Oxford NANOPORE Technologies

Whole-genome sequencing: large genomes

GETTING STARTED

nanoporetech.com

Introduction

Until recently, large eukaryotic genomes have proved challenging to sequence and assemble with traditional sequencing methods. Producing long and ultra-long reads, high-throughput nanopore sequencing is enhancing large genome assembly, enabling resolution of even the most challenging regions.

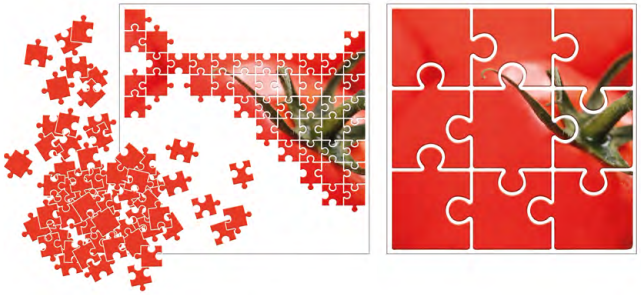
Full characterisation of plant and animal genomes via high-throughput sequencing is revolutionising many areas of research. In human and veterinary health, sequencing enables unprecedented insights into disease, whilst investigating genetic variation in plants allows for the maintenance of diversity and selection of desirable traits, such as high yield and resistance to pathogens. Whole-genome sequencing also helps shed light on the evolutionary relationships between organisms.

Plant and animal genomes are incredibly diverse, with an up to 10,000-fold difference in their lengths: the *Paris japonica* plant genome, at 152 Gb, is almost 50 times larger than the human genome[1]. Ploidy also varies, from diploid up to decaploid. The size and complexity of such genomes has, until recently, limited the capacity to sequence them: of the ~400,000 known plant species, only ~900 (0.2%) have had their genomes sequenced[2,3]. Even some of the most well-studied genomes remain incomplete; until 2020, 8% of the human genome remained unresolved[4], predominantly due to the presence of repeat-rich DNA and large structural variants, which are challenging to resolve via short-read technologies.

Long and ultra-long nanopore sequencing reads allow genomes to be sequenced in fewer, longer fragments, with greater overlap, enabling easier genome assembly — much like building a jigsaw **(see Figure 1)**.

Figure 1. Long nanopore reads simplify genome assembly

Like a jigsaw puzzle with large pieces, long DNA sequencing reads are much easier to assemble than short reads. The tomato genome is approximately 1 Gb in length, which equates to 10 million short reads of 100 bp or 2,000 long reads of 500 kb. The high proportion (60%) of repetitive DNA further complicates assembly when using short sequencing reads.



10 million 'pieces' (short reads)

2,000 'pieces' (long reads)

1. Pellicer, J., Fay, M.F. and Leitch, I. J. The largest eukaryotic genome of them all? Botanical Journal of the Linnean Society. 164(1) (2010).
2. Royal Botanic Gardens Kew. 2020. State of the world's plants and fungi. Available at: https://www.kew.org/science/state-of-the-worlds-plants-and-fungi [Accessed: 07 February 2022].
3. NCBI National Center for Biotechnology Information. Genomes information by organism. Available at: https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/ [Accessed: 07 February 2022].
4. Nurk, S. et al. The complete sequence of a human genome. bioRxiv 445798 (2021).

With nanopore sequencing, there is no upper read length limit; the current record read extends over 4 Mb[5]. Long reads can span structural variants, including insertions, inversions, and duplications, and so greatly enhance their resolution (see **Figure 2**). They also make possible the characterisation of highly repetitive sequences such as repeat expansions, centromeres, and telomeres. In May 2020, the Telomere-to-Telomere (T2T) Consortium published their assembly of the human haploid cell line CHM13, in which they utilised ultra-long nanopore sequencing reads to span repetitive regions that could not be resolved using other technologies, including a 450 kb read that spanned a ribosomal DNA array end-to-end[4]. The group's work represents the first complete sequence of a human genome.

Table 1 describes some of the key benefits of nanopore technology for whole-genome sequencing.

