Helminth Bioinformatics - Latin America & the Caribbean
18–24 May 2025
Universidad de la Republica, Uruguay

Module: Long-read sequencing of parasitic helminth genomes

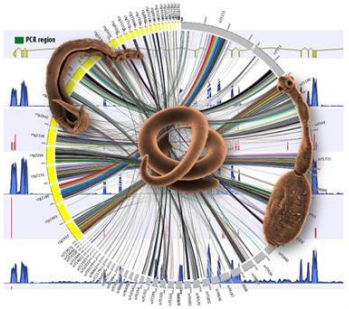
Laura Kamenetzky

Professor, Department of Physiology and Molecular and Cellular Biology, Faculty of Exact and Natural Sciences, University of Buenos Aires (UBA)

Researcher, National Scientific and Technical Research Council (CONICET)

May, 2025

Module aims

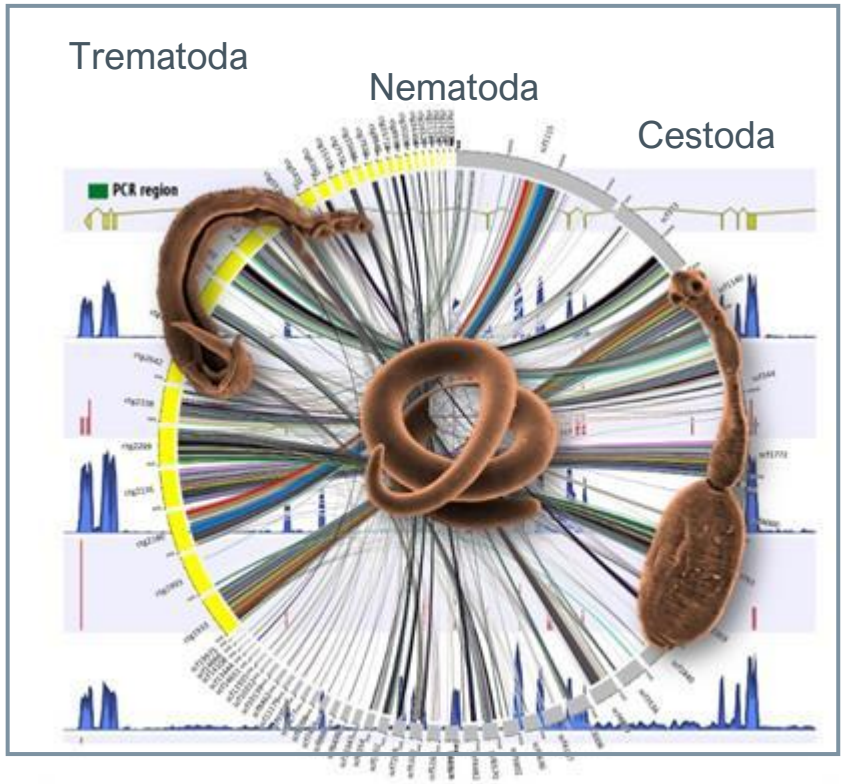


You will learn how to:

- Make your own DNA sequencing library
- Obtain long sequencing reads
- Filter raw sequence data by quality
- Extract biological information from the high quality sequence data
- Perform *de novo* genome assembly

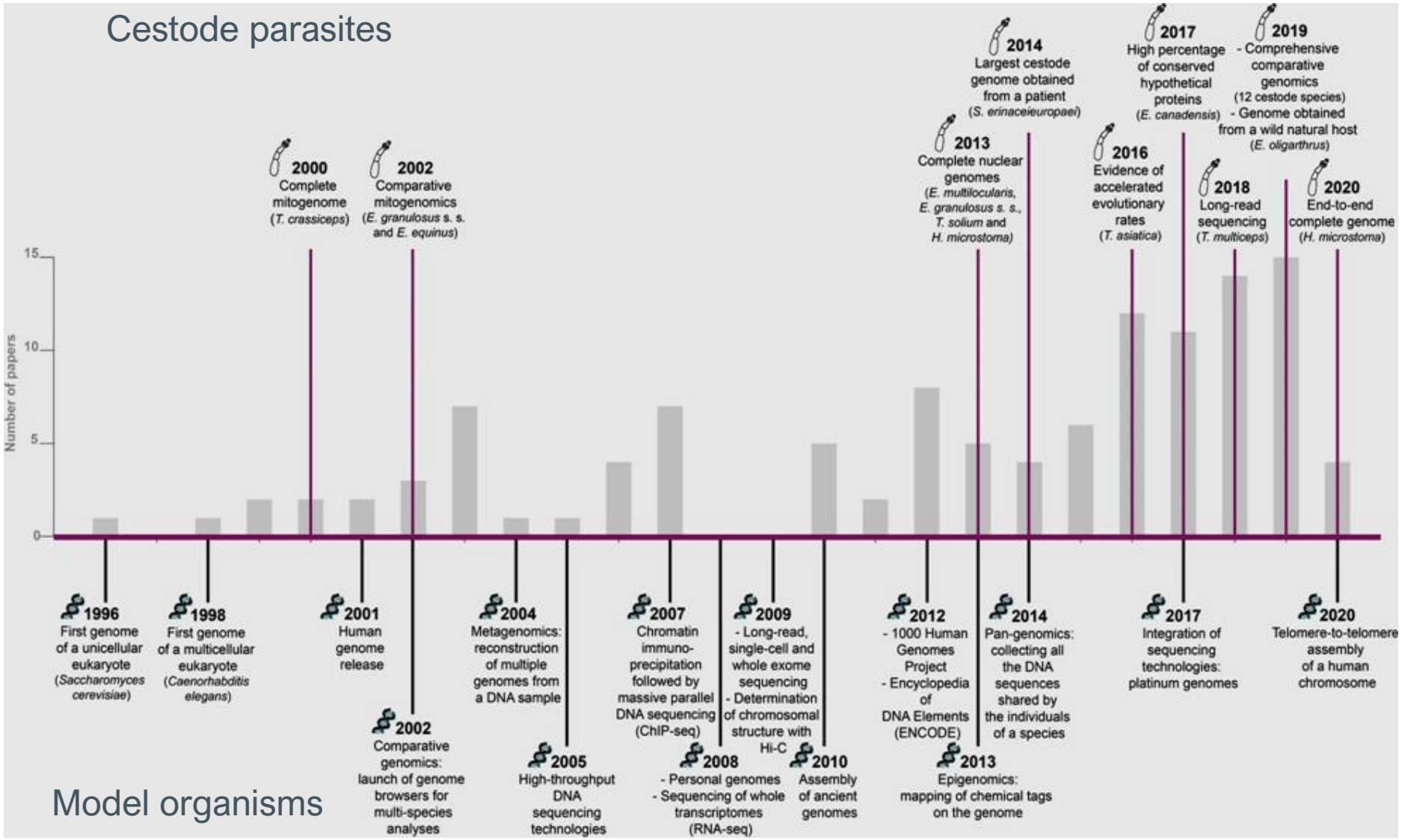
Helminth Genomics today

- Several helminth genomes are available



- We currently have numerous helminth genomes available
- A genome is defined as all of the DNA assembled into discrete chromosomes
- The most complete genomes provide higher quality information for the study of parasite helminths
- The high quality genome information can be used for the study, control, and prevention of the diseases caused by helminth parasites

Timelapse of Cestode Genomics



The first cestode genomes were obtained more than 10 years after the human genome

The difficulty in obtaining complete genomes of helminths is related, among other factors, to the limited access to parasite samples and the fact that there are few research groups dedicated to helminth genomics.

However, some cestode genomes are complete from end to end, such as the genome of *Hymenolepis microstoma*

Kamenetzky et al., 2022

Specialized DataBase: WormBase ParaSite

2025

http://parasite.wormbase.org/

206 Nematoda

69 Platyhelminthes

WormBase ParaSite Version: WBPS19 (WS291) - Archive: WBPS18

Search WormBase ParaSite...

e.g. *O. volvulus*, PRJNA60051, WBGene00262434, Bma-eat-4, eat-4 or metallopeptidase

Genome List | BLAST | BioMart | REST API | Downloads | Tools | WormBase | Login | Register | Help and Documentation

Genome List

Contents

Nematoda (206)
Platyhelminthes (69)

Nematoda

Show All entries | Show/hide columns (8 hidden) | Filter

Species Name	Clade	BUSCO ANNOTATION	N50	Genome Size	Number of Coding Genes
Wuchereria bancrofti	Clade III		12,368,652	88,416,250	11,166
Wuchereria bancrofti	Clade III		9,872	76,991,470	13,058
Trissonchulus sp. WLG1_4	Clade II		51,761	421,514,493	24,757
Trissonchulus latispiculum	Clade II		50,902	642,239,299	35,701
Trileptium ribeirensis	Clade II		44,214	738,795,508	26,621
Trichuris trichiura	Clade I		11,299,416	80,573,711	12,671
Trichuris suis	Clade I		443,734	71,056,402	14,261
Trichuris suis	Clade I		503,034	74,234,559	14,436
Trichuris suis	Clade I		1,322,386	63,839,715	9,831
Trichuris muris	Clade I		28,941,788	111,837,642	14,995
Trichinella zimbabwensis	Clade I		205,645	50,795,114	14,764
Trichinella sp. T9	Clade I		212,690	49,075,249	13,111
Trichinella sp. T8	Clade I		239,129	49,331,761	14,919
Trichinella sp. T6	Clade I		158,103	50,853,738	15,241
Trichinella spiralis	Clade I		212,546	50,025,957	14,737

Despite the challenges involved in obtaining helminth genomes, there are currently over two hundred helminth genomes available in

