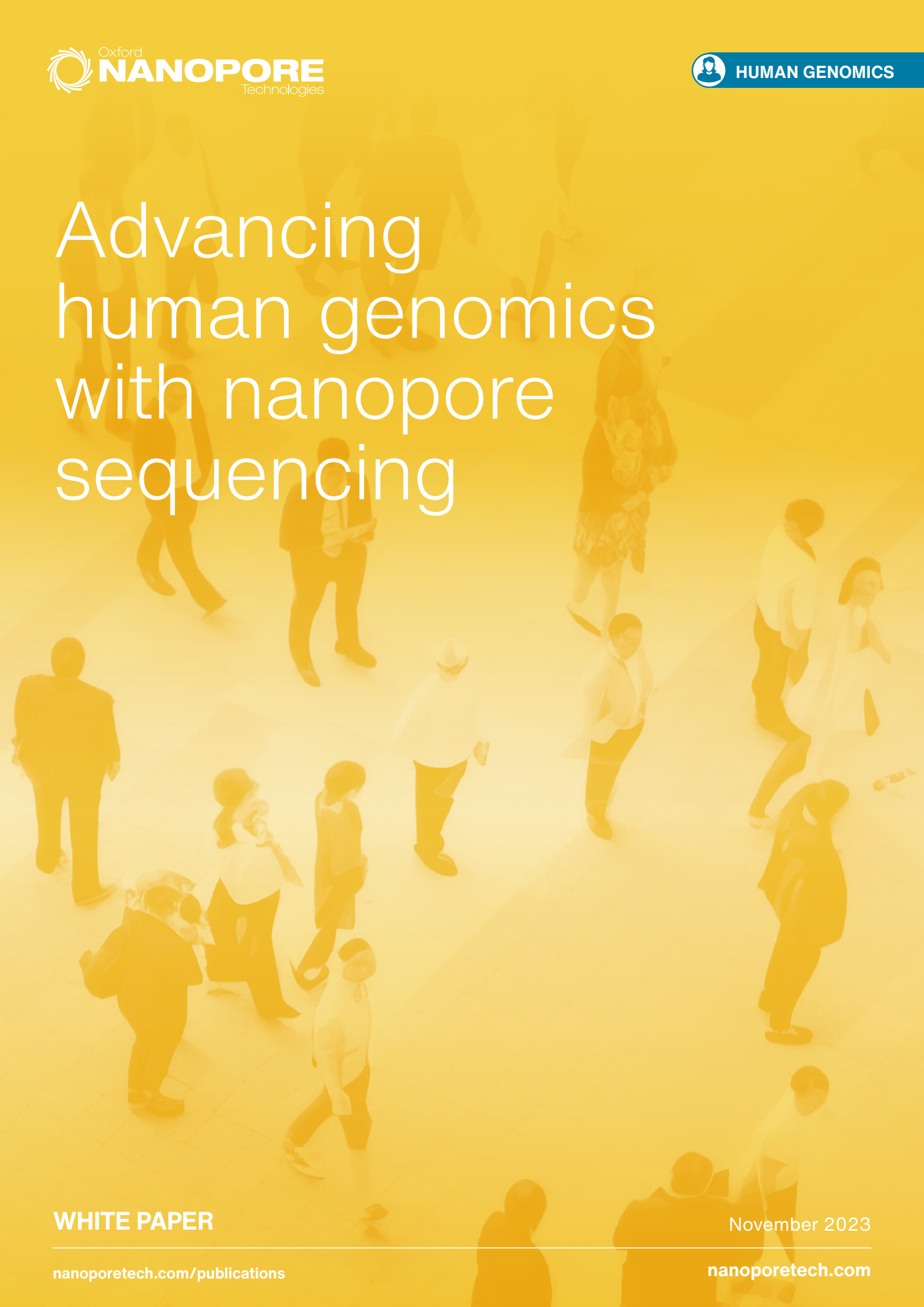
Oxford NANOPORE Technologies

HUMAN GENOMICS

Advancing human genomics with nanopore sequencing



WHITE PAPER

November 2023

nanoporetech.com/publications

nanoporetech.com

Contents

Introduction

High-throughput sequencing technologies have revolutionised the field of human genomics, allowing researchers to more easily investigate and understand biological processes and their impact. Using these technologies, researchers can analyse entire genomes or specific targeted regions of interest, with further functional insights garnered through the characterisation and quantification of RNA transcripts and isoforms. Together, these capabilities have provided unprecedented insight into human genetic diversity and its implications in health and disease.

Despite offering significant advancement in terms of speed and resolution over the older techniques of Sanger sequencing and microarrays, respectively, there are still a number of limitations inherent to traditional short-read, high-throughput sequencing platforms. From closing genome gaps to characterising full-length transcripts, this review will present how long and ultra-long reads, delivered by nanopore sequencing technology are being used to address these limitations, resulting in new biological insights. Specific case studies reveal how researchers are applying the benefits of nanopore technology to a variety of sequencing techniques, including whole-genome, targeted, and RNA sequencing.

The advantages of nanopore sequencing for human genomics

Whole-genome sequencing without the holes

To date, the majority of large genomes sequenced have utilised short-read technologies, which require the DNA to be heavily fragmented prior to sequencing. While these technologies allow rapid and relatively cost-efficient genome analysis compared to older sequencing methodologies, their inherent read-length limitations (typically 150–300 bp) preclude the analysis of vast stretches of DNA corresponding to repetitive regions and large structural variants (SVs)[1,2,3].

The required DNA fragmentation step effectively loses relative positional information, meaning that during genome assembly, these small fragments must be pieced back together through overlap with other short fragments. For repetitive regions, this can be particularly challenging and may result in the collapsing of potentially long regions of repeats down to much shorter lengths, leaving gaps in the assembly (**Figure 1**)[4]. It is for these reasons that there are still an estimated 36,794 regions of the human genome that are intractable to assembly and interpretation using short-read sequencing technologies. These so-called 'dark' regions include protein coding exons from 748 genes, some of which are in important disease-associated loci[5].

> ***'Oxford Nanopore's long-read sequencing capability creates a window into parts of the genome that have been out of reach, as well as giving us a much better handle on structural variants that confer risk of a wide variety of diseases'***
>
> **Kári Stefánsson, CEO, deCODE Genetics[6]**

In the same way, the presence of some SVs, such as deletions, insertions, duplications, inversions, and translocations may be missed when using short sequencing reads alone. It is well established that some repeat regions and SVs are associated with human health and disease (e.g. ageing[7], tandem repeat disorders[8], autism[9], epilepsy[10], and cancer[11]), and with large SVs being 30 times more likely to affect gene expression than single nucleotide variants (SNVs), their routine characterisation is essential[12].

Unlike short-read sequencing technologies, nanopore sequencing processes the entire length of the DNA fragment that is presented to the pore. Complete fragments of thousands of kilobases are routinely processed and ultra-long read lengths over 4 Mb have been achieved[13]. Clearly, such long reads are more likely to span entire regions of repetitive DNA and SVs. As a result, nanopore sequencing provides a more complete view of genetic variation, detecting up to five-fold more SVs when compared to short-read technologies[14].

To support the annotation of human variants, Oxford Nanopore offers a streamlined best practice data analysis workflow, EPI2ME™ wf-human-variation, which incorporates SV, single nucleotide polymorphisms (SNPs), copy number variation (CNV), and short tandem repeat (STR) expansion genotyping.

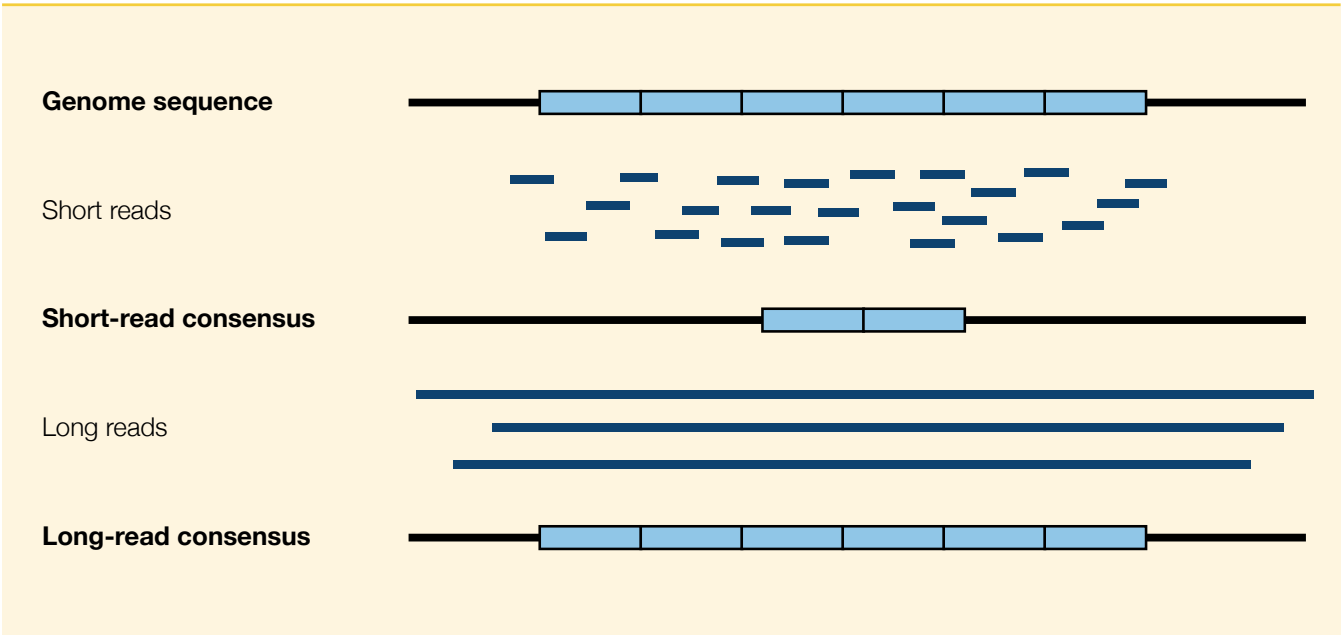


Figure 1
Schematic highlighting the advantages of long reads in *de novo* assembly of repetitive regions. Long read lengths are more likely to incorporate the whole repetitive region (blue boxes) allowing more accurate assembly. Image adapted from Kellog[18].

