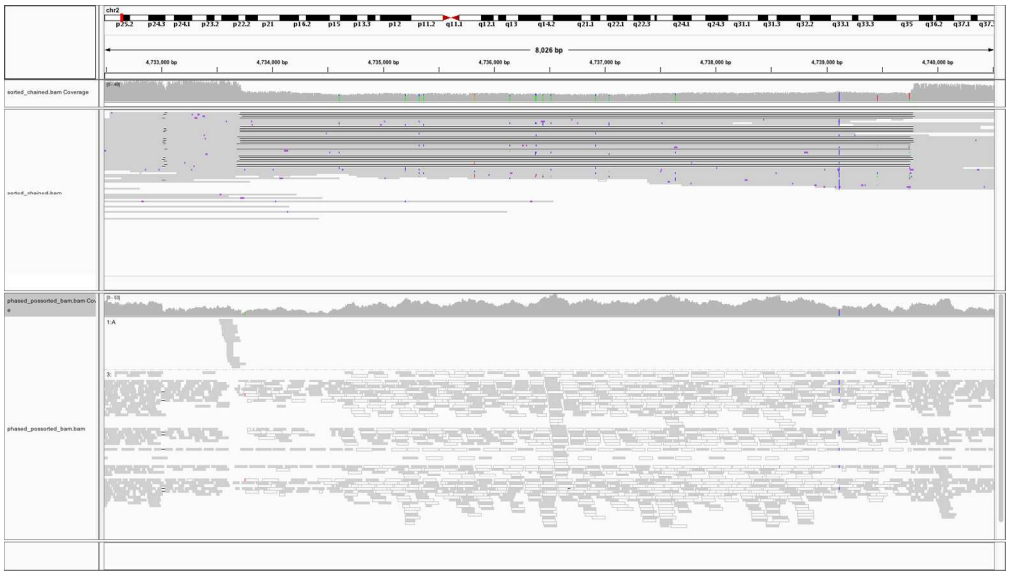


Fig.3 | Long reads span and call variations that short reads cannot. IGV (http://software.broadinstitute.org/software/igv/home) image of (top) PacBio reads from a sample sequenced as part of the GDAP project. The reads span a 6kb heterozygous LINE-1 element deletion and show clear depth variation. (bottom) Illumina reads from the same sample unable to be clearly mapped around the deletion with reads in white indicating where reads were unable to be uniquely mapped.

148x83mm (300 x 300 DPI)

Downloaded from https://academic.oup.com/hmg/advance-article-abstract/doi/10.1093/hmg/ddy177/4996216
by Wellcome Trust Centre for HG user
on 24 May 2018