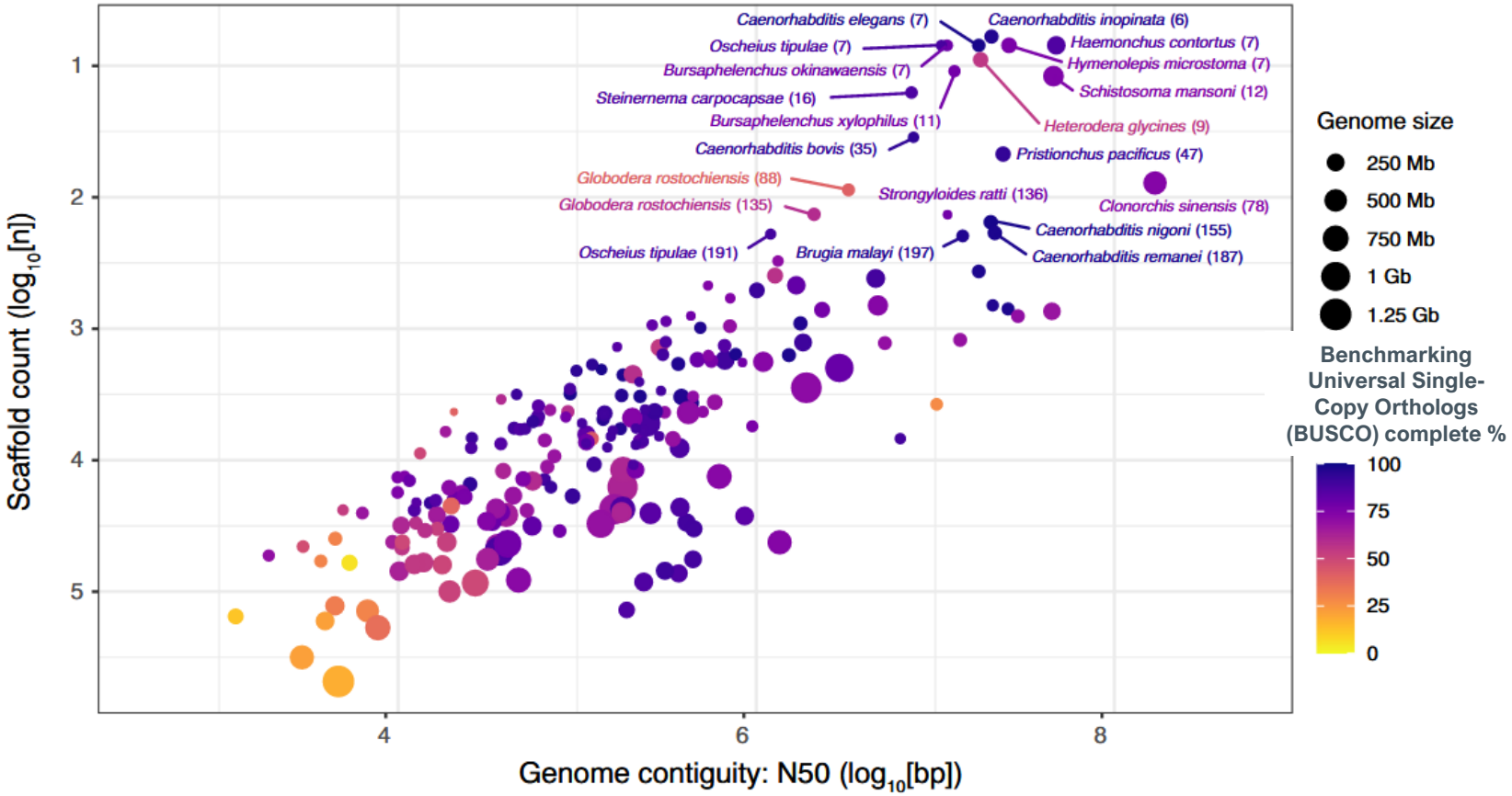
WormBase Parasite database

.

Has powerful analysis tools for rapid access to the genome sequences and annotations that will be used in this course

Completeness of helminth genome assemblies

Completeness of helminth genomes available from WormBase ParaSite release 17



Only a few of the available genomes are of high quality, represented by the points in the upper right corner

Most of the genomes are computationally represented by thousands of contigs and scaffolds, which provide different levels of ‘completeness’ in the genomes

Adapted from Doyle, 2022

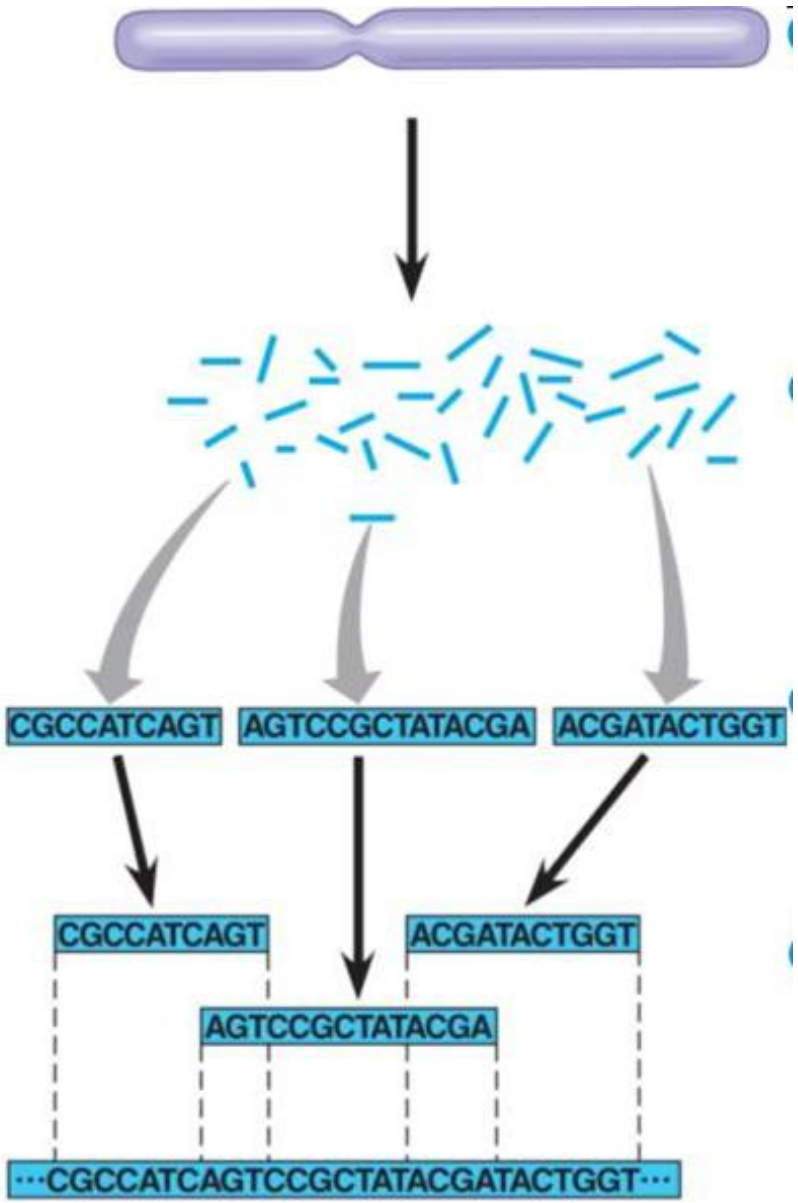
Complete genome process: four main steps

1. **Biological sample:**
   high quality DNA

2. Library construction:
cut the gDNA into overlapping fragments short enough for sequencing

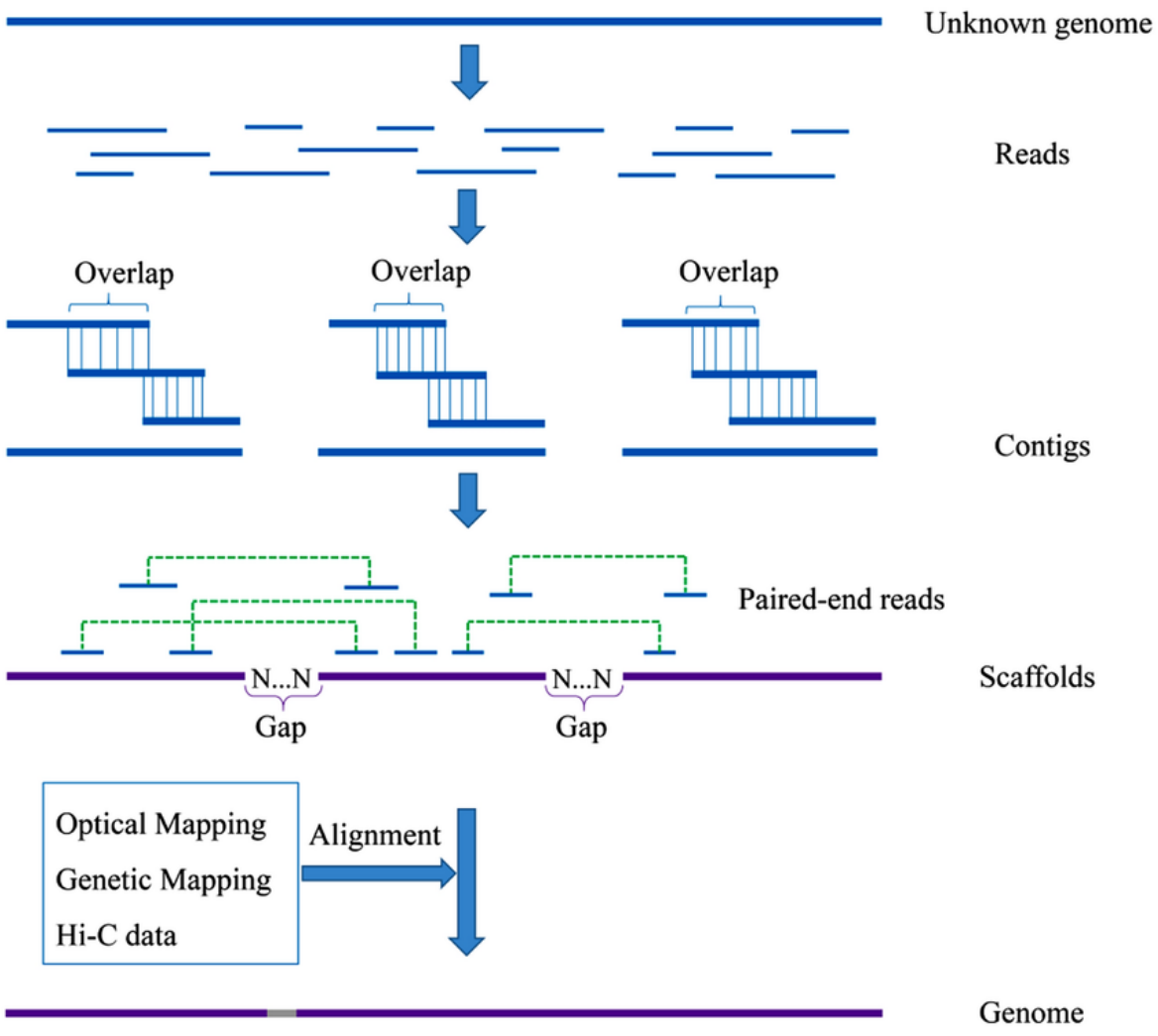
3. Sequencing
load the flow cell onto sequence machine and read each DNA fragment

