WORKFLOW: ANIMAL GENOME ASSEMBLY

Assembling animal genomes using long nanopore sequencing reads

The release of the first animal genome sequence, for *Caenorhabditis elegans*[1], instigated a new era of animal genome sequencing and the acquisition of a wealth of knowledge about animal biology, evolution, and biodiversity[2]. However, despite such progress, only around 0.2% of animal species have had their genomes sequenced, and because many assemblies have been derived using short-read sequencing technology, many remain incomplete[2].

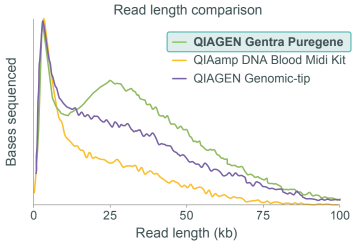
Compared to short-read data, long and ultra-long nanopore sequencing reads enable the resolution of repeat-rich sequences and large-scale structural variants, and demonstrate a lack of bias in GC-rich regions, supporting the assembly of high-quality, highly contiguous animal genomes. These strengths, in combination with the high output and throughput of PromethION™ sequencing devices, allow nanopore technology to further our understanding of genetic variation across the animal kingdom, and therefore may ultimately advance breeding and conservation efforts.

Here we present a simple workflow for animal genome assembly from a mammalian blood sample using PromethION Flow Cells.

EXTRACTION: obtaining high molecular-weight DNA

Find more extraction protocol recommendations for your sample type, from avian blood to insect and reptilian tissue: **community.nanoporetech.com/docs/prepare**

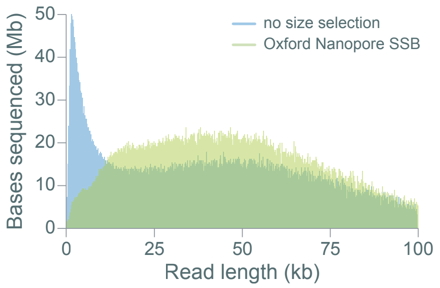
Selecting a suitable extraction method for obtaining high molecular-weight DNA greatly depends on sample type. For DNA extraction from mammalian blood, we recommend the **QIAGEN Gentra Puregene Blood Kit**, which we have found to generate high sequencing yields and long read lengths.



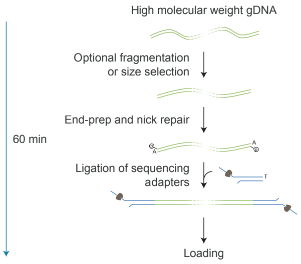
LIBRARY PREPARATION: selecting a kit

Find out more about size selection options: **community.nanoporetech.com/extraction_method_groups/size_selection**

Obtaining long sequencing reads is important for genome assembly, as it maximises the overlap between reads at the downstream analysis stage. We suggest aiming for a read N50 of 25–35 kb. To this end, if the extracted gDNA is dominated by fragments of <10 kb in length, we recommend performing size selection. This can be achieved using our **size selection buffer (SSB)**, which enriches for molecules >~10 kb.



Ligation Sequencing Kit workflow


High molecular weight gDNA
Optional fragmentation or size selection
End-prep and nick repair
60 min
Ligation of sequencing adapters
Loading

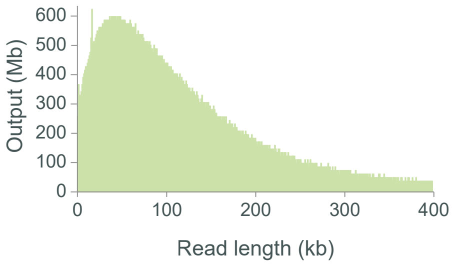
To prepare gDNA for nanopore sequencing, we recommend the **Ligation Sequencing Kit**, providing the greatest yield and control over read lengths. To further increase contiguity and completeness of an assembled genome, and if sample availability allows, we suggest performing an additional run using the Ultra-Long DNA Sequencing Kit and associated workflow. This involves the extraction, library preparation, and sequencing of ultra-high molecular-weight DNA (read N50s >50 kb).