In this guide, we introduce the different approaches and advantages of sequencing large genomes with Oxford Nanopore technology. For information on the sequencing of small genomes, please see the Whole-genome sequencing: small genomes Getting started guide, found in the Resource centre on our website.

Table 1. Advantages of nanopore technology for whole-genome sequencing

Easier assembly
Longer reads, with greater overlap, mean fewer fragments to assemble

Structural variant and repeat resolution
Long reads can span entire structural variants and repeat segments in one read

Phasing
Long reads enhance unambiguous allele phasing

Real-time monitoring
Reads can be basecalled and analysed as sequencing progresses; runs can be stopped once a coverage target is reached

Detection of base modifications
PCR-free nanopore sequencing enables the analysis of epigenetic modifications alongside nucleotide sequence from a single dataset, no special library prep needed

Cost-effective and scalable
A range of sequencing devices are available to suit all project sizes

Absence of GC bias
GC bias is virtually absent in native nanopore sequencing data, meaning greater uniformity of coverage compared to short-read assemblies

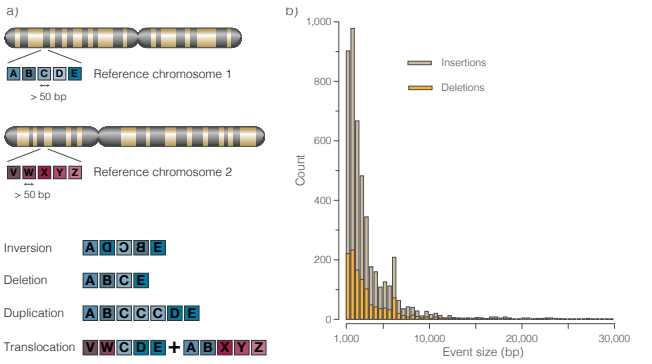


Figure 2. Structural variation a) classes and b) variant size and frequency in the human genome. Image adapted from Huddleston, J. *et al*.[6]

5. Oxford Nanopore Technologies. Ultra-Long DNA Sequencing Kit. Available at: https://store.nanoporetech.com/ultra-long-dna-sequencing-kit.html [Accessed: 07 February 2022].

6. Huddleston, J. et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Research 27(5):677-685 (2016).

Nanopore sequencing devices

Nanopore sequencing devices are scalable to your needs, offering sequencing solutions for genomes of every size. The MinION™ outputs 10s of Gb of data per run, and with its portable design, it can be used anywhere, from the lab bench to a field setting. The flexible GridION™ and ultra-high throughput PromethION™, each with in-built compute, are game changers in large genome sequencing. These devices are ideal for high-output and high-throughput sequencing experiments.

Figure 3 illustrates the size of genomes that can be sequenced to high depth of coverage on each nanopore sequencing device. However, these devices are also highly flexible. As well as enabling the sequencing of viral and bacterial genomes or targeted libraries, the small Flongle™ Flow Cell is also ideal for the QC of large genome libraries prior to sequencing on higher-throughput flow cells. The GridION, with the capacity to sequence on up to five individually addressable MinION Flow Cells, can be used to sequence multiple smaller genomes, or used to full capacity to generate high depth of coverage of a larger genome when needed. Finally, the ultra-high-throughput PromethION devices, available configured for up to 24 or 48 PromethION Flow Cells, enable the on-demand sequencing of large plant and animal genomes to high depth of coverage and the capacity for population-scale projects. With streamlined workflows and real-time analysis, data can be rapidly obtained and even analysed as soon as sequencing starts.

