wellcome connecting science

Mitochondrial Assembly.
Genome skimming strategy.

Mg. Agustin Baricalla
Universidad Nacional de San Luis.

May 2025

$$(dNMP)_n + dNTP \rightarrow (dNMP)_{n+1} + PPi + H^+$$

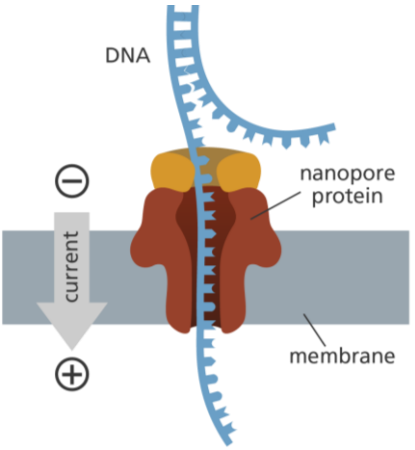
DNA chain (Nanopore)

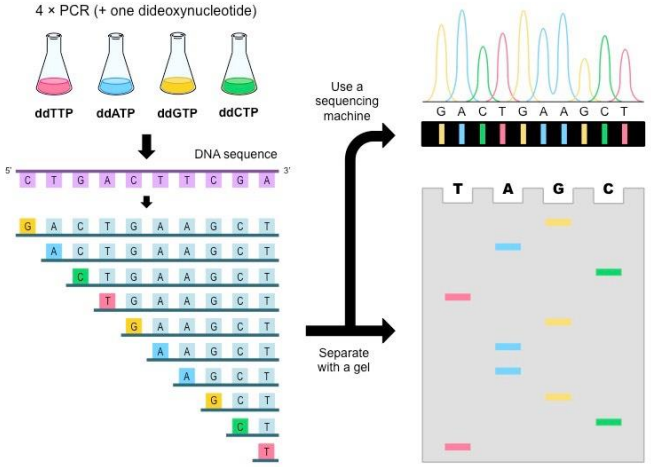
Nucleotide (modified dye)

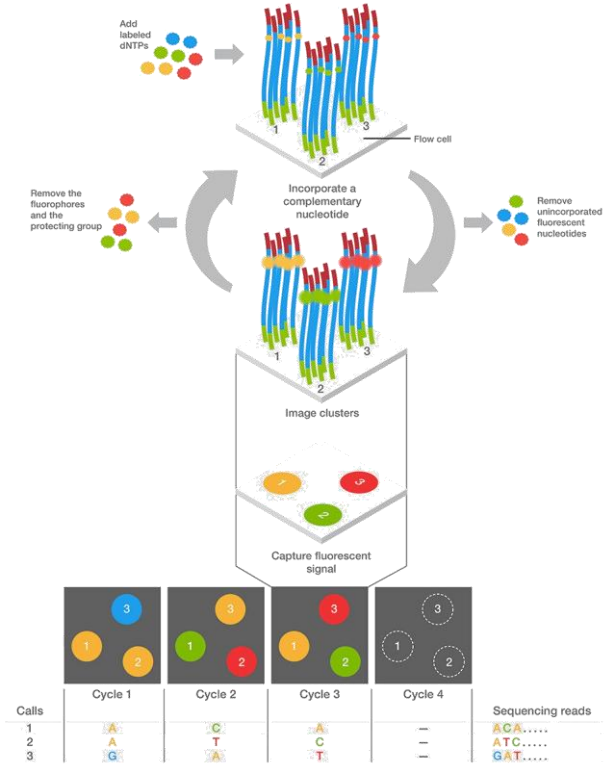
Illumina, Sanger, Pacbio

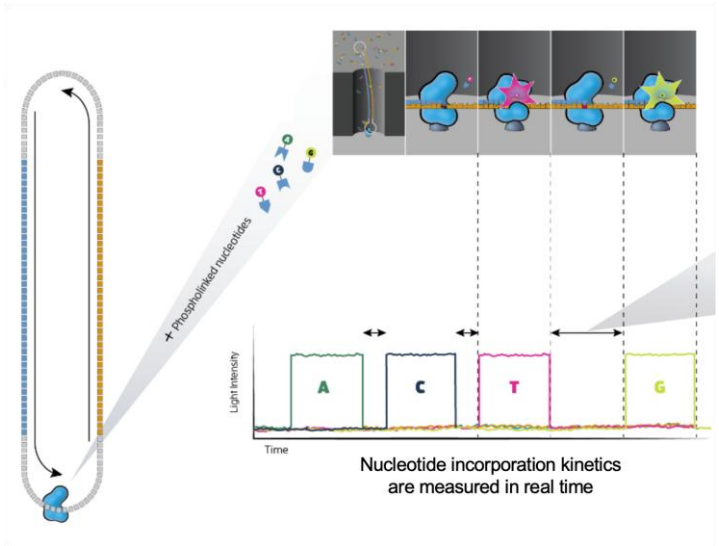
Pyrosequencing

Ion Torrent





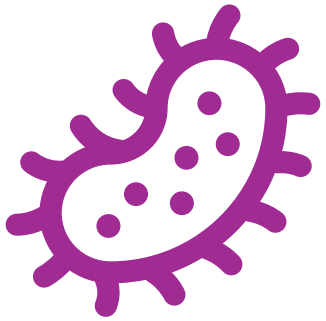




Step I.
The sample.

Previous knowledge of the biology of the organism.

Preliminary considerations - Sample



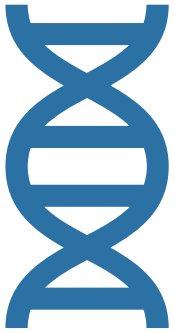
Nuclear + organellar + others

Extraction

Inhibitors (!)

Ploidy?

XY | X0 | ZW (?)