4. Assembly:
order the reads into one overall DNA sequence with computer software



Human 3 ×10^{9} base pairs
C. elegans 8×10^{7}
E. coli 4×10^{6}

Complete genomes



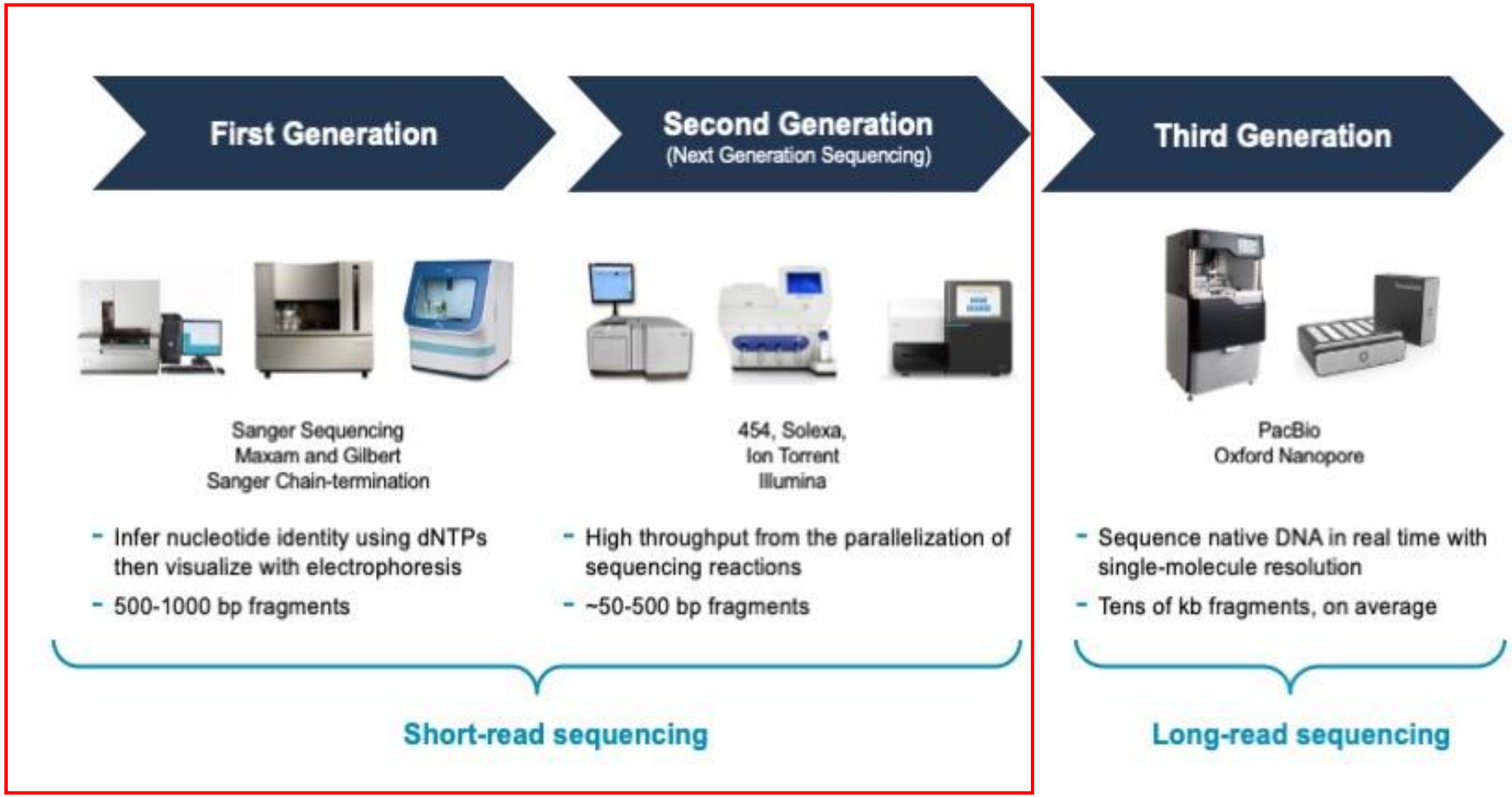
Different stages of genome assembly

The **contig assembly stage**: reads produced by the sequencer are ordered based on sequence identity at their ends.

The **scaffold assembly stage**: contigs are arranged based on their identity using additional paired reads, leaving gaps with no information.

The final stage is the one that produces **end-to-end complete genomes**, assembled into chromosomes.

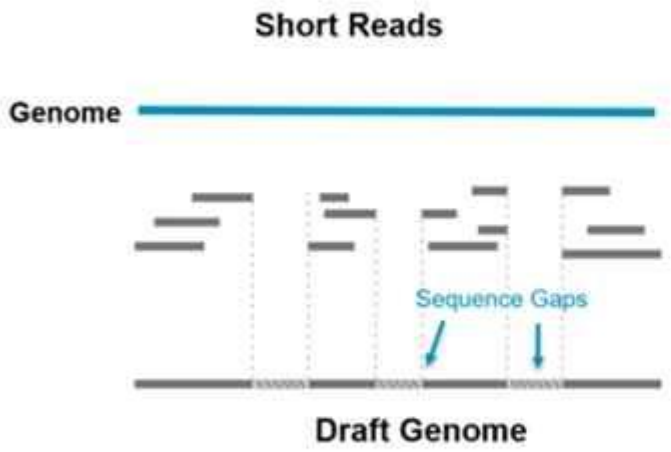
Advances in sequencing technology



Next generation sequencing (NGS)

Second generation

50-500 pb

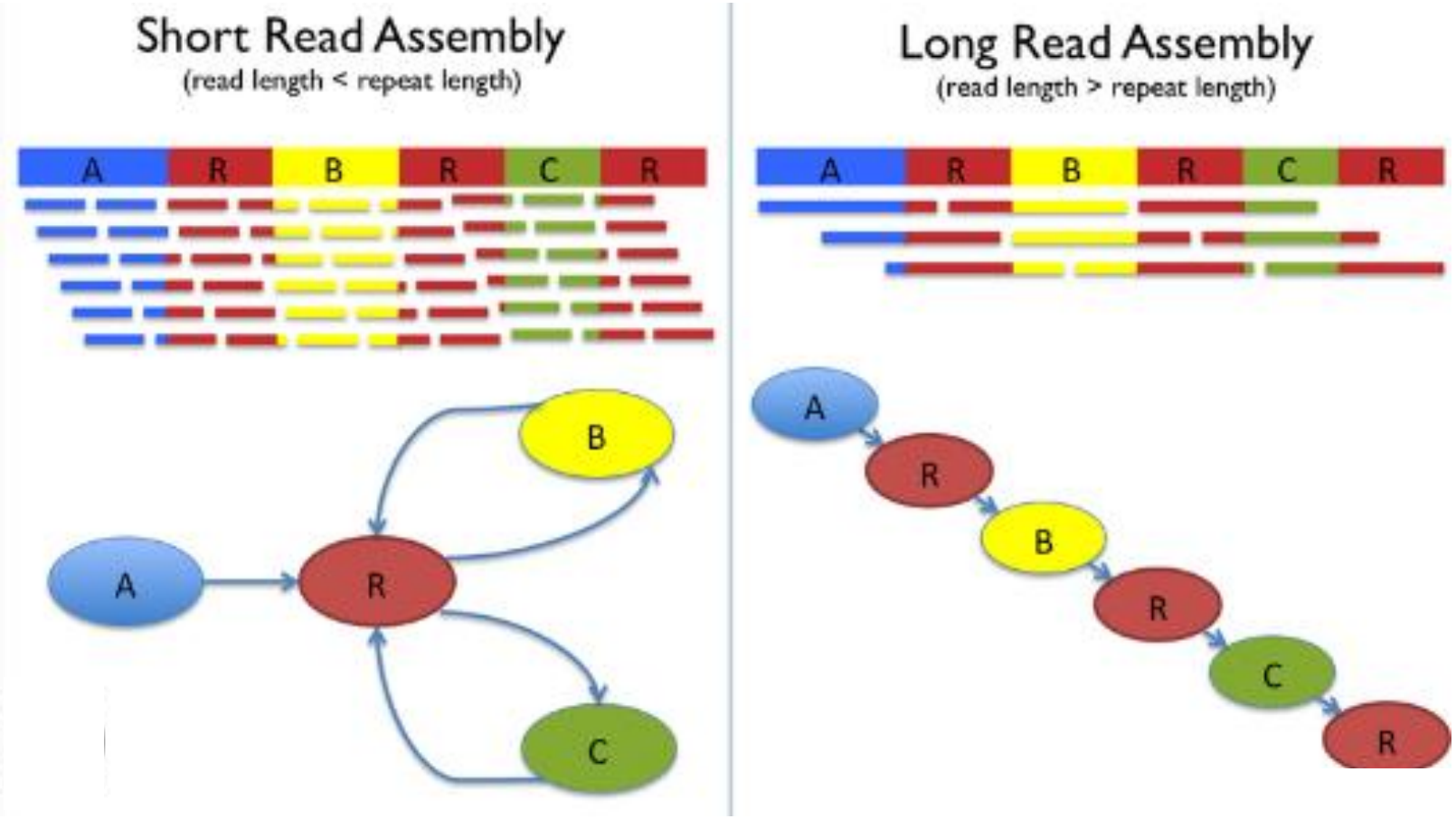


Missing sequencing leads to missed genes and limits biological interpretation

Short reads lack the context needed to resolve long repeats making it difficult to produce accurate and complete genome assemblies

Short reads are often not long enough to span entire repetitive regions, which means that the assembler cannot determine where in the genome the reads truly belong.