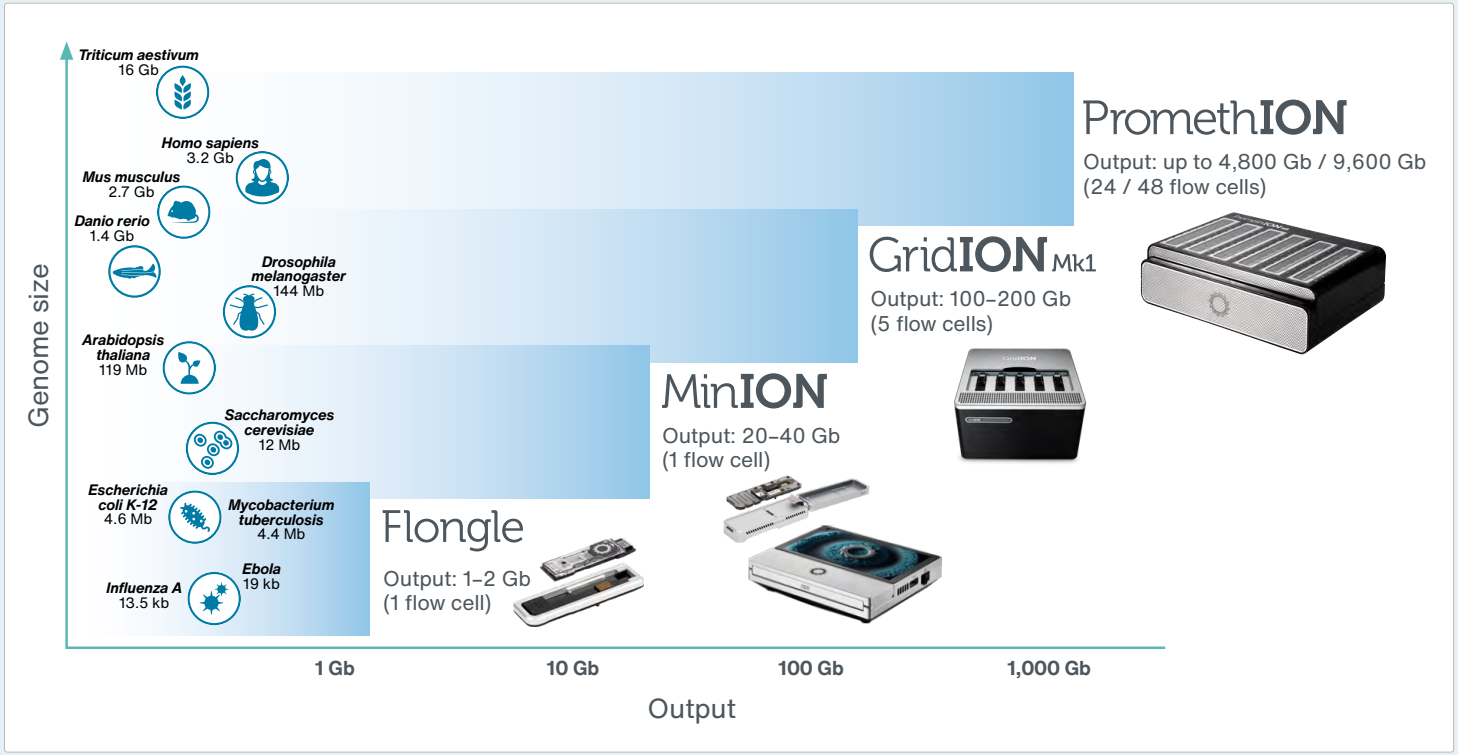


Figure 3. Genome size versus device and output. Outputs may vary according to library type, run conditions, etc.

Large genome assembly: which approach do I choose?

Oxford Nanopore provides end-to-end solutions for multiple genome sequencing approaches. By combining long nanopore sequencing read data with that of ultra-long reads, Pore-C chromatin conformation capture information, or both, a large genome can be assembled and scaffolded to high contiguity and completeness.

Start here

Ligation Sequencing Kit	
Input recommendation	1 µg dsDNA
Preparation time	60 minutes
Read length	Equal to fragment length
Typical throughput	● ● ●
Recommended depth of coverage	≥40x

Optional: increase contiguity and continuity

Ultra-Long DNA Sequencing Kit	
Input recommendation	Find details for your sample type at **store.nanoporetech.com/ultra-long-dna-sequencing-kit.html**
Preparation time	~200 minutes + overnight elution
Read length	50–100+ kb N50, reads up to 4+ Mb
Typical throughput	● ● ○
Recommended depth of coverage	≥20x
See page 8 for more information	

Optional: scaffold for best contiguity

Method: Pore C With: Ligation Sequencing Kit	
Input recommendation	Find details for your sample type at **community.nanoporetech.com/extraction_methods/re-pore-c**
Preparation time	3 days
Read length	~4–5 kb N50*
Typical throughput	● ● ○
Recommended depth of coverage	≥20x
See page 7 for more information	

Highly contiguous, reference-quality assembly

*Read length N50 observed using NlaIII for restriction digestion and SPRI size selection to enrich for fragments >1.5 kb.