For the latest information on nanopore data analysis tools, including analysis tutorials, visit nanoporetech.com/analyse.

Using nanopore sequencing, highly contiguous genome assemblies have been generated for many large and complex organisms that had previously been deemed inaccessible to modern sequencing methods[15,16]. Nanopore technology was essential to the generation of the first fully complete human genome assembly for the cell line CHM13, where ultra-long nanopore sequencing reads facilitated the resolution of highly repetitive regions such as telomeres, centromeres, and chromosome arms, which were previously intractable to alternative technologies[17].

The capacity to generate ever more complete and accurate human genome assemblies allows — for the first time — the opportunity to capture the true genetic diversity of the human species through genome sequencing of large cohorts representing diverse populations to generate pangenome references (see Case study 1).

Applying nanopore sequencing to even larger cohorts has allowed researchers to rapidly and comprehensively identify SNVs, SVs, and haplotypes with potential clinical significance in neurodegenerative disease across thousands of clinical research samples in genome-wide association studies (see Case study 2). Furthermore, the long, direct sequencing reads delivered by nanopore technology are also permitting researchers to identify modified bases (e.g. methylation) alongside nucleotide sequence, in a single assay (Case studies 2 and 3).

Rapid analysis of targeted regions

For researchers wishing to study specific genomic loci, a targeted sequencing approach is commonly employed. The ability to focus on the regions most likely to provide relevant data reduces cost, allows a higher depth of coverage, and simplifies analysis.

A range of targeted sequencing methodologies are available with nanopore sequencing, from bespoke assays to targeted panels and whole exome approaches. Nanopore sequencing has been successfully utilised with both PCR and hybrid capture-based target enrichment strategies.

As discussed for whole-genome sequencing, the long reads provided by nanopore technology offer a range of advantages for researchers interested in targeted sequencing. For example, it is possible to sequence much larger regions in a single read, which allows improved characterisation of highly repetitive regions and SVs. Furthermore, long sequencing reads allow the phasing of alleles and variants, which is extremely challenging when using short-read sequencing[19].

> ***'The challenge of aligning short reads to regions with high homology is often not fully appreciated'***
>
> **Leija-Salazar, M. *et al.*[19]**

Nanopore sequencing also offers the unique capability of adaptive sampling, an on-device target enrichment methodology that allows researchers to target specific genomic regions without upfront wet lab sample enrichment.

Adaptive sampling leverages real-time basecalling to selectively sequence or reject DNA molecules as they pass through the nanopore. Rejection of molecules that fall outside of the chosen criteria (e.g. non-homology to target sequence) is achieved through reversing the current applied to the individual nanopore, thereby making it available to sequence an alternative DNA fragment. By removing the need for sample amplification, genomic regions resistant to amplification due to sequence composition can be readily analysed. Furthermore, base modification information can be captured and called alongside the nucleotide sequencing in a single assay (see Analysis of base modifications, page 9).

Many researchers are now assessing the utility of adaptive sampling to provide rapid, comprehensive, and cost-effective genomic characterisation of disease-linked genes[20,21].

Find out more about applying adaptive sampling to your research at nanoporetech.com/adaptive-sampling.

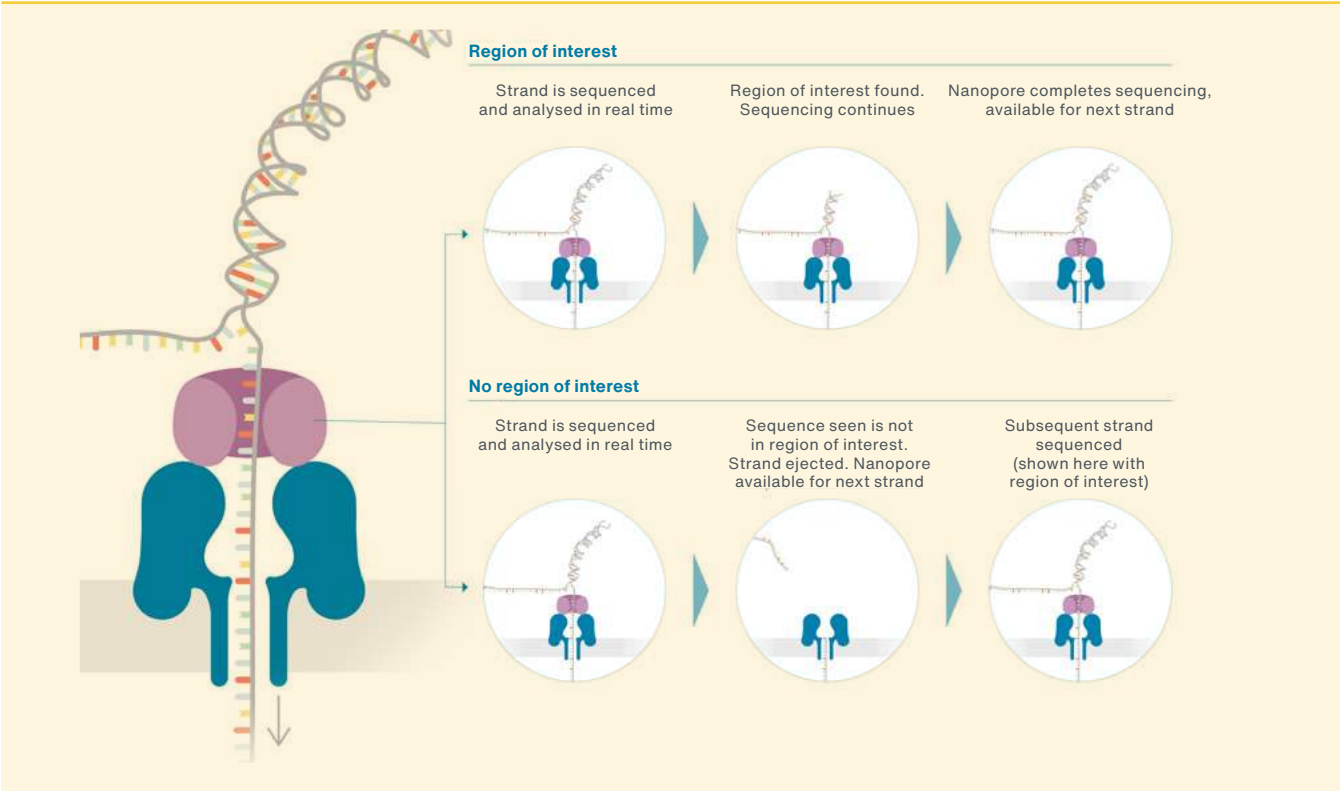


Figure 2
Adaptive sampling is a unique, on-device approach to targeted sequencing that requires no upfront library enrichment steps. Using real-time basecalling, DNA fragments can be accepted or rejected for further sequencing based on their initial sequence composition. A flexible list of regions to be enriched or rejected provided as a .bed file is required. Adaptive sampling can be implemented in advance of, or even during, a run to increase coverage of specific targets.

Full-length RNA transcripts, isoform characterisation, and accurate quantification

Due to the fragmentation required by most traditional sequencing methods, accurate assembly of complete transcripts is exceptionally challenging, especially in instances where a read maps to more than one location. With nanopore technology, entire RNA molecules are processed regardless of their length, allowing the sequencing of complete transcripts in single reads (Case studies 4 and 5). The advantages of nanopore technology are now also being applied to single-cell transcriptome studies, revealing the transcriptomic heterogeneity in health and disease at the cellular level (Case study 5).

In addition to reducing multiple-locus alignment issues, long, full-length nanopore sequencing reads provide a significant advantage in the analysis and correct identification of alternative splicing (**Figure 3**). When using short reads, different transcript isoforms must be computationally reconstructed; however, a study by Steijger *et al.*[22] revealed that automated transcript assembly methods failed to identify all constituent exons in over half of the transcripts analysed. Furthermore, of those transcripts with all exons identified, over half were incorrectly assembled. Such complications are further compounded where reads from highly similar transcripts, such as those of paralogous genes, are under investigation. Rare isoforms could also remain altogether undetected[23].

Utilising long RNA sequencing reads, a team of researchers from New York Genome Centre and The Broad Institute, USA, recently conducted the largest transcriptomic study to date. This study identified 93,718 alternative RNA transcripts, with 77% of these being novel transcript structures. The long reads obtained in this study facilitated further analyses of the effect of rare variants caused by SNPs, providing a better understanding of the exact molecular events associated with genetic variants that potentially contribute to diseases and traits (see Case study 4).