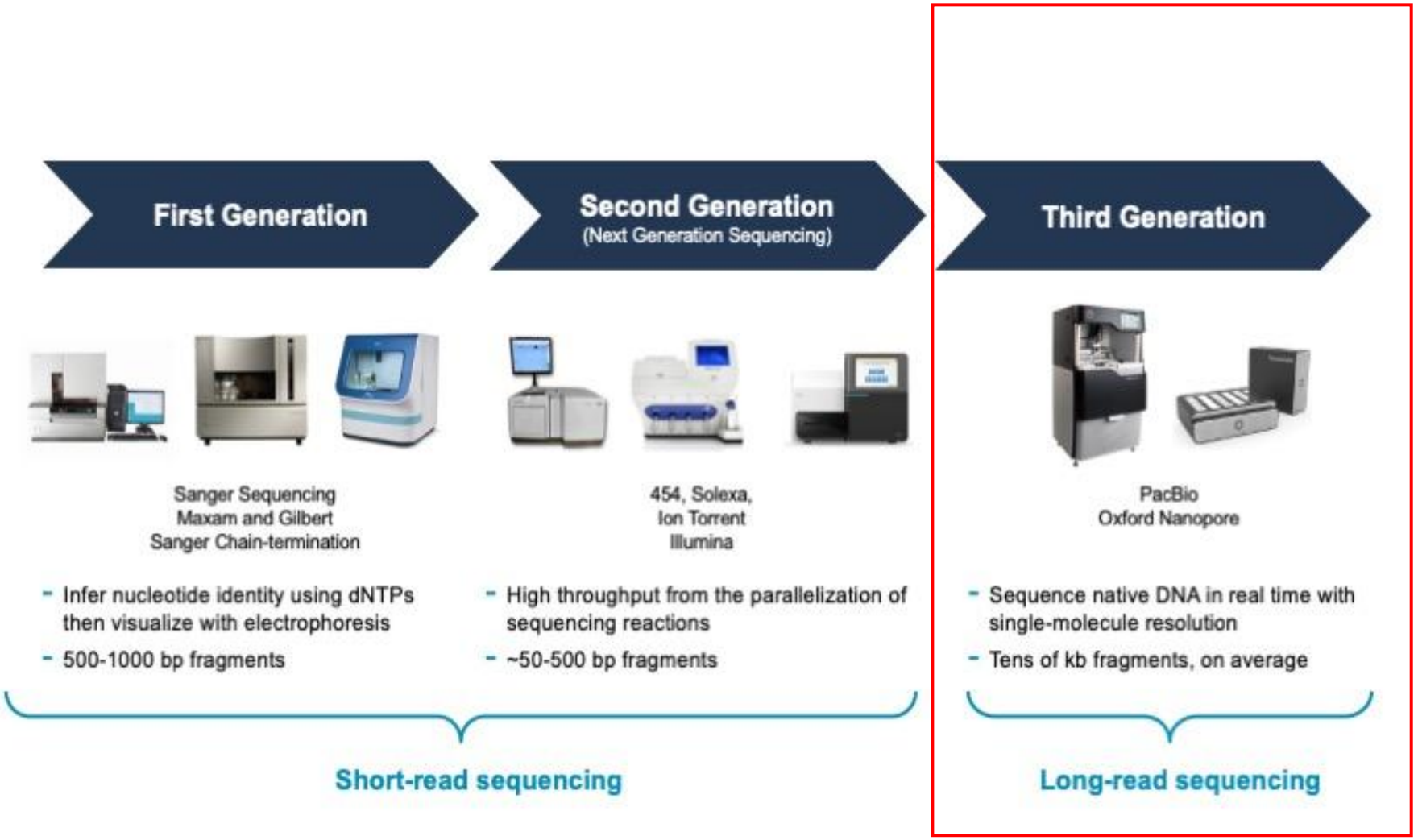
This leads to:
1.Collapsed repeats.
2.Fragmented assemblies
3.Misassemblies



Long-read sequencing enables the unambiguous placement of repetitive sequences, as the reads are often long enough to span entire repeat regions.

Allowing the assembler to place them **unambiguously**, r educing gaps and misassemblies

Advances in sequencing technology



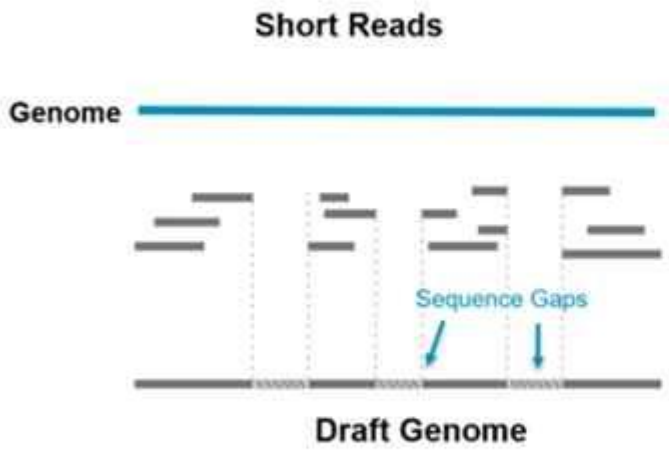
Next generation sequencing (NGS)

Second generation

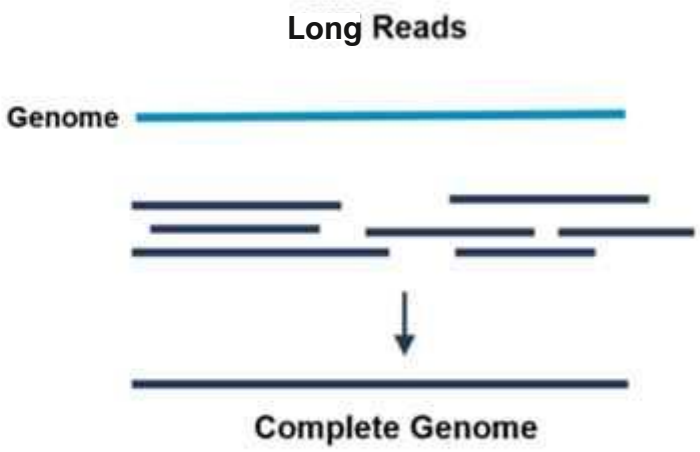
50-500 pb

Third generation

10-100 Kb



Missing sequencing leads to missed genes and limits biological interpretation



A comprehensive structural, functional and organizational picture of the genome

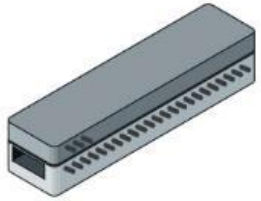
Long read sequencing devices

PacBio sequencers use a technology called **Single Molecule Real-Time (SMRT) sequencing**

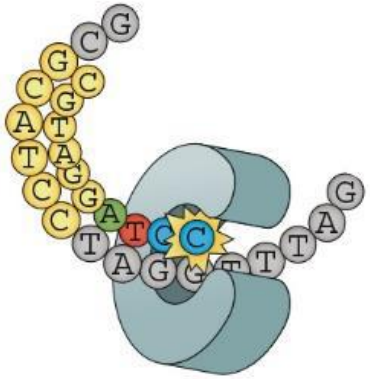
A single DNA polymerase molecule is located in a tiny well and copies the DNA incorporating fluorescent nucleotides

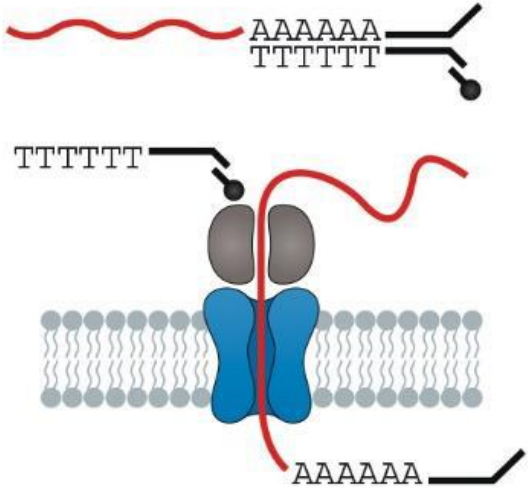
This allows the sequencing of **long reads** in real time

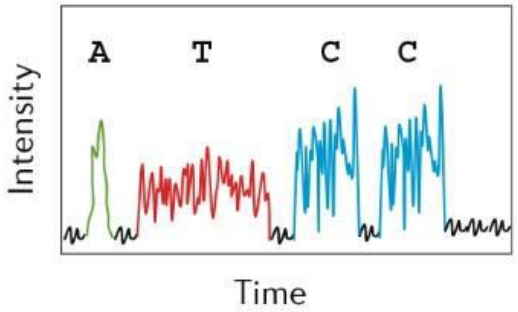
Pacific Biosciences

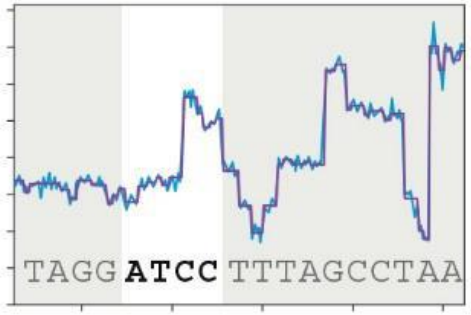


Oxford Nanopore



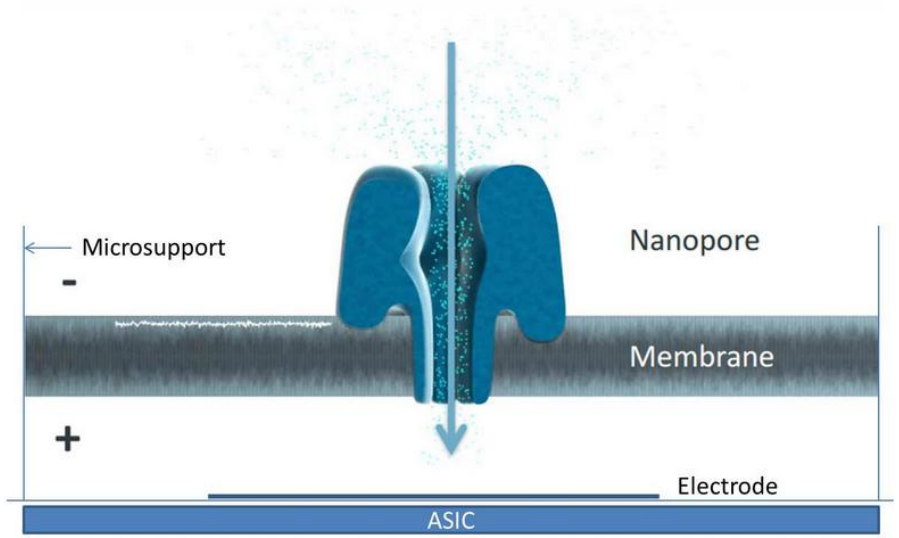






In this course, we will generate helminth genome data and perform *de novo* assembly of the mitogenome using Nanopore technology

ONT uses a strand sequencing method by **Nanopore sensing**



Nanopore sensing is the detection of a molecule coming into contact with a tiny hole, the nanopore protein.

The contact may be the molecule **passing through** the nanopore.

To sense the molecule, the nanopore is set in a **electrically-resistant membrane** so that an ionic current can pass through the nanopore when a voltage is applied across the membrane.