QC

Length

Sequencing

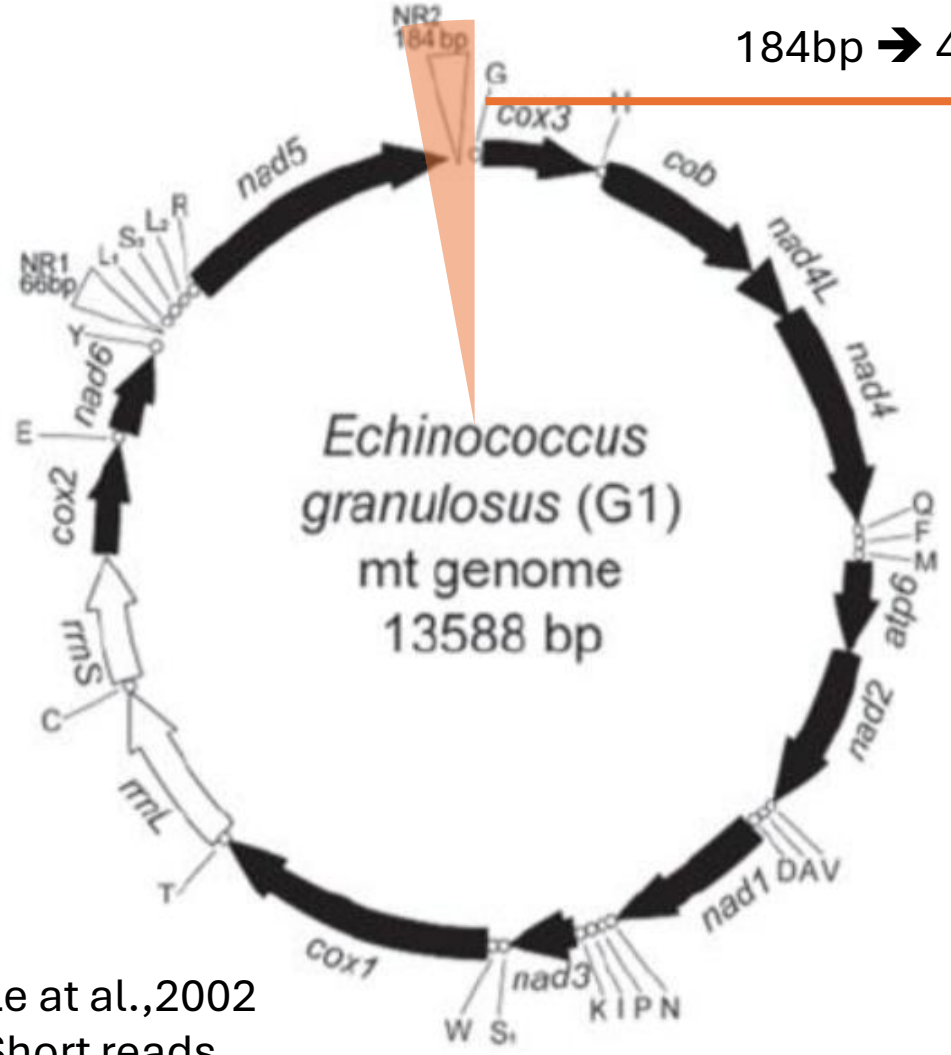
wellcome connecting science

Mitochondrial genome assembly

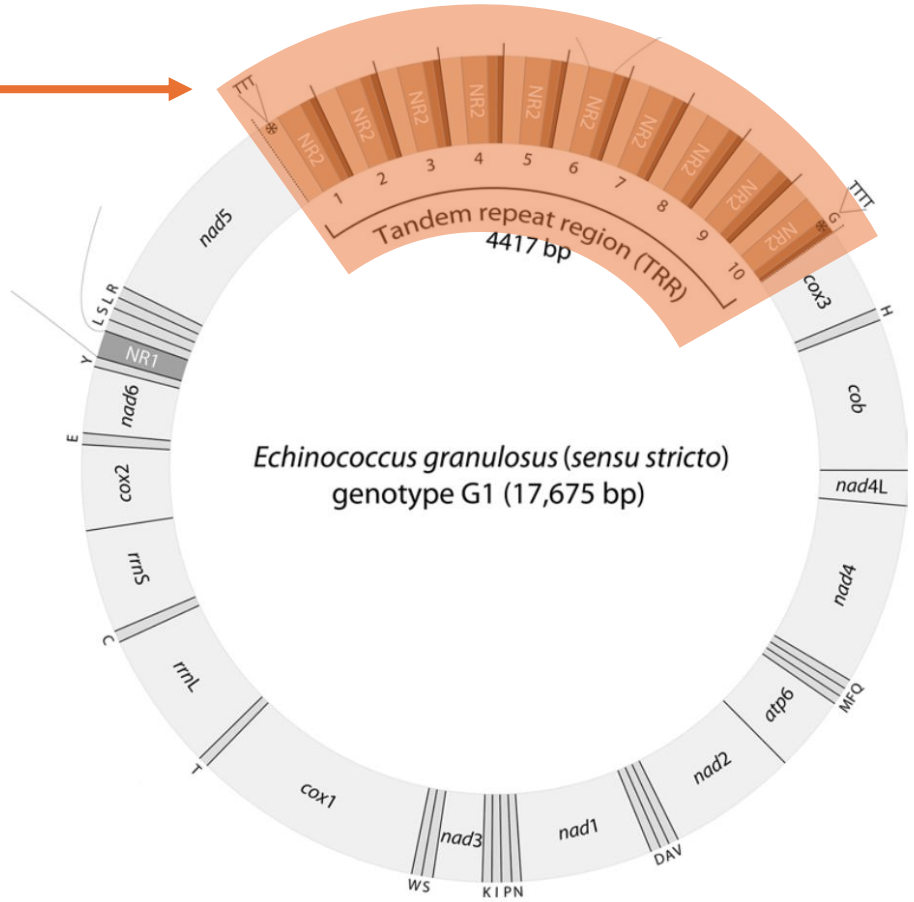
May 2025

Why we choose long reads for mitocondrial assembly?
mtDNA *Echinococcus granulosus:*