From sample to answer

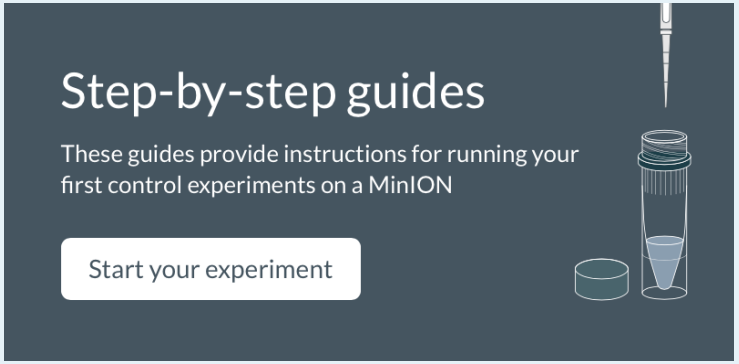
QUESTION

SET-UP

This is my first nanopore sequencing experiment. Where do I start?

Firstly, you will need to set up your sequencing device, download the required software, and then prepare and run a control sequencing experiment. This checks that everything is working as it should, and helps to familiarise you with our library prep and sequencing workflow. Our step-by-step guides take you through this entire process, with easy-to-follow instructions for every step of the way.

View our step-by-step guides:
community.nanoporetech.com/getting_started

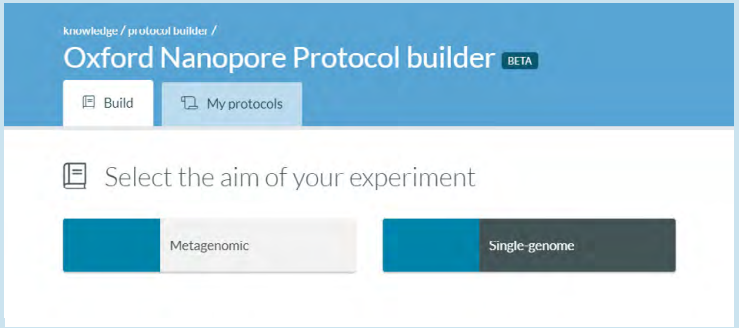


PLANNING

How do I design my protocol?

The Oxford Nanopore Protocol Builder is an interactive tool that enables you to generate your own end-to-end protocol, with application-specific advice encompassing DNA extraction, library prep, sequencing, and data analysis.

Create your bespoke whole-genome sequencing protocol:
community.nanoporetech.com/knowledge/protocol_builder



From sample to answer

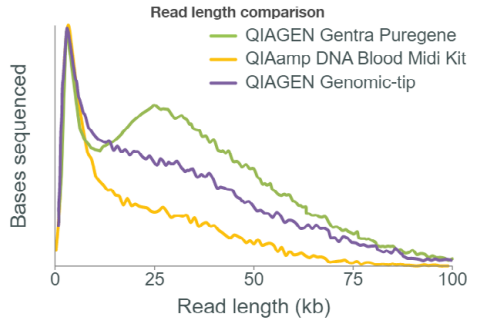
EXTRACTION

How can I best extract high-quality DNA from my sample?

The ‘Prepare’ Documentation section of the Nanopore Community features recommended DNA extraction methods and comparisons for a wide range of sample types, including mammalian, avian, fish, reptilian, and plant samples. It also features data on the effects of carryover of contaminants, such as phenol and ethanol, on library prep efficiency, and guidance on size selection. If performing genome assembly, we would recommend size selecting for long fragments prior to sequencing, to facilitate the downstream assembly process.