Disruption of the current occurs when the molecule and nanopore come into contact with each other, and this disruption can be **measured**

Oxford Nanopore Technologies (ONT)

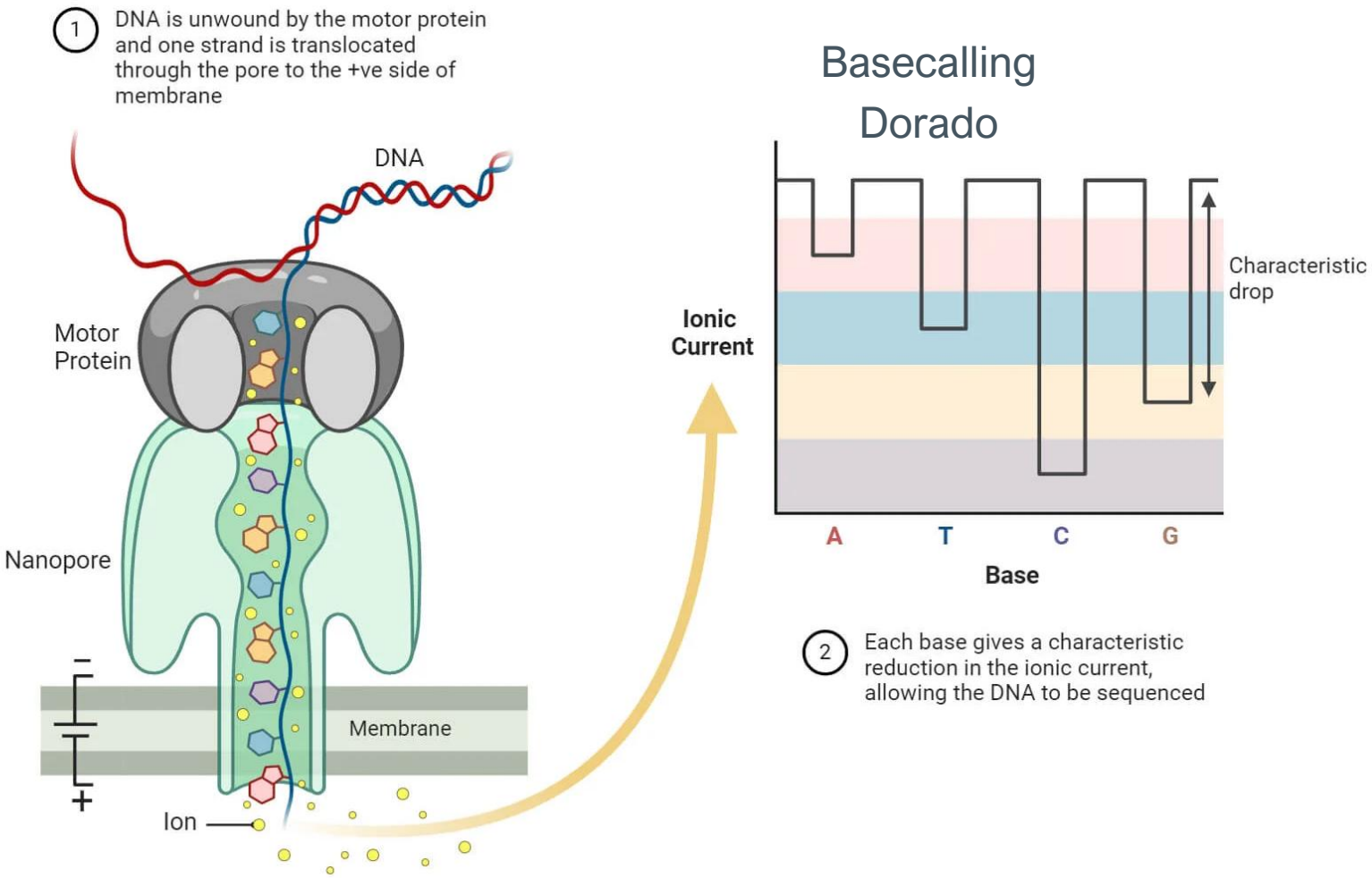
Intact DNA strands are processed by the nanopores and are **sequenced in real-time**

When DNA passes through the pore, this event creates a characteristic **disruption in current**

Measurement of that current makes it possible **to identify nucleotides** in the DNA (G, A, T, C)

The translocation into the nanopore is controlled by the inclusion of a **Motor protein**.

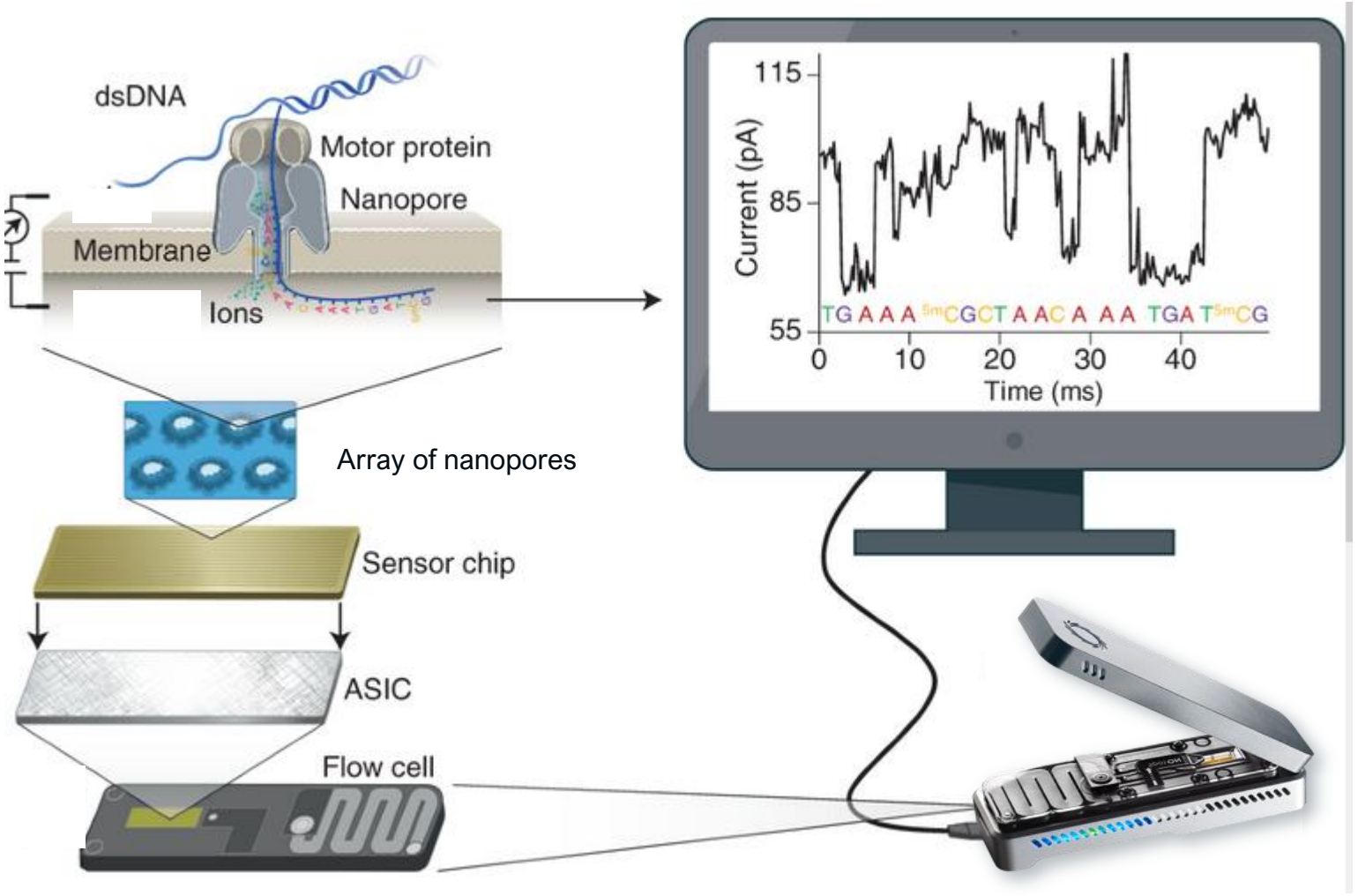
The Motor protein is provided on a leader adapter that is attached to the end of the double-stranded DNA template during **Sequencing Library construction**



Oxford Nanopore Technologies (ONT)

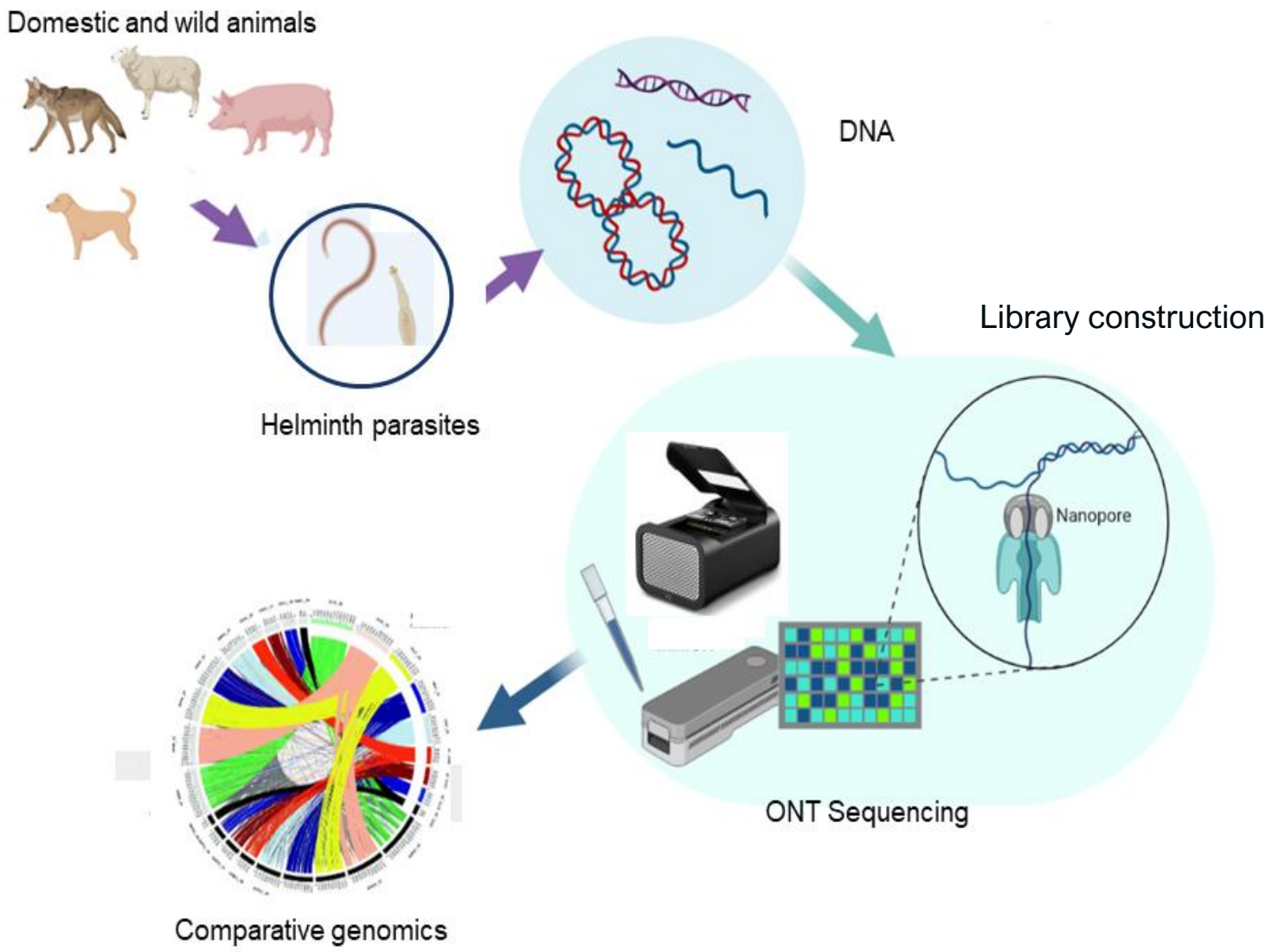
allows the real-time acquisition of hundreds of thousands of reads ranging from 10 to 100 kb