Learn more about whole-genome sequencing and assembly, including how to incorporate ultra-long reads, in our Getting Started guide: nanoporetech.com/resource-centre/guide-whole-genome-sequencing-large-genomes

SEQUENCING: generating high yields of long reads with PromethION

Find out more about the PromethION range:
nanoporetech.com/products/promethion

Typical read length distribution obtained from a library prepared with the Ultra Long DNA Sequencing Kit and sequenced on a PromethION Flow Cell



Assuming an animal genome of ~2–4 Gb in length, we recommend sequencing to a minimum depth of coverage of 30x of 25–35 kb reads; this can be achieved by sequencing on one PromethION Flow Cell. However, sequencing up to a depth of 60x will likely improve assembly contiguity for genomes consisting largely (≥50%) of repetitive DNA — which is a common characteristic of mammalian genomes[3]. For this, a total of two PromethION Flow Cells may be needed. If also sequencing a library prepared with the Ultra-Long DNA Sequencing Kit, we recommend sequencing this on a single PromethION Flow Cell.

We recommend basecalling in high accuracy (HAC) or super accuracy (SUP) mode; SUP basecalling may yield slight improvements to assembly contiguity over HAC, but at the expense of a 2–3x increase in basecalling compute intensity.

Output can be maximised by performing a nuclease flush and loading fresh library every 24 hours.

PromethION sequencing devices have the capacity to run up to 2 (P2), 24 (P24), or 48 (P48) PromethION Flow Cells at any time, providing ultimate flexibility and adaptability to your sequencing needs. For lower throughput requirements, the P2 Solo can be plugged into a GridION™ to expand sequencing capacity, or the stand-alone, compact P2 device can quickly deploy PromethION-scale sequencing to any lab. The P24 and P48 sequencers deliver higher throughput, ideal for larger projects.

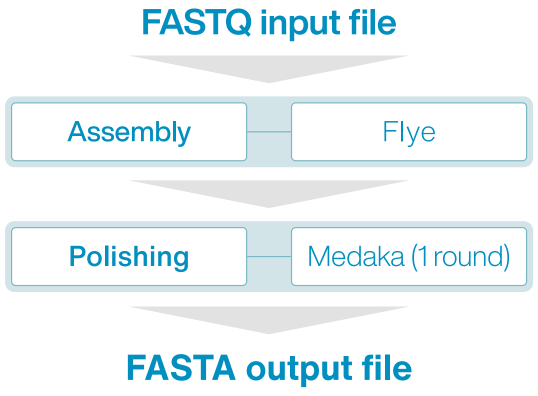
Find out about nanopore sequencing service providers: nanoporetech.com/services/providers

ANALYSIS: selecting an assembly tool

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

To assemble animal genomes, we suggest using the third-party *de novo* assembly tool **Flye**[4]. This analysis package represents a complete pipeline, taking raw nanopore reads as input, and producing polished contigs as output. We also advise one round of additional polishing of the assembly with **Medaka**[5]. These tools can both be found on GitHub.

Regarding analysis runtime, assuming an animal genome of ~2–4 Gb, assembly with Flye would require 1–2 days (based on an AWS instance, with 1 TB RAM and 128 CPU threads). Polishing with Medaka would require an additional day.



FASTQ input file

Assembly — Flye

Polishing — Medaka (1 round)

FASTA output file

Find out more at: **nanoporetech.com/applications/animal-genomics**

References:

1. The *C. elegans* Sequencing Consortium. Science. 282:2012-2018 (1998).
2. Hotaling, S. et al. PNAS. 118(52):e2109019118 (2021).
3. Haubold, B. and Wiehe, T. BMC Bioinformatics. 7:541 (2006).
4. Kolmogorov, M. et al. Nat. Biotech. 37:540-546 (2019).
5. Oxford Nanopore Technologies. Medaka. Software available at: https://github.com/nanoporetech/medaka

Twitter: @nanopore
www.nanoporetech.com