184bp ➔ 4417bp

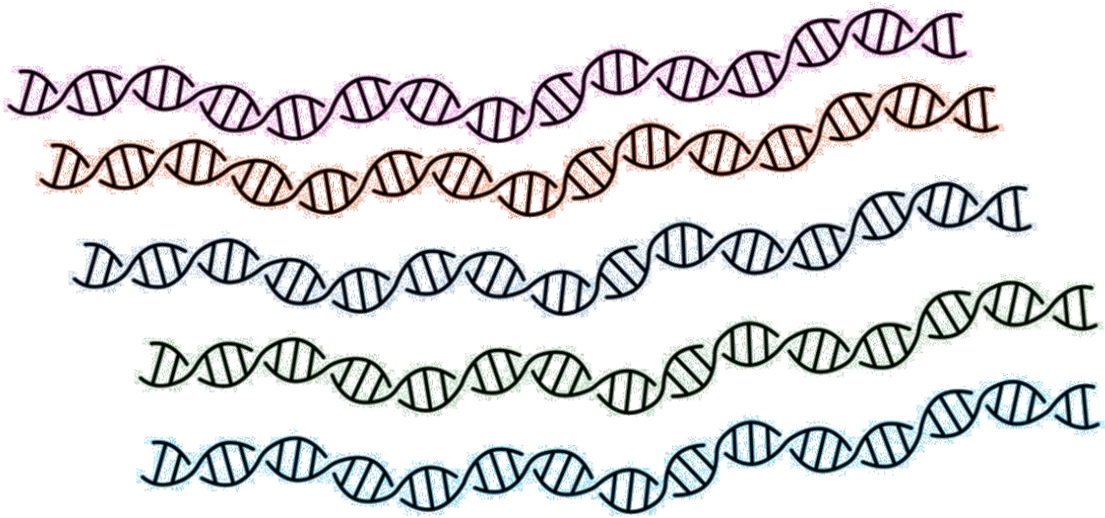


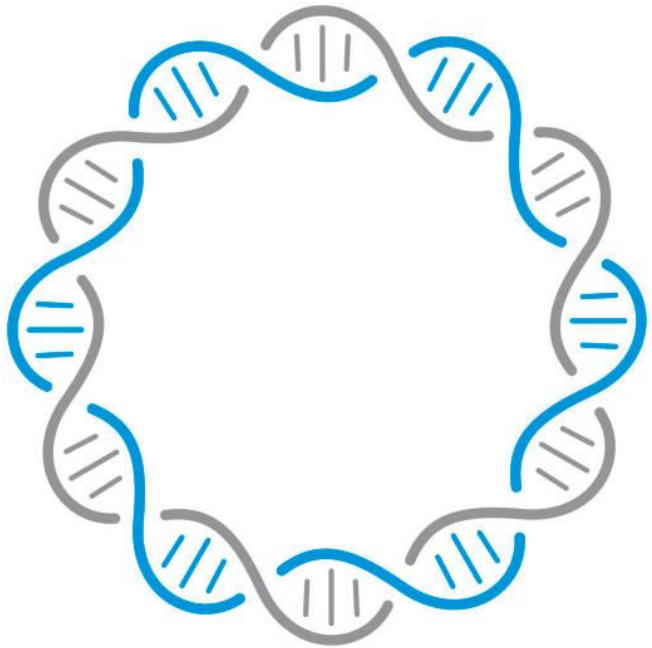
Le at al.,2002
Short reads



Kinkar et al.,2019
Long -reads

Mitochondrial genome assembly strategies





Reads:
Nuclear
Mitochondrial

Closely-related reference

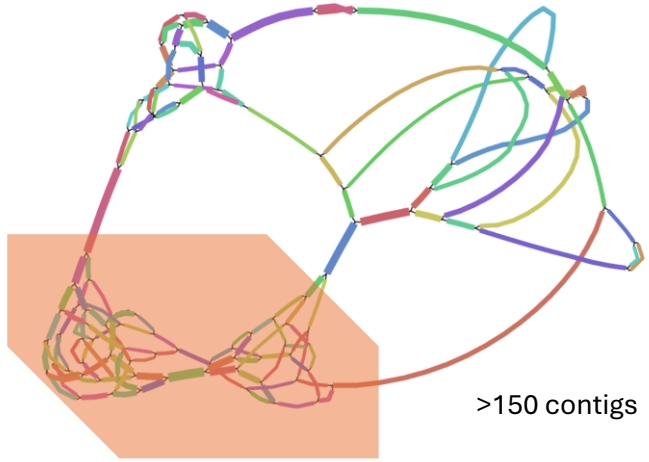
Genes or mitogenomes

Mapping
Alignment
Extension

Re-Mapping
Re-Alignment
Extension

Consensus

MitoBim

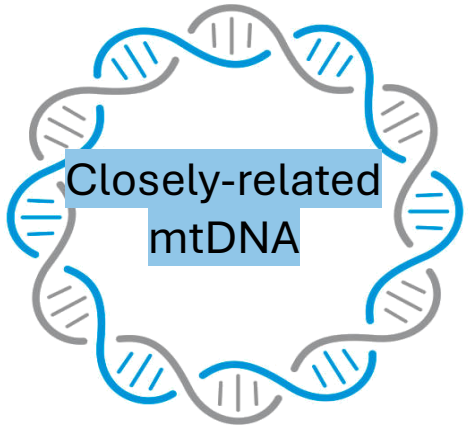


>150 contigs