- ✓ **Does not require DNA fragmentation**
- ✓ **Does not require amplification**
- ✓ **Reads the native molecule**
- ✓ **Allows selection of the genomic region of interest:** Adapting sampling
- ✓ **Scalable:** Sequencing is scalable because the flow cells can have fewer or more nanopores depending on the amount of data that needs to be generated



The flow cell is a disposable element of the sequencing platform that provides the fluidic interface between the nanopores and the electrodes

Sequencing workflow



Helminth parasites from infected animals are isolated

High-quality DNA is purified

Sequencing is performed on ONT devices

Genomes are assembled and analysed

Practical session: Workflow

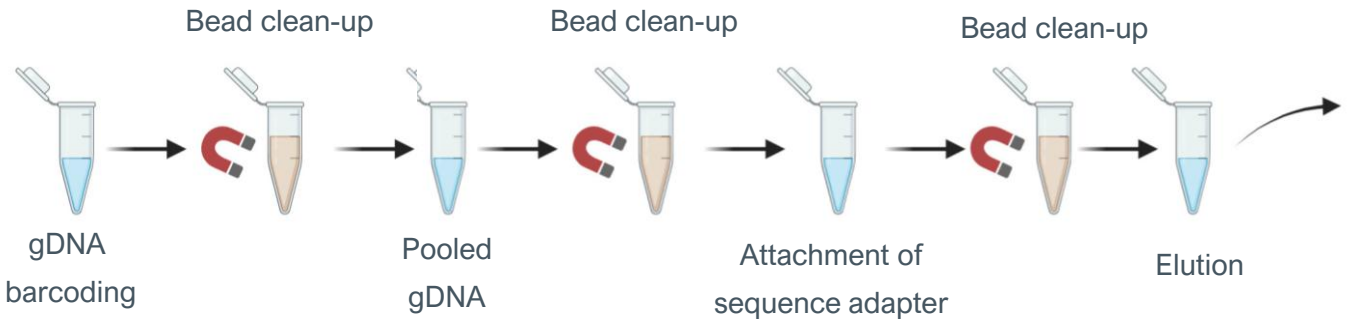
Monday morning

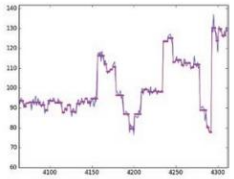
1. Library construction

Overnight

2. Sequencing

3. Basecalling





pod5 file

..TTAGGCCACTGGT..

fastq file

Monday afternoon

4. Mitogenome assembly with training data: *Dioctophyme renale*

Practical session

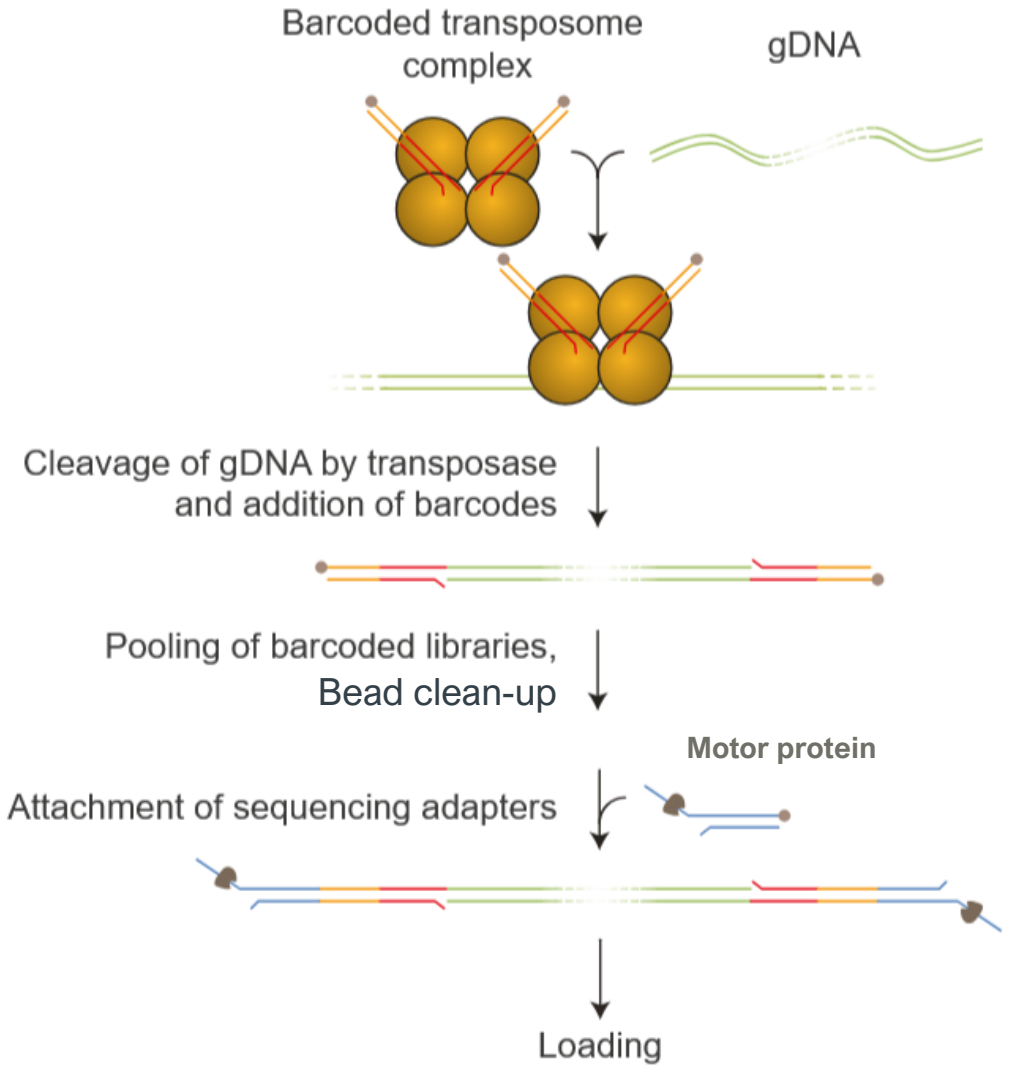
Sequencing Library construction using ONT Rapid Kits

The DNA input is undergoes a transposase step attaching barcodes to the ends simultaneously

The samples can be pooled and a bead clean-up performed

The sequencing adapters are then attached to the samples and the library can be sequenced

< 60 min



1. Barcode Attachment

2. Adapter Attachment

https://nanoporetech.com/document/chemistry-technical-document#introduction-to-library-preparation-chemistry

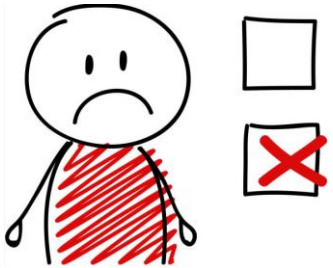
Practical session

Critical step: Bead clean-up

Shiny Bead: Ready for drying

Matte Bead: Ready for elution

Cracked Bead: Risk of low yield



Training data: *Dioctophyme renale*

WCSCourses / Helminth_Bioinformatics_2025 Public

forked from WCSCourses/Helminth_Bioinformatics

Code | Pull requests 1 | Actions | Projects | Security | Insights

Files

main

Go to file

- course_data_2025
- course_manuals_2023
- course_modules_2025
  - Mitogenomic_assembly
    - Mitogenome_assembly_guide....
    - README.md
  - module_R_refresher
  - module_WBPS_1_Basics
  - module_WBPS_2_Tools
  - module_genetic_diversity
  - module_phylogenomics
  - module_transcriptomics
  - module_unix_refresher
  - README.md
- presentations_2023
- seminar_slides_2025

Helminth_Bioinformatics_2025 / course_modules_2025 / Mitogenomic_assembly / Mitogenome_assembly_guide.pdf

csaslett Added PDF of assembly guide

2.99 MB

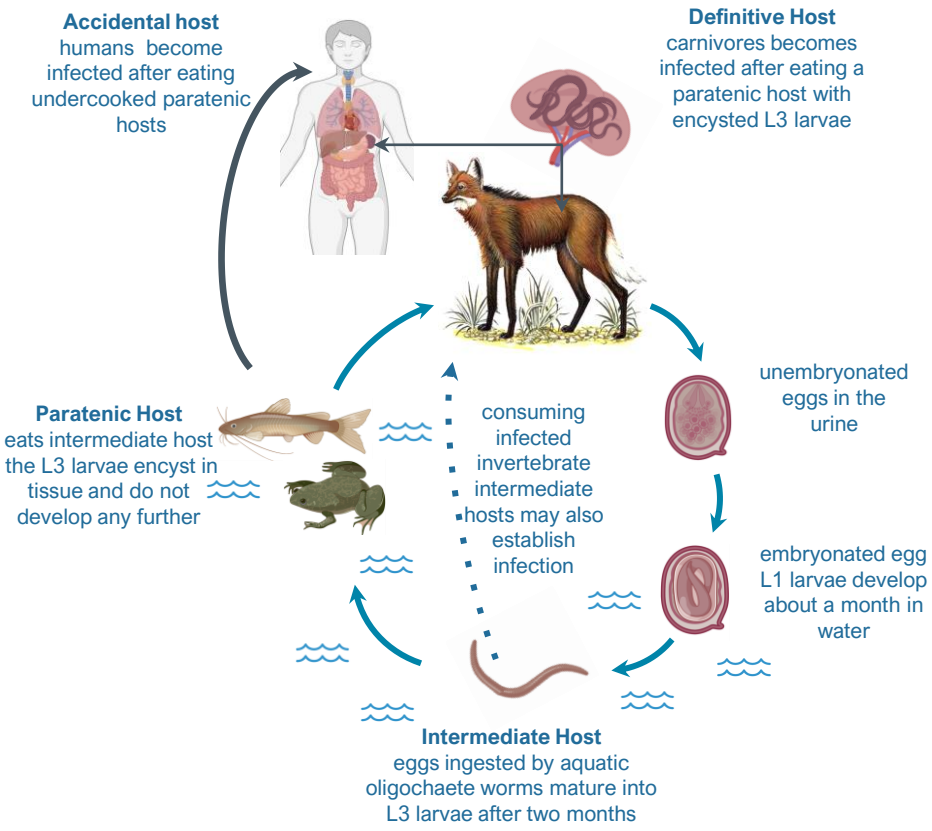
The input data are reads that have already been processed by a basecaller.