> ***'...we anticipate that a high-resolution characterization of the transcriptome with long-read data will be an important approach for the discovery of regulatory mechanisms of disease-associated variants'***
>
> **Glinos, D. *et al.*[47]**

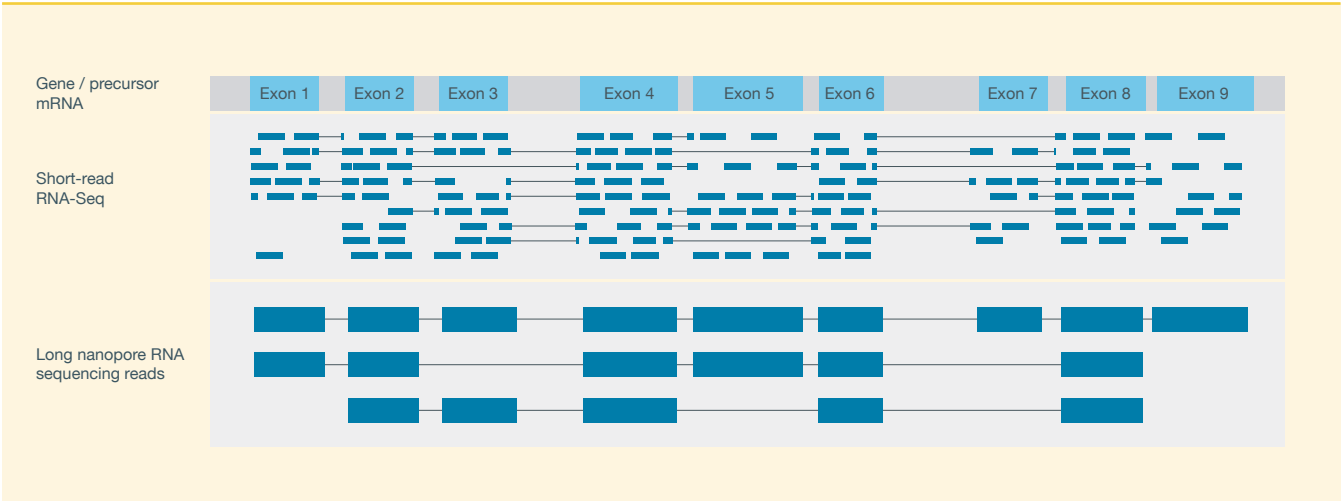


Figure 3
Alternative splicing can give rise to numerous mRNA isoforms per gene, which in turn can alter protein composition and function. The short reads generated by traditional RNA sequencing techniques lose positional information, making the correct assembly of alternative mRNA isoforms challenging. Long nanopore reads can span full-length transcripts, simplifying their identification.

Direct RNA sequencing

Until recently, sequence-based analysis of RNA required the conversion of RNA to complementary DNA (cDNA), a process that can introduce bias through reverse transcription or amplification.

These issues can be exacerbated by the use of traditional short-read sequencing technologies, which are known to exhibit GC bias, where sequences with low or high levels of GC content are underrepresented (**Figure 4**). The amplification step required to generate cDNA also results in the loss of all modified base information. Such base modifications are known to have a role in modulating the activity and stability of RNA and are therefore of increasing interest to researchers. Nanopore sequencing overcomes all of these challenges through the facility for direct RNA sequencing — delivering unbiased, full-length, strand-specific RNA sequences[24]. Using nanopore technology, full-length transcripts over 20 kb in length can be sequenced in a single read[25].

A further benefit of direct RNA sequencing is the ability to accurately measure poly-A tail length[26]. In eukaryotes, messenger RNA (mRNA) is augmented with a series of adenosine bases at the 3' end known as the poly-A tail. These tails can vary in size, with the largest being over 250 nucleotides in length and therefore beyond the typical analysis capabilities of short-read sequencing technologies[26]. Research suggests that poly-A tail length is an important factor in post-transcriptional regulation and further study may provide new insights into gene expression and disease[26].

> ***'Deep long-read RNA sequencing will be necessary going forward to fully comprehend the clinical relevance of individual isoforms for a "single" gene'***
>
> **Aguzzoli Heberle, B. *et al*[45]**

Utilising nanopore sequencing, researchers demonstrated the utility of nanopore direct RNA sequencing in quantifying RNA and poly-A tail abundance, whilst also characterising poly-A tail composition. Their results revealed mRNA tail length to be dynamically regulated at an isoform-specific level during invertebrate embryogenesis, providing key insights into developmental post-transcriptional regulation in a simple, robust, and accurate workflow[27].

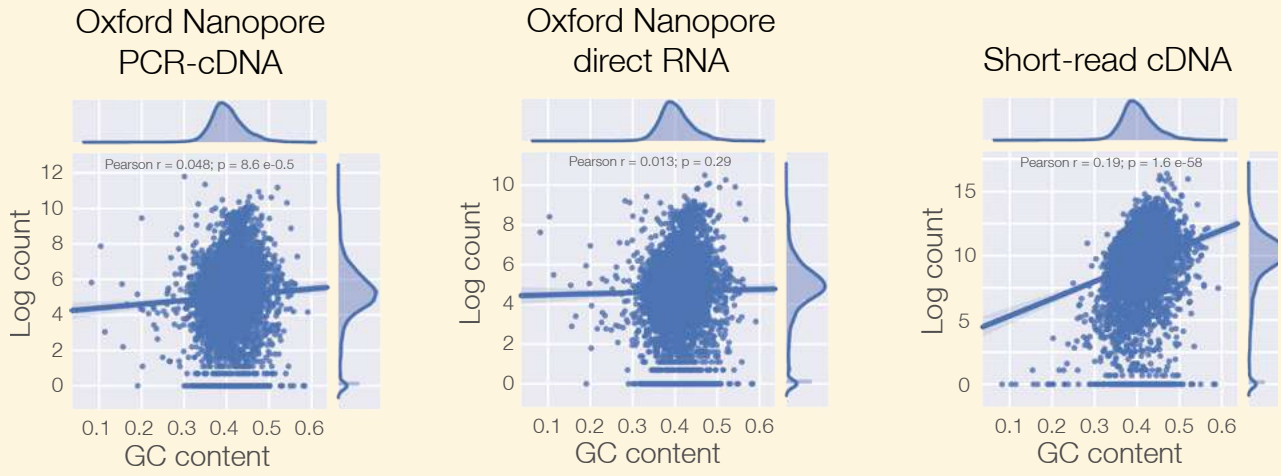


Figure 4
Sequencing workflows that incorporate amplification are vulnerable to sequence-specific biases. Yeast transcriptome libraries were prepared using two nanopore sequencing techniques (cDNA-PCR and direct RNA) and a typical short-read cDNA technique. In all cases, GC bias in the nanopore data sets was lower than in the short-read data set[24].

Analysis of base modifications

The importance of base modifications such as 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC), and N6-methyladenosine (m6A) on gene expression and function is becoming increasingly apparent. For example, 5mC has been linked to many human diseases, including neurological disorders[28] and cancer[29], and may offer significant potential as a diagnostic and prognostic indicator.

The requirement for nucleic acid amplification in traditional short-read sequencing technology erases these base modifications, meaning they cannot be detected without additional time-consuming and often inefficient sample processing methods[30,11]. In addition, short-read based epigenetic studies omit important genomic regions that are recalcitrant to amplification, including centromeres, tandem repeats, and chromosome arms[31].

Nanopore sequencing does not require amplification or strand synthesis, allowing both the base and its modification to be detected in the same sequencing run, with 5mC and 5hmC base modification calling integrated into the MinKNOW™ operating software. Unlike array-based methylation analysis technologies, which typically survey up to 850,000 CpG sites across the human genome, nanopore sequencing can access all 28 million CpG sites[32]. To date, researchers have utilised nanopore sequencing to detect a number of modified bases from both DNA and RNA, including 5mC[33,34], 5hmC[35], m6A[33], pseudouridine[30], and 7-methylguanosine (m7G)[30].

> ***'Nanopore can detect a significant number of sites that short-read assays miss, many of which occur in promoter and exonic regions that are potentially of biological significance.'***
>
> **Foox, J. *et al.*[36]**

Recently, researchers demonstrated the utility of ultra-long nanopore sequencing reads in base modification detection by generating the most complete human methylomes to date (Case study 3). Utilising the ultra-long nanopore sequencing reads of native DNA, the researchers were able to characterise previously inaccessible repeat-rich regions of the genome in tandem with base modification calling of the same dataset[31]. The ability to characterise base modification and sequencing with a single technology offers a time and cost-efficient alternative to the traditional methodologies.

Phasing of variants

As researchers search for greater understanding of genome function, haplotype phasing (determining the maternal or paternal inheritance of an allele) can deliver important insights into health, disease, and evolution. Phasing can also resolve compound heterozygosity, where both alleles of a specific gene or region possess a different variant, which has increasing applications in research for the advancement of precision medicine[37]. However, the nature of short-read sequencing technologies has made it extremely challenging to unambiguously assign parental origin for variants separated over large genomic regions.

> ***'we show that it is possible to achieve state-of-the-art small and structural [variant] calling performance using only [Oxford Nanopore] reads produced by a single flow cell at high throughput'***
>
> **Kolmogorov, M. *et al.*[34]**

Researchers at the UC Santa Cruz Genomics Institute demonstrated how nanopore sequencing can deliver human genome assemblies with megabase-scale phase blocks, including fully phased methylation calls[34]. Furthermore, the nanopore workflow developed by the researchers offered improved SV calling accuracies and superior SNP calling accuracies when compared to short-read sequencing technology. The nanopore sequencing workflow developed for this research is currently being used to sequence thousands of genomes as part of the National Institutes of Health (NIH) Center for Alzheimer's and Related Dementias (CARD) initiative (see Case study 2).

The long phase blocks delivered by nanopore sequencing further open the possibility of using maternal and paternal methylation markers to trace inherited variants of interest to the parent of origin without reliance on parental sequencing, which is often not possible and can incur additional expense[38].

Cost-effective, scalable, and on-demand analysis in real time

Oxford Nanopore provides a range of devices that provide cost-effective, fully scalable, and on-demand sequencing to suit all research requirements (**Figure 5**).

The PromethION™ family of devices offer the highest sequencing outputs, ideal for high-depth analysis of whole human genomes and characterisation of transcripts from single-cell transcriptomes. Running up to 2, 24, or 48 high-output flow cells, respectively, the PromethION 2, PromethION 24, and PromethION 48 devices provide the ultimate sequencing power and flexibility.

The ultra-high throughput capacity of the PromethION 24 and PromethION 48 devices make them particularly applicable to large cohort studies requiring population-scale sequencing data generation.

> ***'Using a single PromethION flow cell, we can detect SNPs with F1-score better than short-read sequencing.'***
>
> **Kolmogorov, M. *et al*[34]**

Each PromethION Flow Cell typically delivers 100–200 Gb* of sequencing data and can be run independently or in parallel.