Orange: complex repetitive region

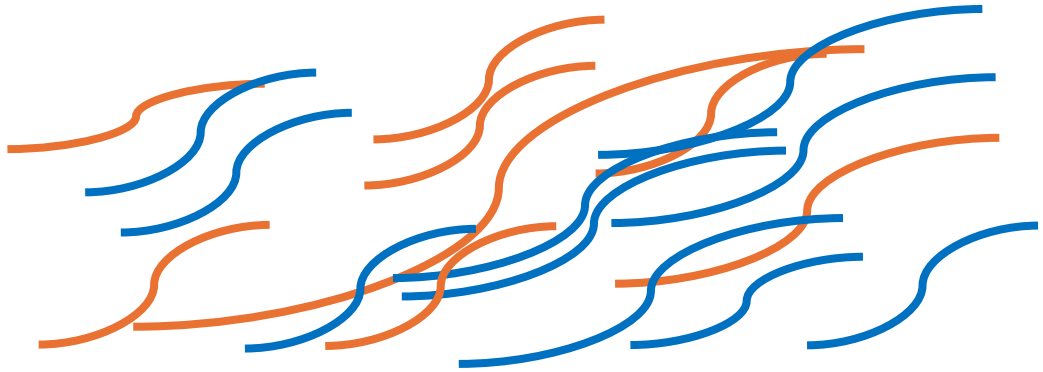
Bandage visualization of mtDNA

Highly repetitive sequences or tandem repeats



↑Nuclear

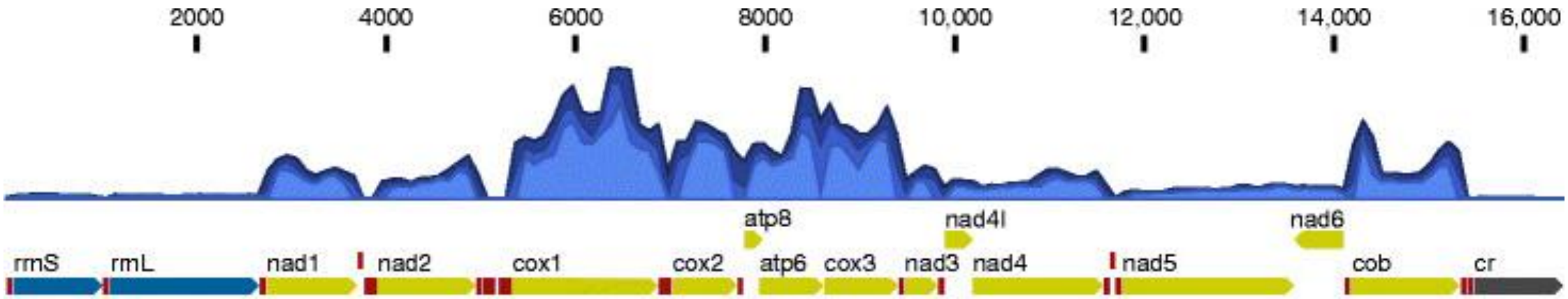
↑↑↑Mitochondrial

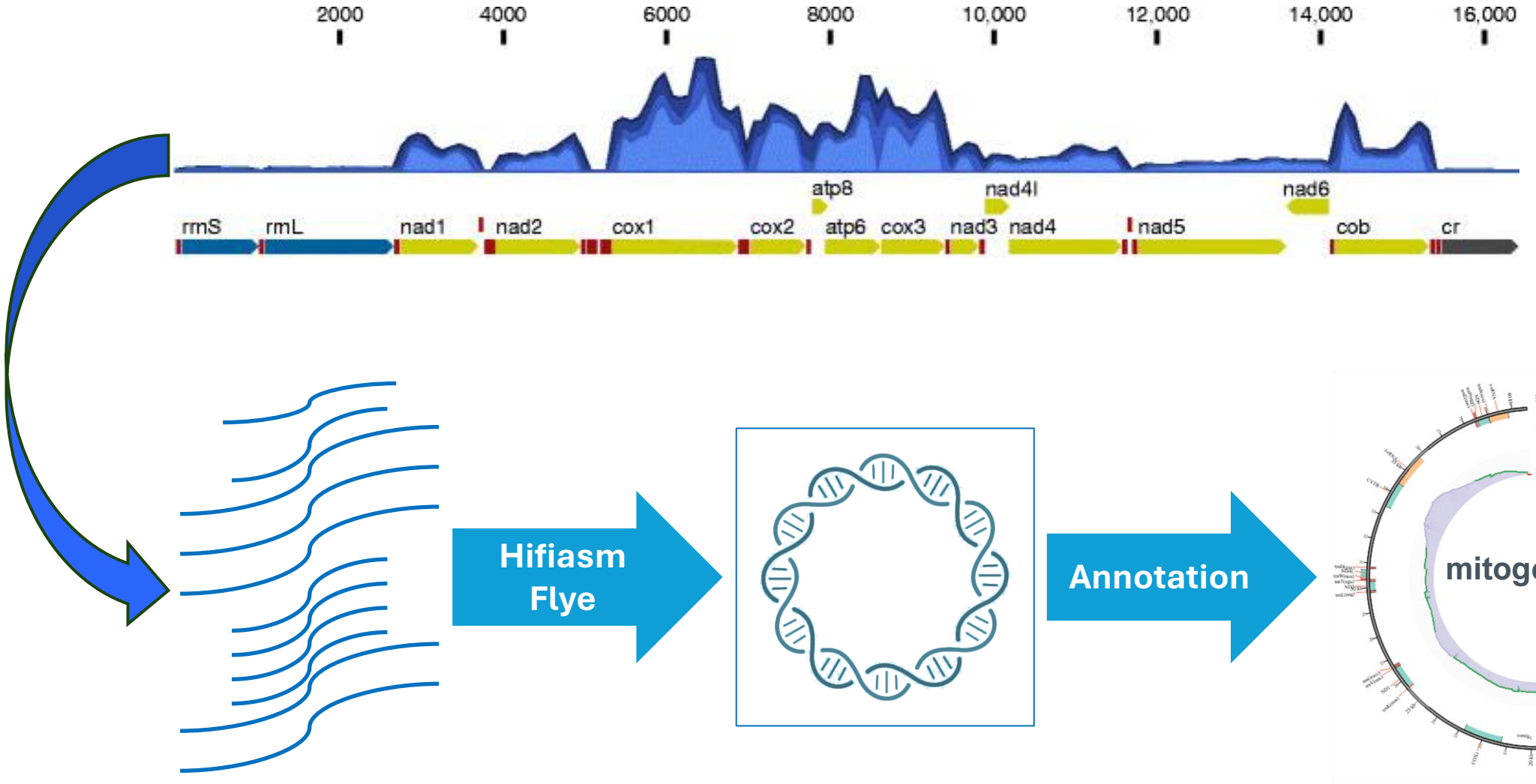


MitoVGP

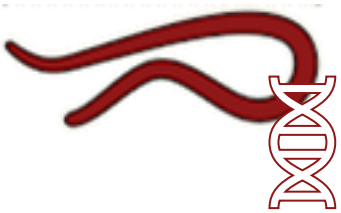
MitoHiFi

Minimap2




2000
4000
6000
8000
10,000
12,000
14,000
16,000
atp8
nad4l
nad6
rmS
rmL
nad1
nad2
cox1
cox2
atp6
cox3
nad3
nad4
nad5