When sequencing DNA or RNA with nanopores, the changes in current caused by the DNA or RNA strand as it passes through the pore are recorded by the MinKNOW™ software running all Oxford Nanopore sequencing devices.
The processive movement of bases through the pore leads to a continuous change in the current, known as a "squiggle". MinKNOW processes the squiggle into real-time reads, each read corresponding to a single DNA/RNA strand.
These reads are written to POD5 files. This raw data contains information not only about canonical bases, but also about base modifications, such as methylation. ()
These raw data obtained during sequencing are processed and converted to FASTQ format () which is one of the most widely used formats in bioinformatics and genomics.

Monday afternoon

we will carry out the assembly of the mitogenome using nanopore reads previosuly obtained

Dioctophyme renale problem



Adapted from https://www.cdc.gov/dpdx/dioctophymiasis/

Nematode parasite commonly known as the "giant kidney worm" and considered the largest parasitic nematode of terrestrial vertebrates described to date

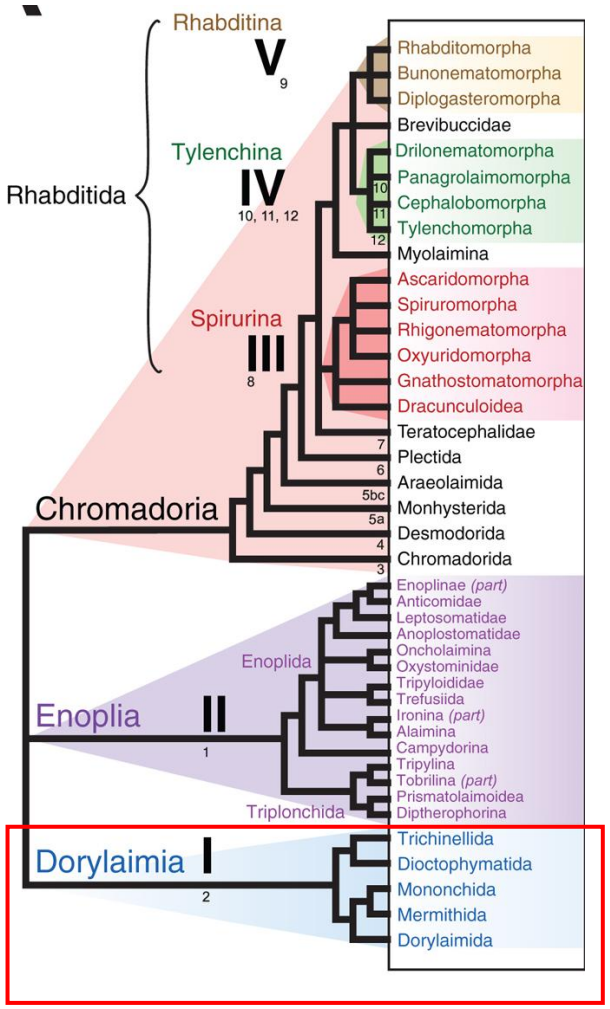
Health problem in dogs and threatened wildlife living near aquatic environments

High risk of causing infections in human populations in riparian areas

Hard to sample, obtained from surgeries or roadkill wildlife (degraded DNA)

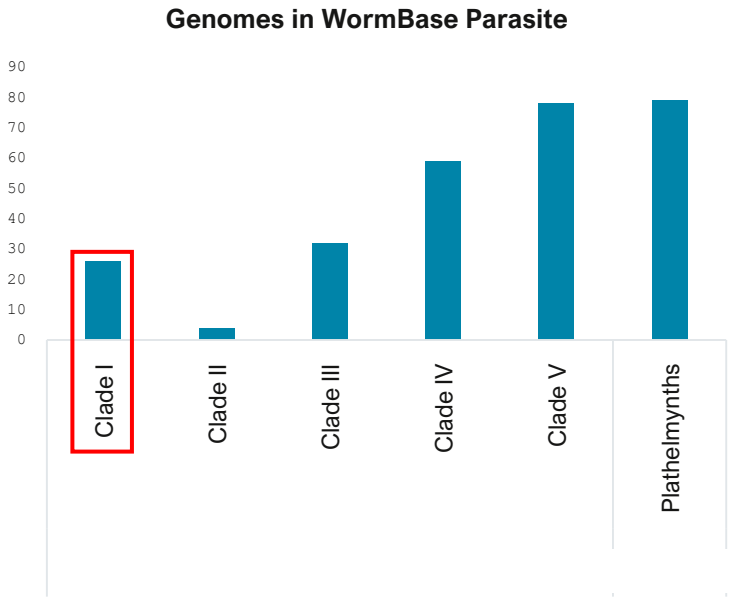
There is little molecular information on this organism (no genome nor transcriptome)

Dioctophyme renale



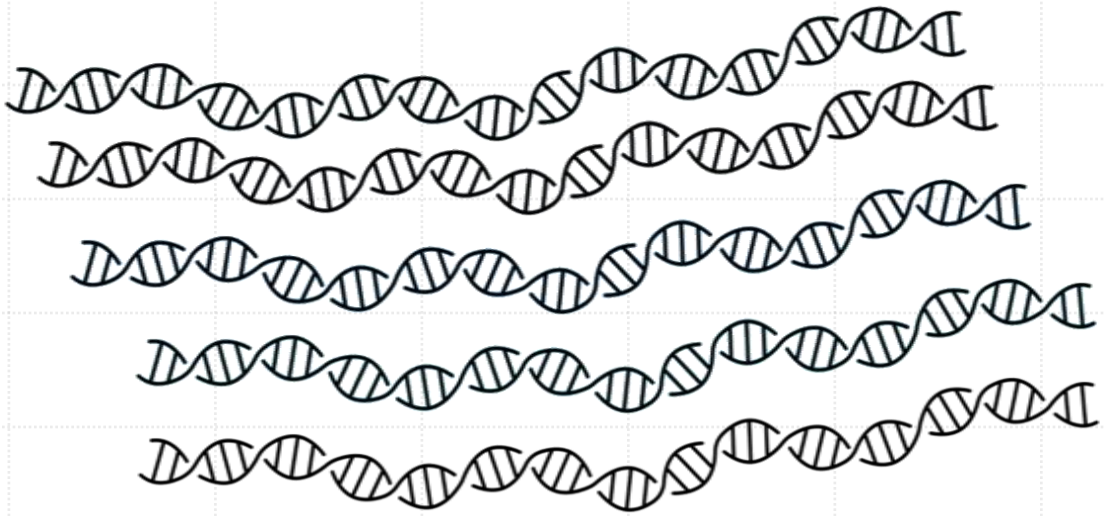
Blaxter M (2014)

Belongs to Clade I, a taxonomic group underrepresented with respect to available genomic data