Unlike traditional sequencing platforms that can require large capital investments, researchers can access the benefits of nanopore sequencing from $1,000 for a MinION™ Starter Packs (including MinION device, a flow cell, and sequencing reagents). MinION Flow Cells typically deliver up to 15–35 Gb* of data, ideal for targeted sequencing applications, low-coverage whole-genome analysis, and isoform-level gene expression studies (**Figure 5**).

The GridION™ device can run up to five MinION Flow Cells independently or in parallel, offering a flexible, on-demand platform suitable and affordable for any lab.

For smaller, more frequently performed analyses, Oxford Nanopore offers Flongle™, a low-cost flow cell adapter for MinION and GridION devices, which delivers up to 1–2 Gb* of data. Flongle Flow Cells also offer a low-cost method of checking library quality prior to running large-scale studies.

A range of streamlined library preparation kits are available, including the facility for sample multiplexing to increase throughput and reduce costs. Oxford Nanopore library preparation kits are highly suited to workflow automation and scripts are available for many automation platforms.

For a detailed comparison of nanopore devices and purchase options, visit **nanoporetech.com/products/specifications.**

Figure 5
Oxford Nanopore sequencing platforms (from left to right): Flongle, a flow cell adapter for MinION and GridION devices; the portable MinION Mk1B; GridION, with capacity for five Flongle or MinION Flow Cells; PromethION 2 Solo; PromethION 2 Integrated; and the high-throughput PromethION 24 and 48 devices.

* Flow cell output will vary according to library type, run conditions, and other sample-specific factors. For more information on data output, visit **nanoporetech.com/products/specifications.**

Case studies

CASE STUDY 1

Capturing global genomic diversity in the human pangenome

While the human reference genome has been a remarkable enabling resource for the scientific community for the past 20 years, it has always been fundamentally limited by its representation of such a small proportion of human genetic diversity. For applications that require comparing an individual's genome sequence or genetic variants to the reference, the interpretation process can falter for anyone with significantly different ancestry than the small number of samples used in building the reference genome. Indeed, while the current reference (GRCh38) includes data from about 20 people, the majority of it is based on a single person's genome.

To overcome this issue, the Human Pangenome Reference Consortium (HPRC) recently published the first draft of a new, graph-based approach to a reference: the pangenome, which represents a broader range of genetic diversity through the inclusion of sequence data from people of many different ancestries[39].

While the consortium's work continues — they envision expanding this effort to cover 350 individuals in time — the first draft of the pangenome contains phased diploid assemblies from 47 people with diverse backgrounds (**Figure 6**). The project incorporated nanopore data and added 119 Mb of sequence data, including 90 Mb of structural variation, to the GRCh38 human reference genome.

The consortium deployed a number of technologies, including short-read sequencing, long-read sequencing, and optical maps. Collectively, all of this data resulted in assemblies representing more than 99% of the genome sequence with greater than 99% accuracy, measured at both the base-pair level and the structural level.

The team used the PromethION sequencing device from Oxford Nanopore Technologies to generate ultra-long reads from each of the 47 samples. For the 29 samples sequenced entirely by the consortium, the average read length was 28.4 kb; for

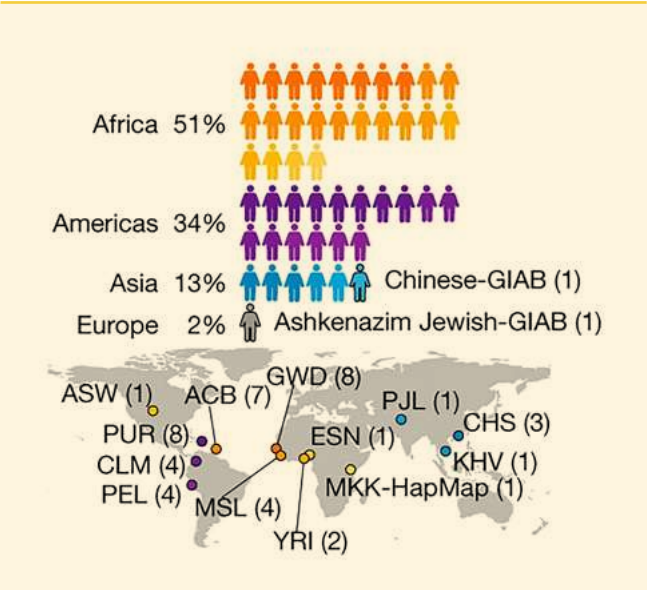


Figure 6
The new pangenome reference was built with genomes from 47 individuals representing diverse genetic backgrounds. Image taken from Liao *et al*[39]. and available under Creative Commons license (creativecommons.org/licenses/by/4.0).

'The human pangenome reference will enable us to represent tens of thousands of novel genomic variants in regions of the genome that were previously inaccessible'

Wen-Wei Liao, Yale University School of Medicine[40]

18 additional samples with prior sequencing data from other sources, the nanopore read length N50 value was approximately 44 kb.

The team compared the results of nanopore sequencing data to data from another long-read sequencing platform. For the other platform, they generated about 40x genome coverage with a read length N50 of 19.6 kb. The same samples were sequenced with a nanopore device to coverage ranging from 10.5-fold to 43-fold. Based on an assessment of quality and mapping accuracy, the team found that while both approaches led to reliable data, the nanopore sequencing data required less coverage to achieve the same level of accuracy.

One of the goals in generating the pangenome was to improve genome interpretation, including variant calls, based on comparing new sequence data to this novel reference. To that end, the scientists assessed the comparison process and evaluated any computational challenges involved in dealing with the more complex graph-based assembly. They reported that '*making the switch to using pangenome mapping is not significantly more computationally expensive and resulted in an average 34% reduction in false-positive and false-negative [small variant] errors compared with using the standard reference methods*'. They further highlighted how '*pangenomes not only improve variant calling but also improve transcript mapping accuracy and detection of ChIP-seq peaks*'.

The pangenome also facilitates analysis of SVs, elements that are typically too long or too complex to be accurately represented in short-read sequence data. Sequencing much longer stretches of DNA in each read, such as through nanopore sequencing, better captures full SVs for downstream analysis, with recent sequencing efforts routinely discovering around 25,000 SVs per human genome[5]. Using a pangenomic method known as PanGenie[41] to analyse short-read data against the nanopore-enabled pangenome, scientists demonstrated that an average of 18,500 SVs could be genotyped in each sample, revealing how long nanopore sequencing reads can further enhance the utility of existing short-read datasets.

Critically, the pangenome is an important resource for addressing the lack of representation of global diversity in clinical research, enabling the characterisation of potentially significant variants in previously underrepresented populations. Already, the resource has been shown to improve read mapping and the calling of small variants. At the SV level, those improvements could be more significant still. The authors noted how '*the pangenome might improve SV genotyping differently across individuals owing to the stronger divergence of the alleles from the reference*' and that '*in the future, the combination of the pangenome and low-cost long-read sequencing should prove to be a potent combination for comprehensive SV genotyping*'.

As well as expanding the pangenome to a larger, more diverse cohort in the near future, the group also plan to move towards telomere-to-telomere genomic assemblies — '*to properly represent the entire genome in almost all individuals*'.

'highly accurate haplotype-resolved assemblies enabled us to access previously inaccessible regions, highlighting new forms of genetic variation and providing new insights into mutational processes such as interlocus gene conversion'

Liao, W.W. *et al.*[39]

CASE STUDY 2

Advancing dementia research through comprehensive analysis of large cohorts

Globally, there are more than 55 million people suffering from dementia, with about ten million more cases added each year[42]. Alzheimer's disease is the most common type, accounting for more than half of all dementia cases.

There is a pressing need to characterise the underlying biology of Alzheimer's disease and other forms of dementia, to potentially support efforts to discover and develop new therapies to treat, cure, or even prevent disease. For such complex conditions, large datasets are needed to produce helpful associations between phenotype and relevant genomic mechanisms.

Large-scale studies have been performed in the past using genotyping microarrays, which do not provide single-base resolution, or short-read sequencing, an approach which has been limited by ambiguous read alignment in the challenging genomic regions associated with cognitive function. These regions can include segmental duplications, repeat expansions, and other large genomic elements that can confound analysis based solely on hard-to-map short reads. To overcome this challenge, researchers need a highly scalable sequencing technique that can produce reads long enough to span large genomic elements for their accurate characterisation — while being cost-effective for utilisation in large-cohort studies.

Now, scientists have successfully demonstrated the use of long nanopore reads as an accessible, scalable approach for sequencing hundreds or thousands of samples from diverse populations to support large-scale studies of dementia. At the National Institutes of Health, USA, researchers have launched a sequencing initiative using long nanopore reads to help unravel the biology behind Alzheimer's disease, Lewy body dementia, and frontotemporal dementia. '*We know that these diseases have a big genetic component*', said Kimberley Billingsley, a scientist at the NIH Intramural Center for Alzheimer's and Related Dementias (CARD), in a presentation at the 2023 London Calling conference[43].

The initiative aims to generate deep datasets of SVs and other variants associated with these forms of dementia through nanopore sequencing of 4,000 human brain clinical research samples. With long nanopore reads, the scientists believe they will be able to identify SVs linked to dementia and resolve highly complex regions such as the HLA locus and the *APOE* gene, which has been linked to an increased risk of Alzheimer's disease.

To pave the way for sequencing thousands of samples, the CARD team first optimised protocols for sample preparation and sequencing of high molecular-weight DNA, data analysis, and data storage and access to make the results available to the scientific community[34,44]. With these protocols in place, they are now sequencing about 200 clinical research samples per month on a PromethION 48 device (**Figure 7**), generating about 30x genome coverage for each sample and achieving read length N50s of about 30 kb.

The CARD computational pipeline that was honed for long nanopore reads generates a harmonised VCF file containing small variants and SVs. Comparing the

Figure 7
With the facility to run up to 48 high-output PromethION Flow Cells independently or in parallel, the PromethION 48 device delivers flexible access to terabases of data, ideal for population-scale sequencing projects.

long nanopore reads with data from a short-read sequencing technology, Kimberley described how the nanopore data '*has a reduced SNP error rate ... especially in those low-mappability regions*'[43].