cob
cr
ALIGNED READS (COVERAGE)
REFERENCE MITOGENOME
Hifiasm
Flye
Annotation
mitogenome

Pipeline

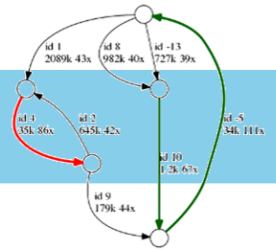


DNA extraction

ONT sequencing

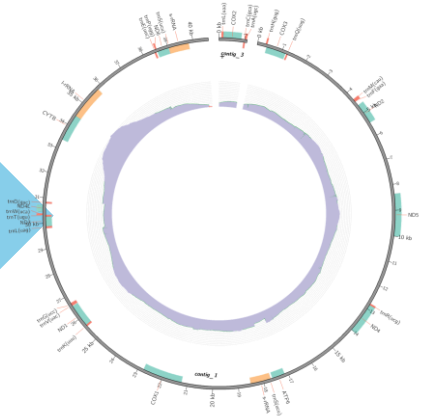
Basecalling & Data QC

Diamond mtDNA reads selection



Flye assembly

Mitos2 annotation



Workflow

Sampling Sequencing

Dioctophyme renale N=2

gDNA EXTRACTION

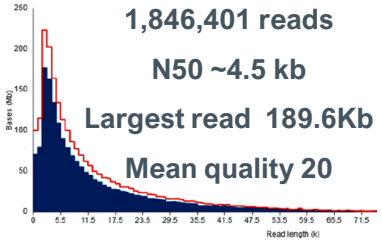
HMW gDNA >40 Kb (nuclear + mitochondrial)