Read more about recommended extraction methods for your sample:
community.nanoporetech.com/docs/prepare

Read length comparison of methods for DNA extraction from rabbit blood samples



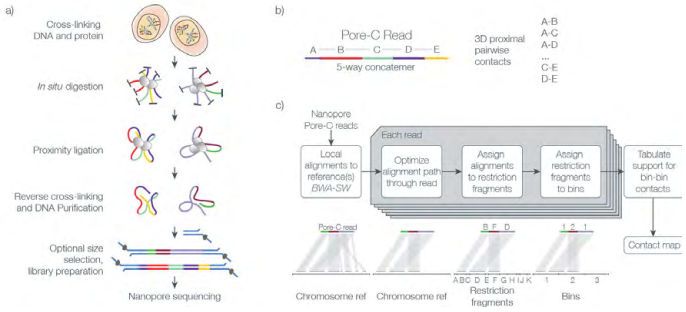
PREPARATION

How does Pore-C data improve large genome assembly quality?

Chromatin conformation capture (3C) methods reveal three-dimensional interactions within a genome. This information can be utilised to orient contigs during assembly; however, the traditional use of short-read-based 3C methods reduces the number of contacts per read, limiting resolution. Pore-C combines 3C with long nanopore sequencing reads, revealing long-range, multi-way contact information, enabling the scaffolding of large genomes to high contiguity. The PCR-free method also allows for the detection of base modifications. Pore-C is an end-to-end workflow, from extraction, to sample prep, to data analysis, with guidance to support you through each step.

Find out more about Pore-C:
nanoporetech.com/applications/investigation/chromatin-conformation

Pore-C: a) laboratory workflow b) multi-contact reads c) overview of bioinformatics workflow



From sample to answer

LIBRARY PREPARATION

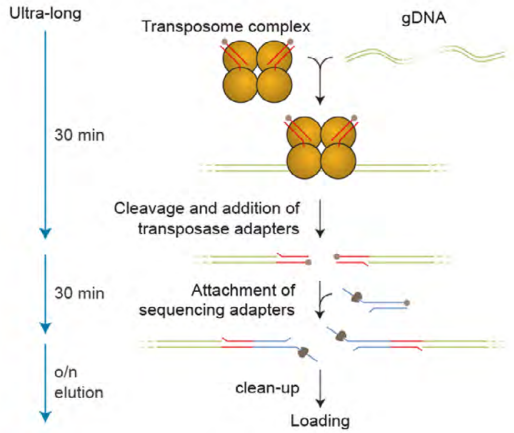
How does the Ultra-Long DNA Sequencing Kit work?

Ultra-long nanopore sequencing reads, which can be defined as those greater than 50 kb in length, can span the most challenging genomic regions, including very long repeat sequences, large structural variants, centromeres, and telomeres. Utilising ultra-long reads in large genome assembly can therefore resolve gaps that cannot be spanned with shorter reads, increasing contiguity and completeness. Ultra-long reads also simplify the process of phasing, with multiple variants or modifications captured per read.

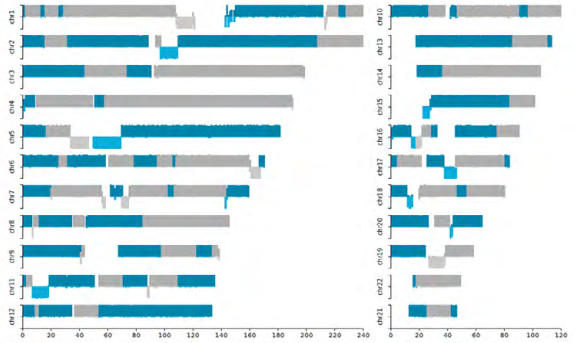
The Ultra-Long DNA Sequencing Kit, used in combination with the NEB Monarch HMW DNA Extraction Kit, is optimised for the preparation of ultra-high molecular weight DNA sequencing libraries. These can produce read length N50s of ~50–100 kb or more, with reads extending beyond a megabase in length. The method requires only ~2 hours hands-on time for sample prep and ~1 hour for library prep, plus an overnight elution step.