The F1 score for SNP detection — a combined measure of precision and recall for variants — was higher for nanopore data than for data from a short-read technology. For SV discovery, the nanopore F1 score was comparable to that of another long-read sequencing platform — but '*with a lower cost and higher throughput*'.

The CARD team sequenced 222 control samples from frontal cortex research specimens. While deeper investigations are ongoing, a preliminary analysis detected more than 80,000 SVs. Most were insertions or deletions, and most represented rare events.

> ***'Nanopore has a reduced SNP error rate compared to [a short-read sequencing technology], especially in those low-mappability regions'***
> **Kimberley Billingsley, NIH Center for Alzheimer's and Related Dementias[43]**

In addition to overcoming alignment ambiguity with long reads, nanopore sequencing offers unique features that CARD scientists are using to comprehensively characterise their samples. As nanopore sequencing does not require PCR, it is possible to directly detect epigenetic modifications in native DNA, enabling the characterisation of both genomic variants and methylation from the same dataset — without the need for any additional library preparation steps. Kimberley said that *'with this data, we can start to differentiate and visualise*

differences in methylation'[43]. Making use of this, the CARD team has generated haplotype-specific and cell type-specific methylation profiling data from the brain clinical research samples and cell lines representing neurons and microglia. The team's computational pipeline can automatically produce *de novo* assemblies, variant files, and methylation calls for all samples analysed.

Utilising the unique accessibility and scalability of nanopore sequencing, the CARD researchers were able to generate very large amounts of data for one brain sample as part of their protocol optimisation work. With 400-fold coverage of the genome from a frontal cortex research sample, the team opted to produce high coverage for other regions of the brain as well. Now, they have a high-definition dataset with 70-fold nanopore sequencing coverage of research samples of the parietal cortex and primary visual cortex, along with 800-fold coverage of the cerebellum. The resource could be useful for understanding somatic as well as germline variants and for querying methylation patterns across regions of the human brain. In an analysis of one region, Kimberley highlighted that *'we've been able to successfully detect low-frequency variants'*[43].

With an optimised pipeline in place, Kimberley described how CARD scientists are now looking to sequence hundreds of samples per month and are targeting biobanks with diverse clinical research samples so they can begin *'sequencing more diverse populations'*. The first group of about 200 clinical research samples came from individuals of European descent; the next 150 samples will come from people of African descent. The team will continue to seek out samples representing diverse ancestries as the project progresses.

In conclusion, Kimberley described how she and her colleagues have *'developed an efficient and scalable wet lab and computational protocol for nanopore long-read sequencing that [serves as a] genuine alternative to short reads for large-scale genomic projects'*[43].

Recently, Heberle *et al.* demonstrated the utility of long nanopore sequencing reads in assessing the role of RNA isoform expression in Alzheimer's disease[45]. Using one PromethION Flow Cell per sample, the team sequenced total RNA from 12 post-mortem aged human frontal cortices, in which six samples were Alzheimer's disease cases and six were control cases.

Their approach identified 53 novel RNA isoforms from clinically relevant genes, some of which were the most highly expressed isoforms for that particular gene. Heberle and colleagues also identified 1,917 clinically relevant genes in the human frontal cortex expressing multiple isoforms, with 1,018 of these displaying different protein sequences and 98 of them being previously associated with brain-related disease. Further to this, the team found 99 differentially expressed RNA isoforms between Alzheimer's cases and controls, which was hidden when looking at overall gene-level expression. Their research highlights the significant knowledge gaps in RNA isoform diversity and how long nanopore RNA sequencing reads enable more accurate quantification of RNA isoforms and their relevance in human health and disease.

> ***'...we believe that long-read RNAseq will be the best tool to assess RNA expression patterns in complex human diseases and to identify new molecular targets for treatment and diagnosis'***
>
> **Heberle, B.A. *et al.*[45]**

CASE STUDY 3

Generating the *'most complete human methylomes to date'*

Until recently, lingering gaps in the human reference genome have limited the ability to perform comprehensive analyses of epigenetic regulation. It was impossible to fully characterise chromatin structure and gene expression, even with the latest GRCh38 build. As these traits are already known to be linked to diseases, such as schizophrenia, lung cancer, and muscular dystrophy, understanding them has been a priority for many researchers.

However, even for finished regions of the reference genome, the use of traditional short-read sequencing technology to detect methylation has been problematic. Samples must be prepared specially for bisulfite sequencing and run separately for methylation analysis, rather than generating both DNA sequence and methylation data in a single run. In addition, short reads cannot span long-range epigenetic interactions or differentiate between haplotypes. Finally, accurately mapping short-read methylation data for highly repetitive regions has been a significant hurdle.

The 2022 release of the first complete human genome from the Telomere-to-Telomere (T2T) consortium — including 225 Mb of novel sequence data based on ultra-long nanopore reads to close the 8% of the genome previously missing — has given scientists an entirely new foundation for understanding epigenetic regulation[17]. By combining that resource with additional nanopore sequencing data, researchers have produced the most comprehensive evaluation yet of methylation across the human genome[31].

In this study, Gershman *et al.* from Johns Hopkins University, USA, and other institutions presented a detailed view of the human epigenome and their novel findings based on long nanopore reads. They chose to use data generated by nanopore sequencing devices from Oxford Nanopore Technologies because traditional short-read data was a poor fit for the highly repetitive sequence added to the new reference genome. In addition, since methylation status and nucleic acid identity can be gleaned from the same data, nanopore sequencing reduces the time and cost needed for an epigenetic analysis workflow.

> ***'emerging long-read technologies offer sequence lengths capable of spanning infrequent unique markers and provide a direct measurement of the base sequence and epigenetic state on single molecules'***
>
> **Gershman, A. *et al*[31]**

Nanopore-based methylation profiling of the T2T reference offered a clear improvement over both the GRCh38 reference and short-read epigenetic analysis of the T2T genome, '*revealing epigenetic patterning of genomic regions that were previously intractable*'. The researchers found that '*methylation profiles [in the T2T genome] using long-read nanopore data demonstrate an increase in the*

genome coverage (32.8 M compared with 29.17 M in GRCh38, omitting chromosome Y) and surveyed more CpGs (10%, 3.18 M) compared with short-read whole-genome bisulfite sequencing'.

In addition, the long nanopore reads clearly addressed the mappability challenge seen with short-read data. Short reads struggled noticeably with important elements such as segmental duplications and satellite DNA, with 165 Mb of sequence enriched for these elements considered highly unmappable using short reads. In contrast, nanopore reads longer than 50 kb enabled accurate methylation detection in these challenging regions.

To deepen their investigation, the team sequenced both the CHM13 cell line used to create the T2T reference and HG002, a lymphoblast cell line. CHM13 captures an early developmental stage and exhibited hypomethylation, while HG002 is later stage, with most of its genome methylated. The authors reported that '*long-read methylomes of distinctive developmental time points surveyed >99% of CpGs, establishing the CHM13 and HG002 methylomes as the most complete human methylomes to date*'.

The team then focused on particular elements of interest, including allele-specific methylation in X chromosome inactivation. This approach was enabled by the long nanopore reads, which, when paired with the T2T reference, '*confer the ability to explore methylation patterns of single molecules*'. They identified genetic regions as hypermethylated or hypomethylated, using that information to distinguish between active and inactive X chromosomes. Looking at the CHM13 cell line, they found several genes that were improperly regulated by the X inactivation process and posited that this was likely due to a failure of that process in this haploid cell line.

Separately, the team also used this approach for a high-resolution analysis of the DXZ4 satellite array, which is associated with X chromosome inactivation and lacks heterozygous polymorphisms. The researchers suggested that '*because satellite arrays are known to be hypervariable in the human population and linked to several human diseases, these results highlight the importance of long-read single-molecule epigenetic studies for understanding disease pathology*'.

In one final investigation, the team explored the methylation profile of centromeres — the last regions of the genome to be finished by the T2T consortium because of their highly repetitive sequence. They identified a hypomethylated centromeric unit across chromosomes, linking its role in '*kinetochore positioning and epigenetic regulation of chromosome segregation' and analysing it in human samples representing diverse ancestries. They reported that 'when combined with findings in other organisms, e.g., maize and medaka, this suggests that the [centromeric dip region] is a conserved, functionally important feature of complex centromeres across vertebrate and plant lineages*'.

Concluding, the researchers noted that '*this analysis provides a framework with which to investigate the most elusive regions of the human genome, granting insights into epigenetic regulation*'.

CASE STUDY 4

New insights into human transcriptome variation

Variation in RNA transcript structure is a key feature of eukaryotic gene regulation and allows for over 230,000 alternative transcripts to be derived from approximately 63,000 human genes[17]. Variation in transcript isoforms is predominantly mediated by RNA splicing; however, genetic variants within splicing regulatory sequences can also lead to aberrant transcript isoforms that can cause disease[46]. Traditional RNA sequencing approaches, which utilise short sequencing reads that must be computationally reassembled, offer limited application to isoform-level transcript studies.

Delivering high outputs of full-length transcript reads, nanopore sequencing offers a unique solution to this challenge — enabling comprehensive characterisation of transcript isoforms. To assess the application of nanopore technology for allele-specific alternative splicing analyses, researchers from the New York Genome Centre and The Broad Institute, USA, sequenced the transcriptome of 88 Genotype-Tissue Expression (GTEx) tissue and cell line samples using MinION and GridION devices[47].

Using RNA obtained from fibroblast cell lines, the team initially compared two Oxford Nanopore library preparation approaches, namely direct cDNA sequencing and cDNA-PCR sequencing. Given the precious nature of the GTEx cell lines and tissues, and the prioritisation of read depth in the study, the cDNA-PCR Sequencing Kit was selected for all subsequent experiments due to its lower input requirements and higher data output.