LIBRARY CONSTRUCTION

gDNA Libraries N=4

QC Assembly

PromethION
Basecalling SUP



1,846,401 reads

N50 ~4.5 kb

Largest read 189.6Kb

Mean quality 20

mtDNA READ SELECTION

diamond

Nematode clade I mitochondrial database

Putative mitochondrial reads

4,664 reads

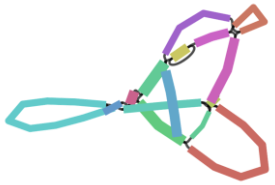
N50 ~4 kb

Largest read 18.1Kb

Mean quality 19

***DE NOVO* ASSEMBLY**

Flye



Mitochondrial Contigs

Annotation Molecular markers design

12 CDS
22 tRNAs
2 rRNAs

MITOS

diamond
FILTERING

Nuclear reads removal

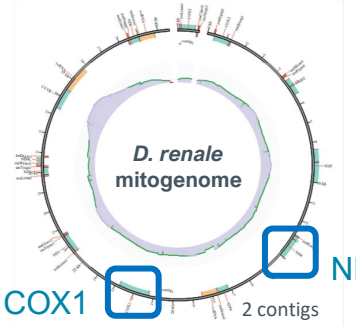
CDS identity > 60%

CDS coverage > 90%

High confidence mitochondrial reads

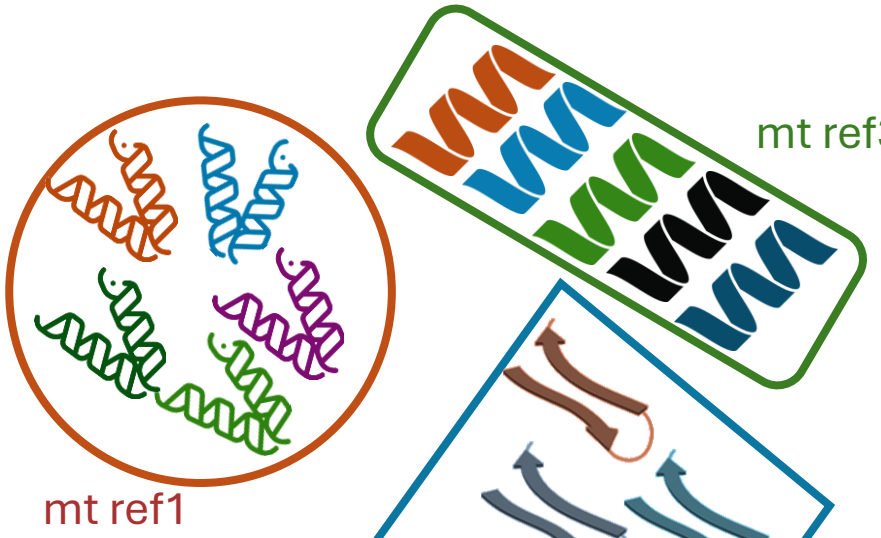
120 reads
N50 ~2.5 kb
Coverage **7.2X**

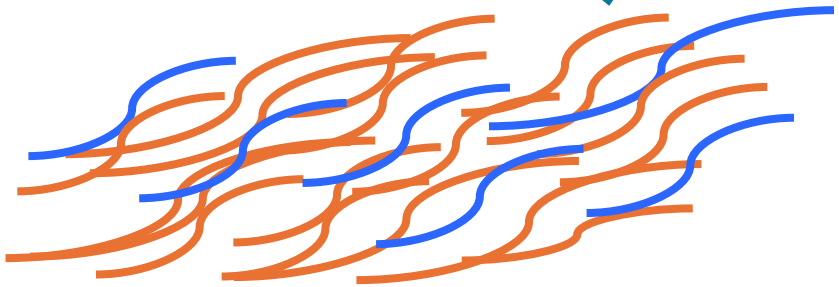
MOLECULAR MARKER DESIGN



BIONFO

Mitochondrial genome *de novo*
Lab10 - pipeline



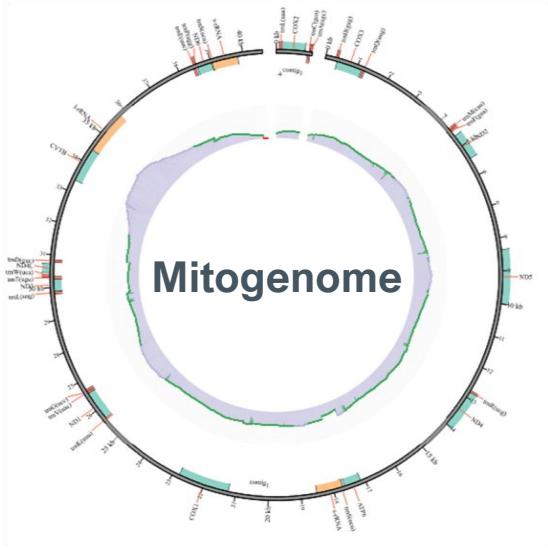


↑↑↑Nuclear
↓Mitochondrial

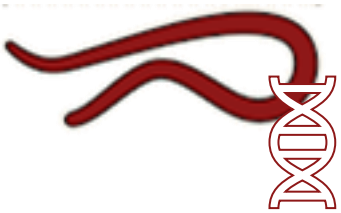
diamond

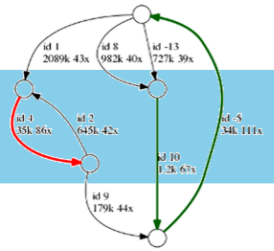
-Based on proteins
-(Non) Closely related
- Reads selection

Flye

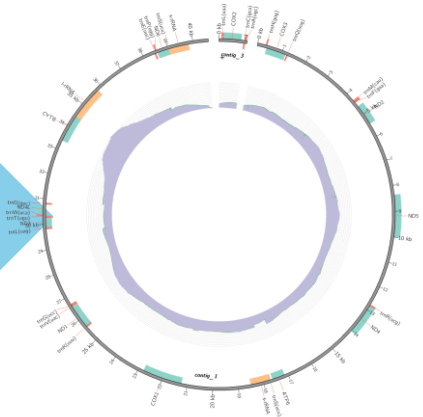


Pipeline





MITOS



DNA extraction

ONT sequencing

Basecalling & Data QC

Diamond mtDNA reads selection

Flye assembly

Mitoz annotation

☑ ☑

Readings files format

FASTQ format

- Simple format of unaligned sequence reads
- Sequences and quality score associated with each base.
- File extension : .fastq o .fq
  - (Compressed fastq.**gz** | fq.**gz**)

Readings files format

FASTQ format

Read ID = '@' sign + sequence identifier

Read sequence

'+' sign for parity

Read Quality by base (ASCII code)

```
@IL16_4408:3:5:17860:13258
CTGGCTCACATACAGGCCAGTATAAAGCGTCTCCTTTAAA
+
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHE
@IL16_4408:3:14:13276:1210
GTAGAAAACAATATAGAAGCCTTAGGACAGAAAACCCTAA
+
HHHHHHHHHHHHHHHHHDHHHHHHHHBGHHHHHHCDCCF>
```

Nanoplot - **QC**

```
NanoPlot
-t 8
--fastq minion.fq
--dpi 300
--N50
-o ./nanoplot
--huge
```

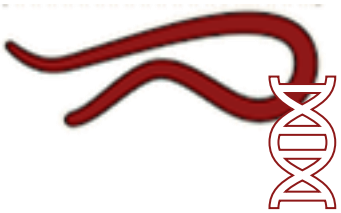
Program (BD creation)
Threads
Fastq file name
Image quality
To show the N50 value mark
Output folder name
Big data file