View the Ultra-Long DNA Sequencing Kit:
store.nanoporetech.com/sample-prep.html

The Ultra-Long DNA Sequencing Kit workflow



Phasing human genome HG002 using ultra-long nanopore sequencing reads



From sample to answer

LIBRARY PREP

Can I multiplex my samples?

Barcoding options are available for the sequencing of multiple samples on the same flow cell. This can be utilised, for example, to sequence two to three human genomes at low-to-medium coverage on one PromethION Flow Cell, or to sequence higher numbers of large genomes where only shallow coverage is required, such as for 'genome-skimming' approaches. To sequence complete genomes, including regions that cannot be amplified, we recommend the PCR-free Native Barcoding Kits, which use a ligation-based approach to attach barcodes to each sample. PCR-based and rapid barcoding approaches are also available.

Find out more about multiplexing kit options:
store.nanoporetech.com/sample-prep.html

SEQUENCING

Can I reuse my flow cell?

Yes — the Flow Cell Wash Kit can be used to remove >99.9% of a sequenced library, leaving the flow cell ready either to sequence a fresh library or to be prepared for storage until it is needed again. This is especially useful when sufficient depth of coverage of a genome is obtained early on in a sequencing run: the run can be stopped when a coverage target is reached, then the flow cell washed and used again for a new sample. By barcoding each library, any residual carried-over reads can then be filtered out in analysis.

View the Flow Cell Wash Kit:
store.nanoporetech.com/expansion-packs.html

DATA ANALYSIS

How can I analyse my data?

A range of tools is available for the assembly of large genomes using long nanopore sequencing reads **(see Table 2)**. Selecting the right tool for your assembly can involve several factors, such as optimising for maximum contiguity or fastest analysis time. New tools and updates to existing methods are frequently shared: these can be found in our Resource centre, which is continuously updated with the latest developments.

Find out more about nanopore sequencing analysis solutions:
nanoporetech.com/analyse

Table 2

Assembler	Advantages	Reference
Flye	Fast, best contiguity and completeness	Kolmogorov *et al.* (2019). doi:10.1038/s41587-019-0072-8
Shasta	Fastest, also has slower low-memory mode	Shafin *et al.* (2020). doi:10.1038/s4187-020-0503-6
Raven	Very fast, lowest memory	Vaser and Šikić (2021). doi:10.1038/s43588-021-00073-4
NextDenovo	Good contiguity (more time required)	github.com/Nextomics/NextDenovo
Canu	Good contiguity (more time required)	Koren *et al.* (2017). doi:10.1101/gr.215087.116

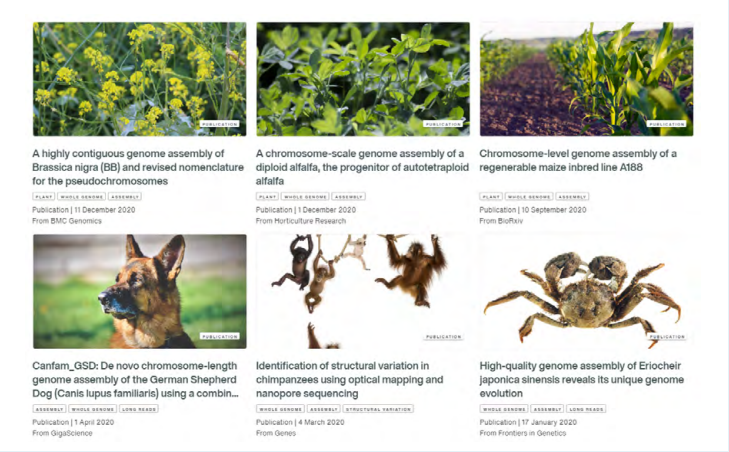
DISCUSSION

Where can I discuss whole-genome sequencing with other Oxford Nanopore users?

The Nanopore Community is a thriving online hub, helping users to get started, share their work and experiences, and collaborate. Members of the Nanopore Community are continually developing new protocols and tools for extraction, library prep, and analysis, for an increasingly diverse range of applications. You can also find the latest publications featuring whole-genome nanopore sequencing in the 'Resources' section of our website.