> ***'The advent of long-read sequencing technologies offers the opportunity to study the role of genetic variation in transcript structure'***
>
> **Glinos, D. *et al.*** [47]

Following basecalling, minimap2[48] was utilised to align the generated reads to the GRCh38 human genome reference. The team obtained a median of over six million raw reads per sample, with 80% aligning to the reference genome. The median read length was 709 bp and 789 bp for raw and aligned reads, respectively.

The analysis tool FLAIR v1.4[49] was used to quantify, define, and identify novel transcripts. The long nanopore reads allowed the researchers to identify 93,718 full transcript structures across 21,067 genes — the largest transcriptome dataset achieved in a single study (**Figure 8**). Of the 93,718 transcripts identified, 77% displayed alternative novel transcript structures, and of these 51% shared at least one splice junction with existing annotated transcripts, highlighting the utility of long nanopore reads for detailed characterisation of novel, previously overlooked transcript isoforms (**Figure 9**). The researchers further noted that the novel transcripts showed tissue-specific expression, highlighting the potential for characterising tissue-specific gene expression and regulation with long nanopore sequencing reads.

Phasing of the haplotypes was facilitated by the long RNA sequencing reads obtained in this study in combination with pre-existing whole-genome

sequencing (WGS) data, which allowed the investigation of the allele-specific effect of nucleotide variants on transcript structure.

To perform allele-specific expression (ASE) and allele-specific transcript structure (ASTS) analyses, wherein the transcript expression was correlated to other transcripts of the same gene, the researchers developed a toolkit specifically for long reads — long-read allelic analysis (LORALS). These analyses allowed the researchers to investigate the effects of rare variants caused by previously intractable SNPs and provide a better understanding of the exact molecular events associated with genetic variants that potentially contribute to diseases and traits.

Long RNA sequencing reads allowed the characterisation of the effect of rare and common variants on transcript structure, whilst overcoming the limitations of short-read WGS by enabling full-length transcript sequencing and facilitating ASTS and ASE analyses of structure[47]. The researchers stated that '*we anticipate that a high-resolution characterization of the transcriptome with long-read data will be an important approach for the discovery of regulatory mechanisms of disease-associated variants*'[47].

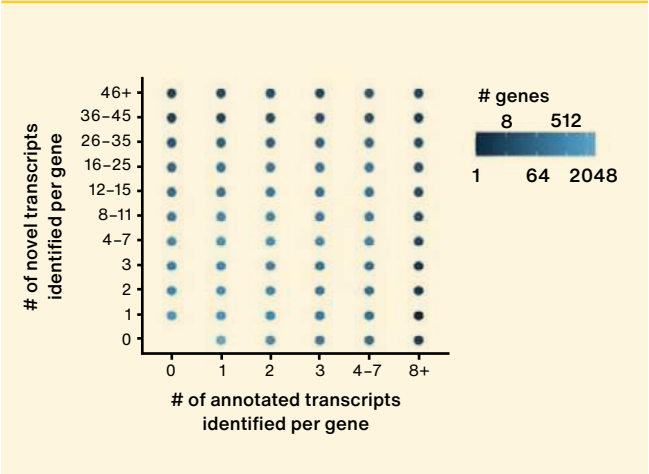


Figure 8
The number of annotated and novel transcripts identified by Glinos *et al*[47]. Over 90,000 novel transcripts were identified, highlighting the utility of long nanopore sequencing reads in generating a more complete transcriptome. More novel transcripts were identified for genes with preexisting annotated transcripts (GENCODE v.26). Figure kindly provided by Dr. Dafni Glinos.

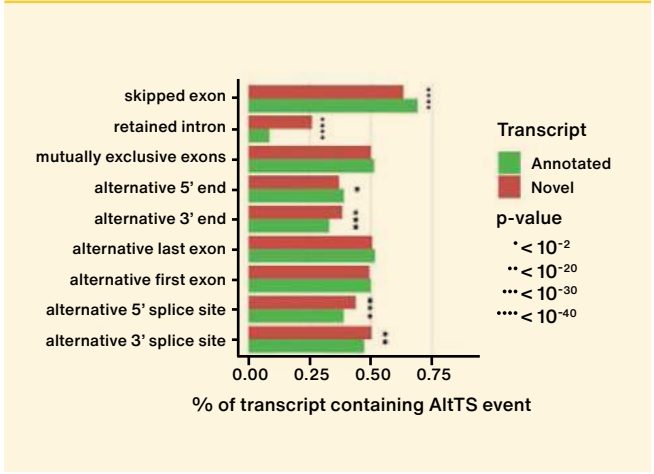


Figure 9
Long RNA sequencing reads facilitated AltTS analyses, which revealed the total number of AltTS events and their proportion in comparison to existing GENCODE v26 annotations. Figure kindly provided by Dr. Dafni Glinos.

CASE STUDY 5

Explore new dimensions of biology with single-cell nanopore sequencing

Single-cell transcriptome sequencing is a powerful tool for profiling transcriptomic heterogeneity in health and disease. However, in the past it has relied on short-read sequencing technologies, which provide little information at the transcript level due to the inherent limitations in linking distal splicing events within the same transcript. With the advent of long-read sequencing technologies, such as that provided by Oxford Nanopore Technologies, we are ushering in a new era of single-cell sequencing, where transcripts can be viewed at the isoform level.

Long nanopore sequencing reads facilitate the pairing of alternative splicing with genetic mutations. This was demonstrated by Dr. Dan Landau's team, based at the New York Genome Centre, USA, who are exploring the relationship between somatic mutations in the splicing factor gene *SF3B1* — which are some of the most prominent drivers in myeloid leukaemias — and their effects on the transcriptome. As a result of aberrant splicing, *SF3B1* mutations manifest in disrupted haematopoietic differentiation, but exactly how this occurred was unknown. The difficulties in studying splice-altering mutations in the myeloid lineage is not least because of the array of progenitor cell types that are present during differentiation, but also because wild type cells and *SF3B1* mutant cells lack any distinguishing cell-surface markers. Also, single-cell analysis with 3'- or 5'-biased short-read sequencing lacks resolution in assessing changes at the isoform level.

To that end, the team developed their single-cell workflow, Genotyping of Transcriptomes-Splice (GoT-Splice); **Figure 10**, for the simultaneous profiling of gene expression, cell surface protein markers, somatic mutation genotyping, and, due to the long nanopore sequencing reads generated on PromethION, they could look at the transcript isoforms within the same single cell[50]. Importantly, they observed a four-fold increase in the number of splice junctions per cell detected using full-length nanopore sequencing, which further '*afforded greater coverage uniformity across the entire transcript, compared to 3'-biased coverage in short-read sequencing*' (**Figure 11**).

They found that *SF3B1* mutations were enriched in cells committed towards erythroid progenitors, which is consistent with the literature on *SF3B1*-driven dyserythropoiesis phenotype. The integration of GoT with nanopore sequencing '*showed that* SF3B1 *mutations exert cell-type specific mis-splicing*'. That is, in the erythroid lineage, there was obvious cell-type-specific 3' cryptic splice site usage in *SF3B1* mutant cells, affecting genes related to the cell cycle[50].

> ***'3' or 5' biased short-read sequencing is limited in its ability to map full-length RNA isoforms and splicing aberrations'***
>
> **Cortés López, M. *et al.*[50]**

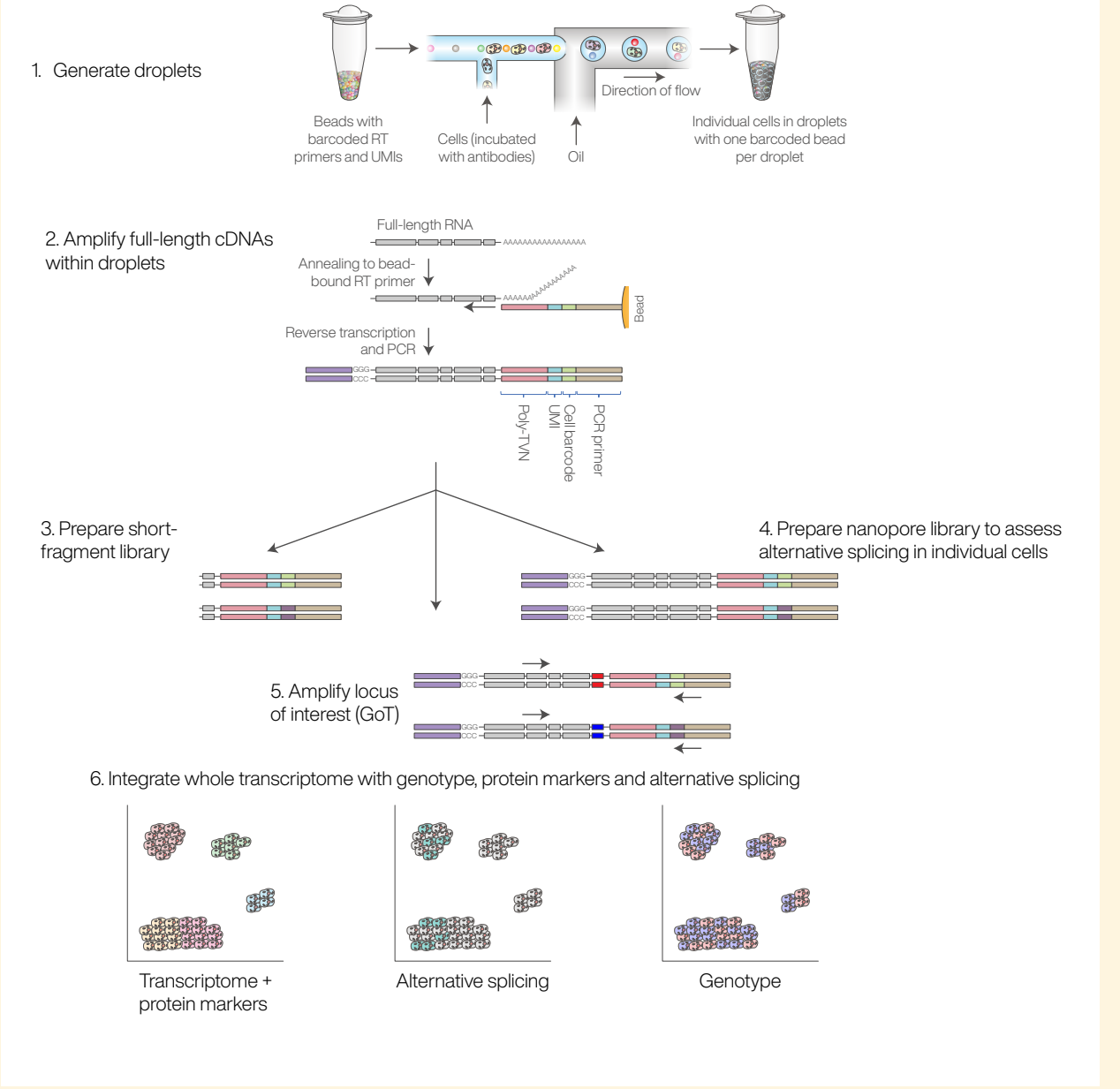


Figure 10

Schematic of GoT-Splice workflow. The combination of Genotyping of Transcriptomes (GoT) with full-length cDNA sequencing using nanopore technology enables the simultaneous profiling of somatic mutations and alternative splicing at single-cell resolution.