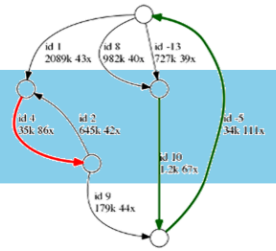
Nanoplot - **QC**

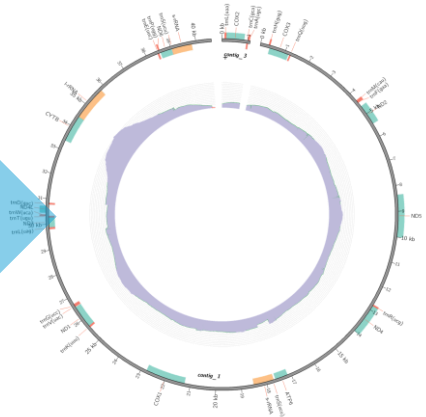
Summary statistics

Metrics	
number_of_reads	26190
number_of_bases	686824347.0
median_read_length	12433.0
mean_read_length	26224.7
read_length_stdev	35320.1
n50	52343.0
mean_qual	18.8
median_qual	21.7
longest_read_(with_Q):1	560959 (13.3)
longest_read_(with_Q):2	540464 (14.5)
longest_read_(with_Q):3	531365 (15.8)
longest_read_(with_Q):4	421500 (18.3)
longest_read_(with_Q):5	373488 (23.4)
highest_Q_read_(with_length):1	32.5 (5063)
highest_Q_read_(with_length):2	31.9 (9176)
highest_Q_read_(with_length):3	30.9 (5644)
highest_Q_read_(with_length):4	29.7 (43859)
highest_Q_read_(with_length):5	29.5 (7944)
Reads >Q10:	26190 (100.0%) 686.8Mb
Reads >Q15:	24320 (92.9%) 632.1Mb
Reads >Q20:	18101 (69.1%) 469.3Mb
Reads >Q25:	1461 (5.6%) 21.5Mb

Pipeline







DNA extraction

ONT sequencing

Basecalling & Data QC

Diamond mtDNA reads selection

Flye assembly

MitoZ annotation

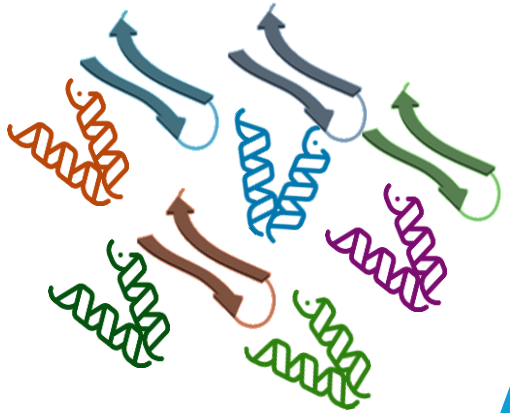
Diamond BLAST

DIAMOND is a sequence aligner for protein and translated DNA searches vs Protein database.

Key features:

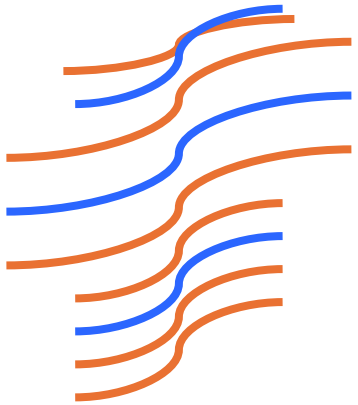
- Pairwise alignment of proteins or DNA.
- Faster than BLAST.
- Allows alignments with reading frame changes for analysis of long reads (error rate tolerance).
- Various **output formats**, including **tabular** and **fasta/q**.

Diamond BLAST

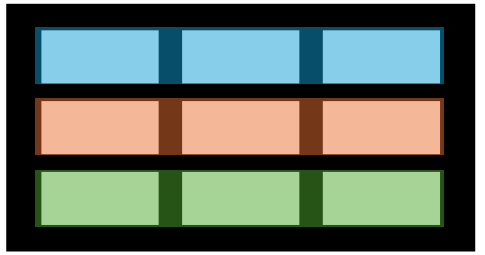


PROTEINS
FASTA

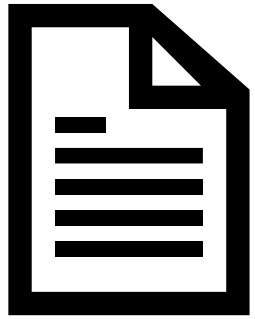
NUCLEIC ACIDS
FASTA/Q



PROTEIC
DATABASE

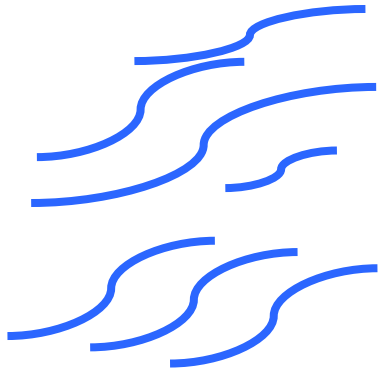


TABULAR FILE



FASTA/Q
MATCHING
SEQUENCES

NUCLEIC ACIDS
FASTA/Q



Diamond - outformat 6

qseqid	sseqid	qlen	slen	pident	length	mismatch	gapopen	qstart	qend	sstart	send	evalue	bitscore	nident	positive	ppos
300262.1_14095	contig_1_ATP6_len=168_17134_17639_+	14095	168	99.4	168	1	0	12145	11642	1	168	3.24E-100	309	167	168	100
38576.1_8924	contig_1_COX1_len=520_21238_22799_+	8924	520	100	422	0	0	68	1333	1	422	5.02E-283	838	422	422	100
599553.1_8749	contig_1_CYTB_len=372_33381_34500_+	8749	372	100	370	0	0	2671	3780	3	372	1.82E-246	732	370	370	100
599553.1_8749	contig_1_ATP6_len=168_17134_17639_+	8749	168	99.4	168	1	0	7962	8465	1	168	1.99E-100	309	167	168	100
455962.1_8005	contig_1_ND4_len=393_12828_14008_-	8005	393	100	259	0	0	4122	3346	135	393	5.77E-152	467	259	259	100
						...										
65512.1_7635	contig_1_ND1_len=299_25659_26558_+	7635	299	100	116	0	0	1955	1608	1	116	3.22E-64	211	116	116	100