The sample has both nuclear and mitochondrial DNA

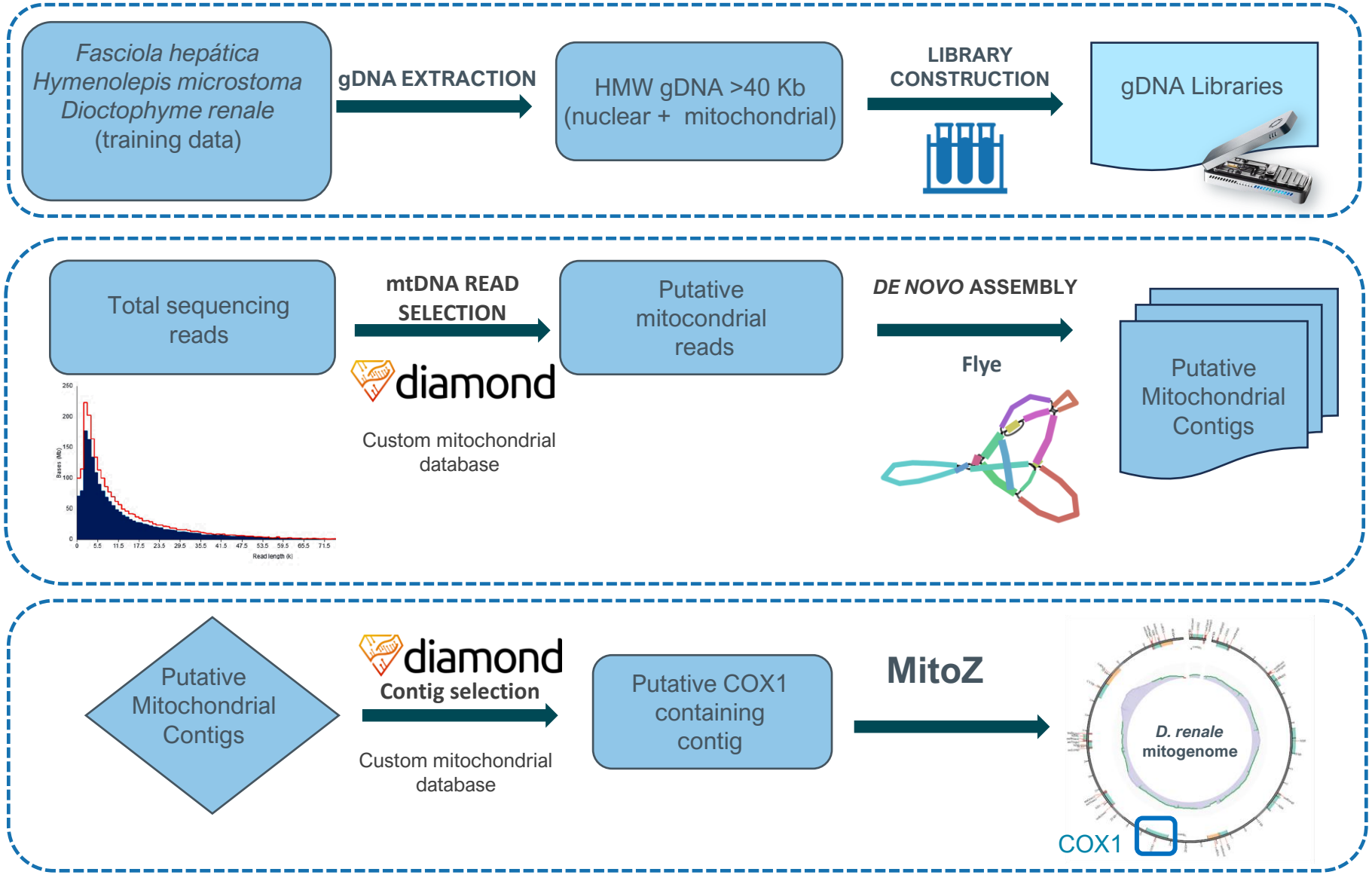
Nuclear: 100-300 Mb



Mitochondrial 15-40 Kb

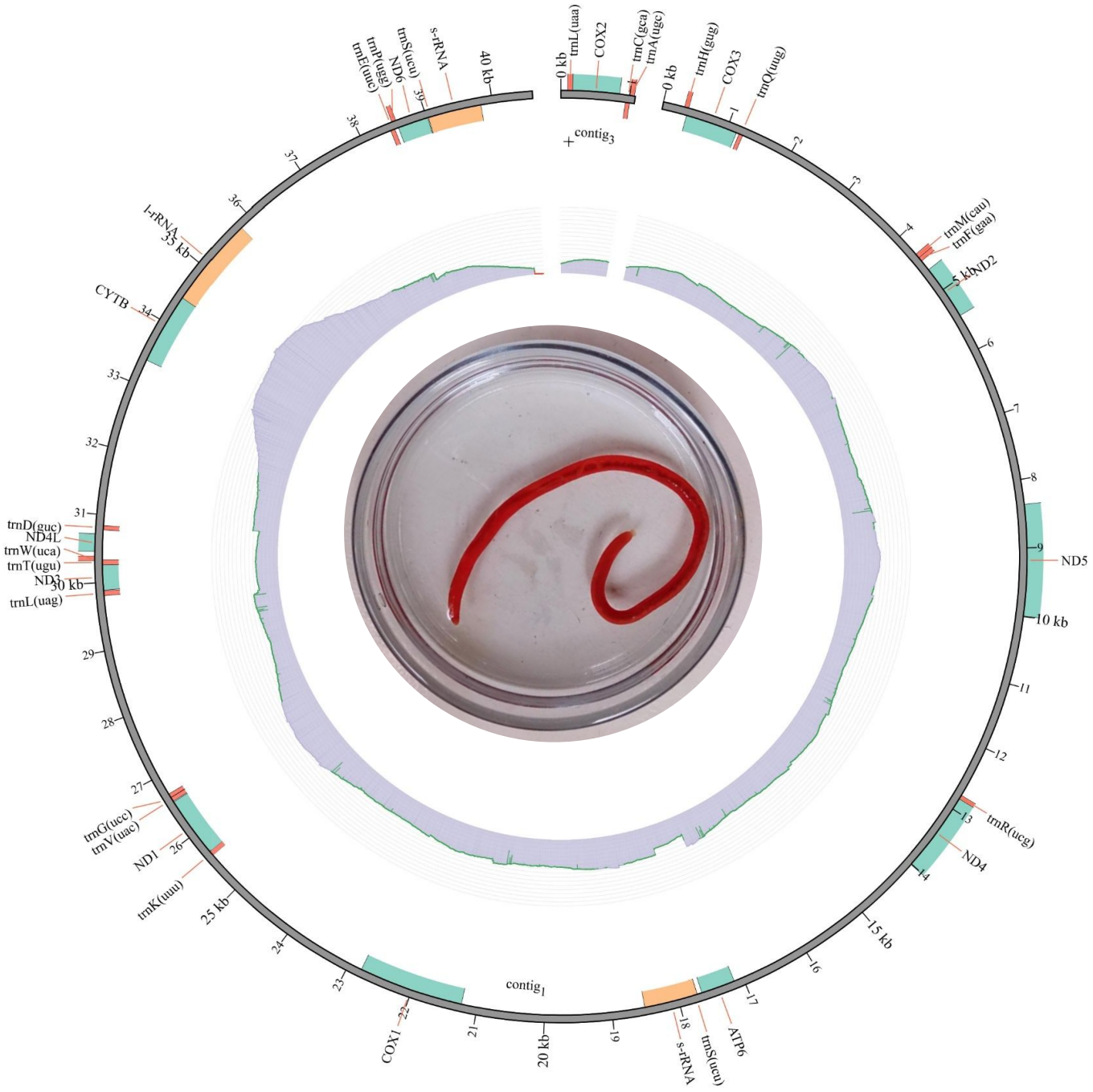
Bioinformatic strategies must be used to separate the nuclear information from the mitochondrial information

Mitogenome sequencing workflow



A similar workflow to the one employed in this course was used to obtained the *Dioctophyme renale* mitogenome

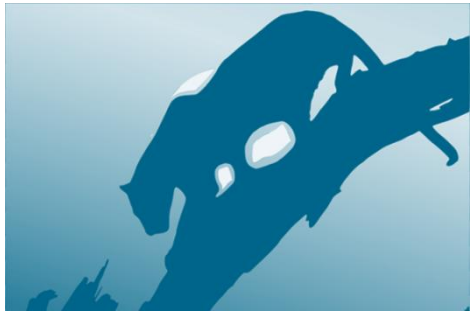
The annotated mitogenome was used to design molecular markers and to determine the phylogenetic relationships of parasites found in different countries and hosts



Parasite genetic variation was analyzed based on the newly generated mitochondrial data

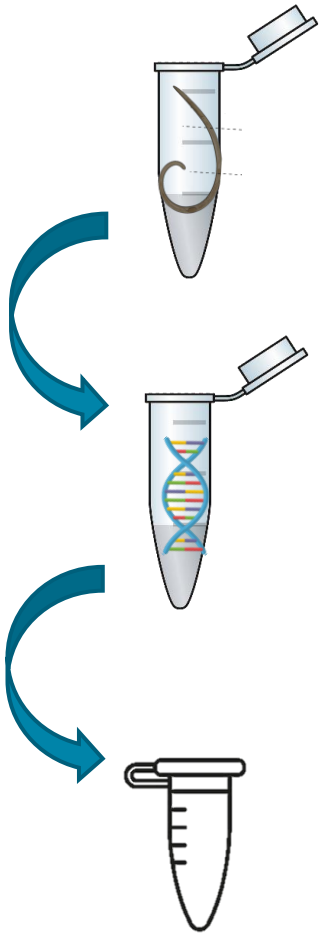
Domestic species

Habitat



Wild species

DNA extraction

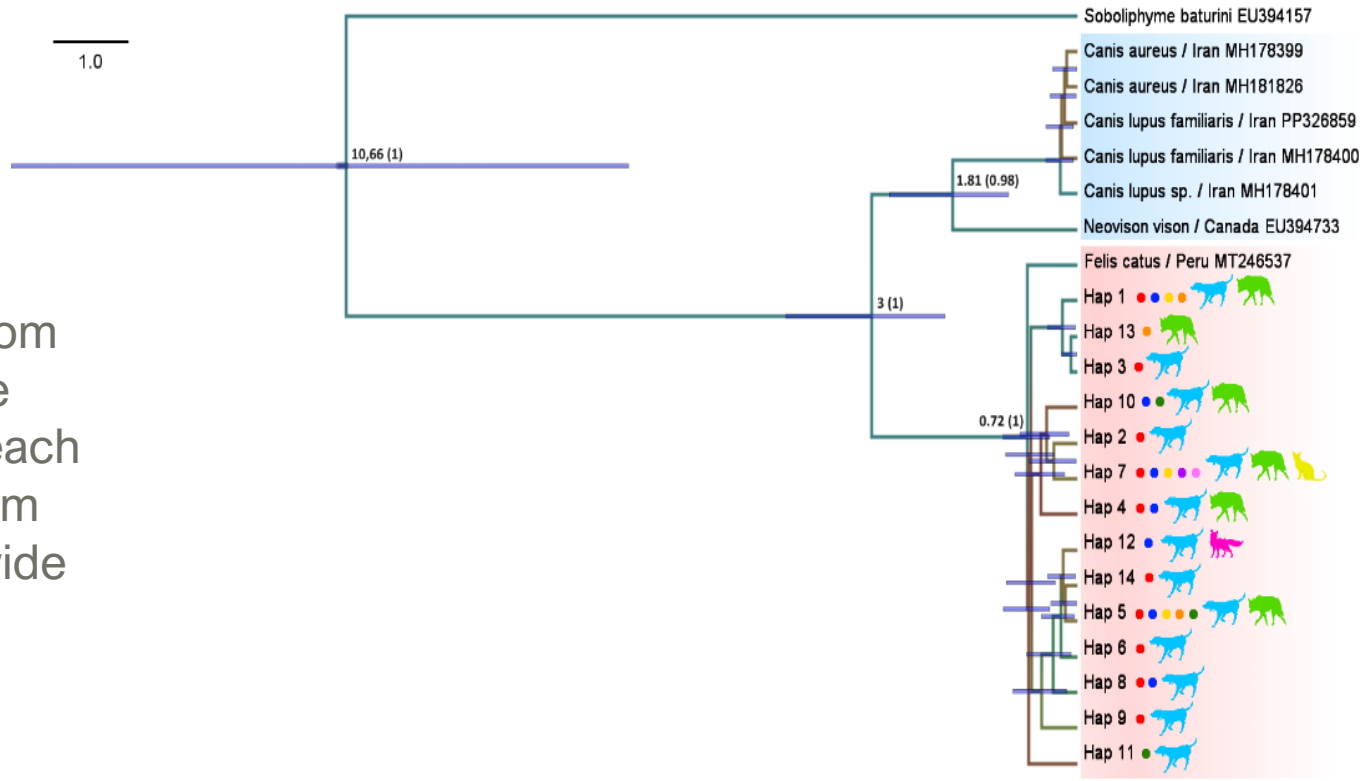


Amplicon-seq
COX1

We collected samples form domestic and wild host species
We amplified by PCR the new COX1 gene obtained from the mitogenome

We determine the phylogenetic relationships of parasites found in different countries and hosts

Parasites samples from South America are genetically closer to each other than those from other regions worldwide



Other regions

South America

The wild and domestic hosts share parasite populations

Chrysocyon brachyurus
(maned wolf)

Lycalopex gymnocercus
(pampas fox)

Felis catus
(cat)

Canis lupus familiaris
(dog)

Arce et al., 2024

Acknowledgments

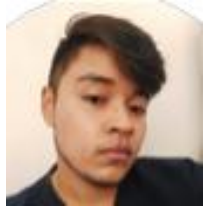
Genomics and Bioinformatics Group iB3

Agustín Baricalla, Mg
Doctoral Fellowship

Natalia Macchiaroli, PhD
Researcher

Juan Arrabal, PhD
Posdoctoral Fellowship

Lucas Arce, Mg
Doctoral Fellowship



Kevin Calupiña, Bq
Magister Fellowship

Ines Sananez, PhD
Posdoctoral Fellowship

Marina Ingravidi, Mgs
Doctoral Fellowship

Thanks!

Collaborators

Prof. Gisela Franchini
INBIOLP-UNLP-CONICET

CONICET

UNITED NATIONS UNIVERSITY
UNU-BIOLAC
Programme for Biotechnology in Latin America and the Caribbean

AGENCIA NACIONAL DE PROMOCION CIENTIFICA Y TECNOLOGICA

UBA: Universidad de Buenos Aires

analytical

Oxford NANOPORE Technologies

Perez Guerrero Foundation

THE ROYAL SOCIETY