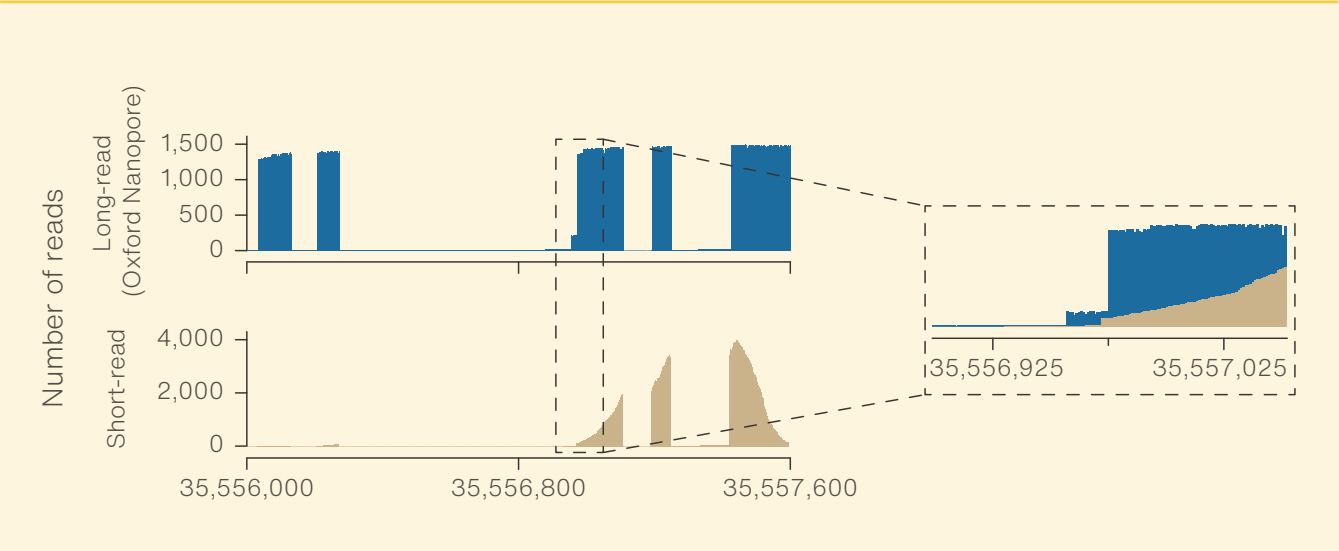


Figure 11
Sequencing coverage plot for the *ERGIC3* gene. GoT-Splice with long nanopore reads allows for a significant increase in the number of junctions captured per cell (blue) and greater sequencing coverage uniformity across splice junctions compared to short-read sequencing (grey).

On the other side of the globe, Dr. Rachel Thijssen from the University of Queensland, Australia, has also been using nanopore sequencing to advance her research into blood cancers, specifically chronic lymphocytic leukaemia (CLL), and the reasons underlying venetoclax drug resistance during relapse[51]. The pharmacologic action of venetoclax is the inhibition of the pro-survival protein BCL2, culminating in apoptosis of the leukaemic cells. Whilst this delivers high remission rates, there is an eventual loss of efficacy. Thijssen *et al.* investigated CLL samples treated with venetoclax monotherapy using a single-cell sequencing approach, to provide an accurate picture of tumour heterogeneity. Like Dr. Landau's team, Dr. Thijssen used long nanopore reads to assess and link mutations to changes in transcriptional read-outs at the single-cell level. They found an array of previously '*unappreciated*' disrupted splicing events that could be attributed to the venetoclax resistance phenotype, including a non-functional transcript of the pro-apoptotic *NOXA* gene. Her findings support the notion that venetoclax therapy should be administered in a time-limited period to prevent the development of full-blown resistance[51].

'when comparing short-read and long-read sequencing, we found a 12.3-fold increase in the number of junctions detected using long-read sequencing, with the majority of junctions (90%) unique to long-read data'

Thijssen, R. *et al.*[51]

Both teams have demonstrated the power of linking somatic mutations to transcriptional changes at the single-cell level, and how long nanopore sequencing reads were pivotal in obtaining full-length transcript isoforms.

Summary

Our knowledge of the human genome, its genes and their function has advanced considerably since the publication of the first human genome sequence in 2003. These advances have been supported by the rapid development of genomic analysis technologies, allowing faster, more detailed, and more affordable genetic analyses. However, the inherent challenges of traditional short-read sequencing technologies limit their ability to fully characterise the whole spectrum of genetic variation.

The long, direct, and real-time sequencing reads offered by nanopore technology delivers a step-change in human genetics research, allowing routine and complete characterisation of highly important genomic events such as structural variation, repetitive regions, phasing, RNA isoforms, and base modifications. Ultra-long nanopore sequencing reads have been shown to be instrumental in finally achieving a complete human reference genome and they are now being applied in large cohort studies, incorporating diverse backgrounds to generate pangenome references that fully capture human genetic diversity.

Using nanopore sequencing, researchers are now unlocking the secrets of the genome — from characterising complete centromeres to discovering new, highly expressed transcript isoforms. As stated by Dr. Karen Miga, Co-chair of the T2T consortium, University of California, USA, '*Using long-read methods, we have made breakthroughs in our understanding of the most difficult, repeat-rich parts of the human genome*'[52].

About Oxford Nanopore Technologies

Oxford Nanopore's goal is to enable the analysis of anything, by anyone, anywhere. The company offers the only sequencing technology to combine scalability — from portable to ultra-high throughput formats — with real-time data delivery and the ability to elucidate accurate, rich biological data through the analysis of short to ultra-long fragments of native DNA or RNA (**Table 1**).

The capacity to generate reads of any length, including long and ultra-long reads in excess of 4 Mb, enables comprehensive analysis of human genetic variation, allowing enhanced characterisation of structural variation, repetitive regions, haplotype phasing, RNA splice variants, isoforms, fusion transcripts, and methylation. These benefits can be applied across whole-genome, whole-transcriptome, and targeted sequencing approaches, allowing researchers complete flexibility in the development of cost-effective genomic analysis workflows.

	Flongle	MinION	GridION	PromethION 2	PromethION 24/48
Read length	Fragment length = read length. Longest read now >4 Mb				
Run time	1 min – 16 hrs	1 min – 72 hrs	1 min – 72 hrs	1 min – 72 hrs	1 min – 72 hrs
Number of flow cells per device	1	1	5	2	24/48
DNA sequencing yield per flow cell*	Up to 2.6 Gb	Up to 48 Gb	Up to 48 Gb	Up to 277 Gb	Up to 277 Gb
DNA sequencing yield per device*	Up to 2.6 Gb	Up to 48 Gb	Up to 240 Gb	Up to 554 Gb	Up to 6.5 Tb/ 13 Tb
Multiplexing	1 – 96 samples	1 – >2,000 samples	1 – >2,000 samples	1 – >2,000 samples	1 – >2,000 samples

* Theoretical max output (TMO). Assumes system is run for 72 hours (or 16 hours for Flongle) at 400 bases/second. Actual output varies according to library type, run conditions, etc. TMO noted may not be available for all applications or all chemistries.

Table 1
A range of nanopore sequencing devices are available to meet the coverage and throughput requirements for all sequencing applications. Data correct at time of print. Visit **nanoporetech.com** for the latest information.

For the latest information about applying nanopore sequencing to your human genetics research, visit nanoporetech.com/human-genetics.