Join the Community discussion and ask the experts here:
community.nanoporetech.com

Find out how nanopore technology is being used for whole-genome sequencing:
nanoporetech.com/applications/whole-genome-sequencing



Case studies

Case study 1: The lungfish: assembling the largest animal genome to date with long nanopore sequencing reads

Lungfishes were able to 'conquer' the land ~400 million years ago. Characterisation of lungfish genomes is of high importance in understanding the evolutionary mechanisms enabling these animals to adapt to life on land; however, until recently, their expected size and predicted high repeat content meant that sequencing and assembling the lungfish genome had been considered impossible[7].

Meyer *et al*. used long and ultra-long nanopore reads reads to tackle the genome assembly of the 'living fossil' Australian lungfish (*Neoceratodus forsteri*)[8]. Using the high-throughput PromethION, they generated ~1.2 Tb nanopore sequencing data. Following polishing and scaffolding using short reads and chromatin conformation capture data, they produced a chromosome-scale assembly, obtaining 27 large scaffolds corresponding to the chromosome sizes estimated via cytogenetics. The genome spanned 43 billion base pairs, representing the largest animal genome assembled to date.

Phylogenetic analysis supported the hypothesis that lungfishes are the closest living relatives of land vertebrates, sharing their last common ancestor ~420 million years ago. The genome was found to feature huge intergenic regions, with ~90% repeat content; the repeat structure, dominated by LINEs, was more similar to the genomes of land vertebrates than those of fishes. Introns represented ~21% of the genome — a similar proportion to that found in the human genome. Comparative genomic analysis suggested the presence of genetic pre-adaptations for life on land, including those for air breathing, olfaction, and the evolution of limbs from their fins — advancing '*our understanding of this major transition during vertebrate evolution*.'

Read the publication:
nanoporetech.com/giant-lungfish-genome

Case study 2: Chromosome-scale assembly of a wheat genome achieves higher resolution for research and breeding

Triticum aestivum, or bread wheat, is a food staple worldwide and, therefore, one of the most economically important cultivated crops. However, the growing population and climate change pose challenges for securing enough wheat for food production; current estimates suggest production must increase by 50% over existing levels by 2050 to meet demand[9]. Strategic breeding programs are critical to addressing these challenges, and these are dependent on a comprehensive knowledge of the wheat genome. Sequencing the wheat genome has historically proven challenging, owing to its large size (15.5 Gb), high repeat content, and hexaploidy[10]; short-read sequencing technologies *'underestimate the repetitive content of the genome and ... can lack tandemly duplicated genes'*[11]. Long sequencing reads offer a solution to that end: using long nanopore reads generated on a PromethION device and subsequent scaffolding via an optical technology, Aury *et al*. produced *'the most contiguous and complete chromosome-scale assembly of a bread wheat genome to date.*[11]' Following polishing with short reads, the assembly was highly complete, with a BUSCO score of 96.6%, whilst the contig N50 of 2.2 Mb demonstrated a *'30-fold improvement over existing chromosome-scale assemblies'*. This high-quality resource could facilitate the rapid selection of agriculturally important traits, aiding breeding programs and enhancing crop yield.

Read the publication:
nanoporetech.com/hexaploid-wheat-assembly

7. Franchini, P. Blog: On the quest to assemble the giant lungfish genome. Available at: https://nanoporetech.com/about-us/news/blog-quest-assemble-giant-lungfish-genome [Accessed: 07 February 2022].
8. Meyer, A. et al. Giant lungfish genome elucidates the conquest of land by vertebrates. Nature 590:284–289 (2021).
9. Tadesse, W. et al. Genetic gains in wheat breeding and its role in feeding the world. Crop Breeding, Genetics and Genomics. 10.20900/cbgg20190005 (2019).
10. Tørresen, O. et al. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Research 47, 10994-11006 (2019).
11. Aury, J. et al. Long-read and chromosome-scale assembly of the hexaploid wheat genome achieves higher resolution for research and breeding. bioRxiv 457458 (2021).

Oxford Nanopore Technologies
Phone: +44 (0)845 034 7900
Email: sales@nanoporetech.com
Twitter: @nanopore

nanoporetech.com

GS_1026(EN)_V3_16Feb22