References

1. Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* 36(4):338-345 (2018). DOI: https://doi.org/10.1038/nbt.4060

2. Conrad, D. F. et al. Origins and functional impact of copy number variation in the human genome. *Nature* 464:704–712 (2010). DOI: https://doi.org/10.1038/nature08516

3. Cretu Stancu, M. et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat. Commun.* 8(1):1326 (2017). DOI: https://doi.org/10.1038/s41467-017-01343-4

4. Cao, M.D. et al. Scaffolding and completing genome assemblies in real-time with nanopore sequencing. *Nat. Commun.* 8:14515 (2017). DOI: https://doi.org/10.1038/ncomms14515

5. Ebbert, M.T.W. et al. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Bio.* 20(1):97 (2019). DOI: https://doi.org/10.1186/s13059-019-1707-2

6. Amgen invests £50 Million ($66 Million) in Oxford Nanopore Technologies. https://www.amgen.com/newsroom/press-releases/2018/10/amgen-invests-50-million-66-million-in-oxford-nanopore-technologies (2018) [Accessed: 10 Oct 2023]

7. Cawthon, R. M., Smith, K.R., O'Brien, E., Sivatchenko, A., and Kerber, R.A. Association between telomere length in blood and mortality in people aged 60 years or older. *Lancet* 361(9355): 393–5 (2003). DOI: https://doi.org/10.1016/S0140-6736(03)12384-7

8. Stevanovski, I. et al. Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. *Sci. Adv.* 8(9):eabm 5386 (2022). DOI: https://doi.org/10.1126/sciadv.abm5386

9. Brandler, W.M. et al. Paternally inherited cis-regulatory structural variants are associated with autism. *Science* 360(6386):327-331 (2018). DOI: https://doi.org/10.1126/science.aan2261

10. Ishiura, H. et al. Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. *Nat Genet*. 50(4):581-590 (2018). DOI: https://doi.org/10.1038/s41588-018-0067-2

11. Xu, L. et al. Long-read sequencing identifies novel structural variations in colorectal cancer. *PLoS genetics* 19(2), e1010514 (2023). DOI: https://doi.org/10.1371/journal.pgen.1010514

12. Ebbert, M.T.W. et al. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Bio.*20(1):97 (2019). DOI: https://doi.org/10.1186/s13059-019-1707-2

13. Oxford Nanopore Technologies. Ultra-Long DNA Sequencing Kit. Available at: https://store.nanoporetech.com [Accessed: 10 October 2023]

14. Beyter, D. et al. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet*. 53: 779–786 (2021). DOI: https://doi.org/10.1038/s41588-021-00865-4

15. Meyer, A. et al. Giant lungfish genome elucidates the conquest of land by vertebrates. *Nature* 590(7845), 284–289 (2021). DOI: https://doi.org/10.1038/s41586-021-03198-8

16. Aury, J.M. et al. Long-read and chromosome-scale assembly of the hexaploid wheat genome achieves high resolution for research and breeding. *GigaScience* 11, giac034 (2022). DOI: https://doi.org/10.1093/gigascience/giac034

17. Nurk, S. et al. The complete sequence of a human genome. *Science* 376(6588):44-53 (2022). DOI: https://doi.org/10.1126/science.abj6987

18. Kellog, E.A. Genome sequencing: Long reads for a short plant. *Nat. Plants* 1:15169 (2015). DOI: https://doi.org/10.1038/nplants.2015.169

19. Leija-Salazar, M. et al. Detection of GBA missense mutations and other variants using the Oxford Nanopore MinION. *Mol. Genet. Genomic Med.* 7(3):e564 (2019). DOI: https://doi.org/10.1002/mgg3.564

20. Miller, D.E. et al. Targeted long-read sequencing identifies missing disease-causing variation. *Am. J. Hum. Genet*. 108(8):1436-1449 (2021). DOI: https://doi.org/10.1016/j.ajhg.2021.06.006

21. Yamada, M. et al. Diagnosis of Prader-Willi syndrome and Angelman syndrome by targeted nanopore long-read sequencing. *Eur. J. Med. Genet*. 66(2): 104690 (2023). DOI: https://doi.org/10.1016/j.ejmg.2022.104690

22. Steijger, T. et al. Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* 10:1177–1184 (2013). DOI: https://doi.org/10.1038/nmeth.2714

23. Martin, J.A. and Wang, Z. Next-generation transcriptome assembly. *Nat. Rev. Genet*. 12: 671-682 (2011). DOI: https://doi.org/10.1038/nrg3068

24. Oxford Nanopore Technologies. Low bias RNA-seq: PCR-cDNA, PCR-free direct cDNA and direct RNA sequencing. Poster. Available at: https://nanoporetech.com/resource-centre/quantitative-rna-seq-pcr-cdna-pcr-free-direct-cdna-and-direct-rna-sequencing [Accessed: 16 October 2023]

25. Timp, W. and Jain, M. Direct RNA cDNA sequencing of the human transcriptome. Presentation. Available at: https://nanoporetech.com/resource-centre/direct-rna-cdna-sequencing-human-transcriptome [Accessed: 16 October 2023]

26. Jalkanen, A.L., Coleman, S.J. and Wilusz, J. Determinants and implications of mRNA poly(A) tail size — does this protein make my tail look big? *Semin. Cell Dev. Biol.* 34: 24–32 (2014). DOI: https://doi.org/10.1016/j.semcdb.2014.05.018

27. Begik, O. et al. Nano3P-seq: transcriptome-wide analysis of gene expression and tail dynamics using end-capture nanopore cDNA sequencing. *Nat. Methods* 20: 75–85 (2023). DOI: https://doi.org/10.1038/s41592-022-01714-w

28. Weng, Y.L., An, R., Shin, J., Song, H., and Ming, G.L. DNA modifications and neurological disorders. *Neurotherapeutics* (4):556-67 (2013). DOI: https://doi.org/10.1007/s13311-013-0223-4

29. Kuschel, L.P. et al. Robust methylation-based classification of brain tumours using nanopore sequencing. *Neuropathol. Appl. Neurobiol.* 49(1), e12856 (2023). DOI: https://doi.org/10.1111/nan.12856

30. Smith, A. M., Jain, M., Mulroney, L., Garalde, D. R., & Akeson, M. Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing. *PloS one* 14(5), e0216709 (2019). DOI: https://doi.org/10.1371/journal.pone.0216709

31. Gershman, A. et al. Epigenetic patterns in a complete human genome. *Science* 376(6588), eabj5089 (2022). DOI: https://doi.org/10.1126/science.abj5089

32. Flynn, R. et al. Evaluation of nanopore sequencing for epigenetic epidemiology: a comparison with DNA methylation microarrays. *Hum. Mol. Genet*. 31(18):3181-3190 (2022). DOI: https://doi.org/10.1093/hmg/ddac112

33. Garalde, D.R. et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* 15(3): 201–206 (2018). DOI: https://doi.org/10.1038/nmeth.4577

34. Kolmogorov, M. et al. Scalable nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. *Nat. Methods* 20: 1483–1492 (2023). DOI: https://doi.org/10.1038/s41592-023-01993-x

35. Oxford Nanopore Technologies. Haplotype-resolved multi-modal analysis of cancer genomes using nanopore sequencing. Poster. Available at: https://nanoporetech.com/resource-centre/haplotype-resolved-multi-modal-analysis-cancer-genomes-using-nanopore-sequencing [Accessed: 31 August 2023]

36. Foox, J. et al. The SEQC2 epigenomics quality control (EpiQC) study. *Genome Biol.* 22(1):332 (2021). DOI: https://doi.org/10.1186/s13059-021-02529-2

37. Ammar, R. et al. Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. *F1000Research* 4(17): (2015). DOI: https://doi.org/10.12688/f1000research.6037.2

38. Oxford Nanopore Technologies. Unprecedented access to haplotype-resolved biology enabled by ultra-long reads and Pore-C. Poster. Available at: https://nanoporetech.com/resource-centre/unprecedented-access-haplotype-resolved-biology-enabled-ultra-long-reads-and-pore-c [Accessed: 10 October 2023]

39. Liao, W.W. et al. A draft human pangenome reference. *Nature* 617: 312–324 (2023). DOI: https://doi.org/10.1038/s41586-023-05896-x

40. NIH. Scientists release a new human 'pangenome' reference. https://www.nih.gov/news-events/news-releases/scientists-release-new-human-pangenome-reference (2023) [Accessed: 16 October 2023]

41. Ebler, J. et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet*. 54: 518-525 (2022). DOI: https://doi.org/10.1038/s41588-022-01043-w

42. WHO. Dementia. https://www.who.int/news-room/fact-sheets/detail/dementia [Accessed: 16 October 2023]

43. Billingsley, K.J. Population-scale nanopore sequencing to further understand the genetics of Alzheimer's disease and related dementias. Presentation. Available at: https://nanoporetech.com/resource-centre/london-calling-2023-population-scale-nanopore-sequencing-further-understand [Accessed: 10 October 2023]

44. Billingsley, K.J. et al. Processing human frontal cortex brain tissue for population-scale Oxford Nanopore long-read DNA sequencing SOP V.2. *protocols.io* (2022). DOI: dx.doi.org/10.17504/protocols.io.kxygx3zzog8j/v1

45. Heberle, B.A. et al. Using deep long-read RNAseq in Alzheimer's disease brain to assess clinical relevance of RNA isoform diversity. *bioRxiv* 552162 (2023). DOI: https://doi.org/10.1101/2023.08.06.552162

46. Park, E., Pan, Z., Zhang, Z., Lin, L. and Xing, Y. The expanding landscape of alternative splicing variation in human populations. *Am. J. Hum. Genet*. 102(1):11-26 (2018). DOI: https://doi.org/10.1016/j.ajhg.2017.11.002

47. Glinos, D.A. et al. Transcriptome variation in human tissues revealed by long-read sequencing. *Nature*. 608(7922): 353-359 (2022). DOI: https://doi.org/10.1038/s41586-022-05035-y

48. Heng, L. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 34(18): 3094–3100 (2018). DOI: https://doi.org/10.1093/bioinformatics/bty191

49. Tang, A.D., Soulette, C.M., van Baren, M.J. et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat. Commun*. 11: 1438 (2020). DOI: https://doi.org/10.1038/s41467-020-15171-6

50. Cortés López, M. et al. Single-cell multi-omics defines the cell-type-specific impact of splicing aberrations in human hematopoietic clonal outgrowths. *Cell Stem Cell* 30:1-20 (2023). DOI: https://doi.org/10.1016/j.stem.2023.07.012

51. Thijssen, R. et al. Single-cell multiomics reveal the scale of multilayered adaptations enabling CLL relapse during venetoclax therapy. *Blood* 140(20):2127-2141 (2022). DOI: https://doi.org/10.1182/blood.2022016040

52. NIH. Researchers generate the first complete, gapless sequence of a human genome. https://www.nih.gov/news-events/news-releases/researchers-generate-first-complete-gapless-sequence-human-genome [Accessed: 10 October 2023]

Oxford Nanopore Technologies
phone +44 (0)845 034 7900
email support@nanoporetech.com
twitter @nanopore
www.nanoporetech.com

WP_1025(EN)_V4_01Nov2023