Diamond - outformat 6

Qseqid	sseqid	qlen	slen	pident	length	mismatch	gapopen	qstart	qend	sstart	send	evalue	bitscore	nident	positive	ppos
65512.1_7635	ND1	7635	299	100	116	0	0	1955	1608	1	116	3.22E-64	211	116	116	100
65512.1_7635	COX1	7635	520	100	422	0	0	6378	5113	1	422	7.27E-286	838	422	422	100

Diamond tabular

qseqid	Query sequence ID.
sseqid	ID of the subject sequence (**the one in the database).**
qlen	Length of the query sequence.
slen	Length of the subject sequence.
pident	Percentage of identity between the query sequence and the subject sequence.
length	Length of alignment.
mismatch	Number of misalignments (differences) in the alignment.
gapopen	Number of gap openings in the alignment.
qstart	Beginning of the alignment in the query sequence.

qend	End of alignment in the query sequence.
sstart	Beginning of the alignment in the subject sequence.
send	End of alignment in the subject sequence.
evalue	E-value, which indicates the number of alignments expected by chance with a score equal or better. Lower values indicate higher significance.
bitscore	Bit score of the alignment, a measure of the size of the alignment.
nident	Number of identities (identical base pairs) in the alignment.
positive	Number of positive base pairs (similar amino acids) in the alignment.
ppos	Percentage of positive positions in the alignment.

Parameter highlights

qseqid	ID of the query sequence.
sseqid	ID of the subject sequence **(the one in the database)**.

evalue	E-value, which indicates the number of alignments expected by chance with a score equal or better. **Lower values indicate higher significance.**
bitscore	Alignment bit score, a measure of alignment size.

https://www.ncbi.nlm.nih.gov/books/NBK62051/

Database preparation

`diamond makedb`	Program (BD creation)
`--threads 4`	Threads to be used
`--db cladoi.dmnd`	Database name
`--in proteins_mt.fa`	Protein file for indexing

Diamond Search

```
diamond blastx
--threads 4 --db cladoi.dmnd
--query minion.fq.gz
--query-gencode 5 --ultra-sensitive
--out mt.outfmt6.tsv --outfmt 6 --header
--unal 0 --alfmt fastq
--al minion_drenale.fq
```

Diamond Search

```
diamond blastx
--query minion.fq.gz
--db cladoi.dmnd
--query-gencode 5
--ultra-sensitive
--out mt.outfmt6.tsv
--unal 0
--alfmt fastq
--al minion_drenale.fq
```

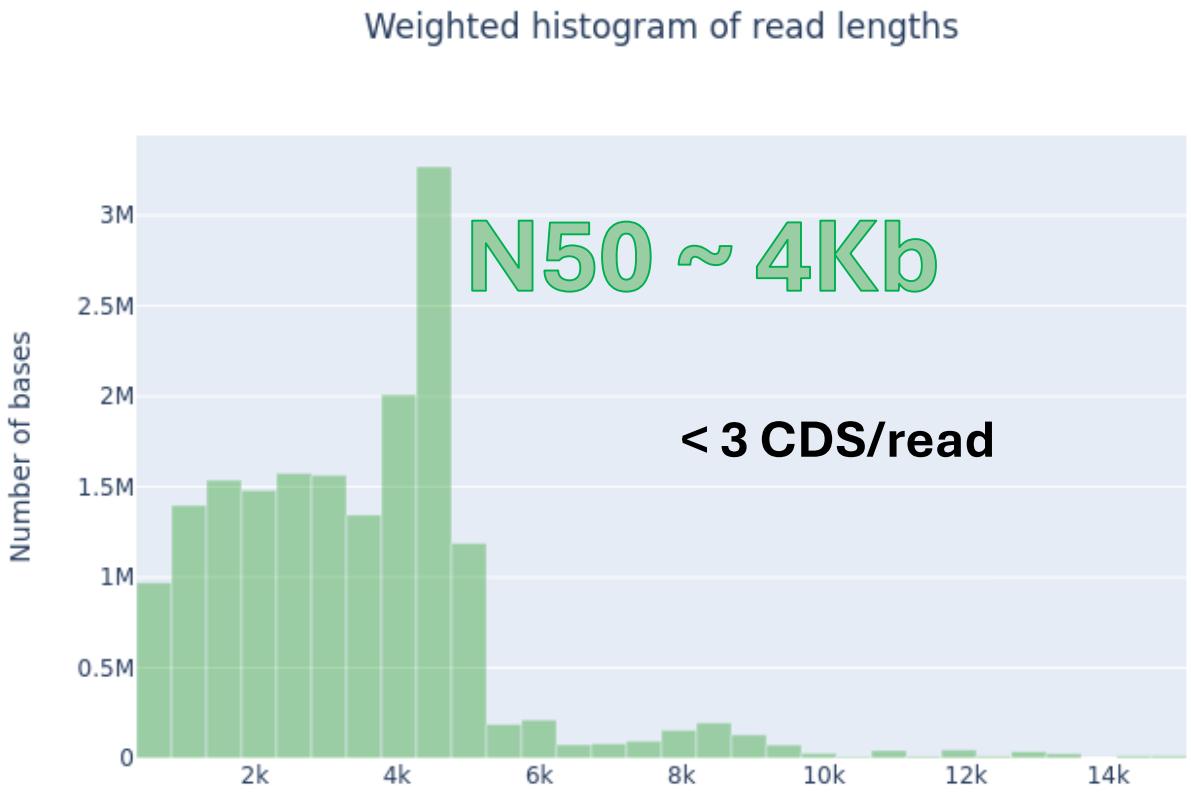
Program (BLAST **X / P**)
File input (**sequenced**)
Database
Codon usage table (**5 –invertebrates, 9-Platyhelminthes**)#
Sensitivity *
Name **Output table**
Non-aligned reads are not saved
Output format(fastA/Q)
Output file name

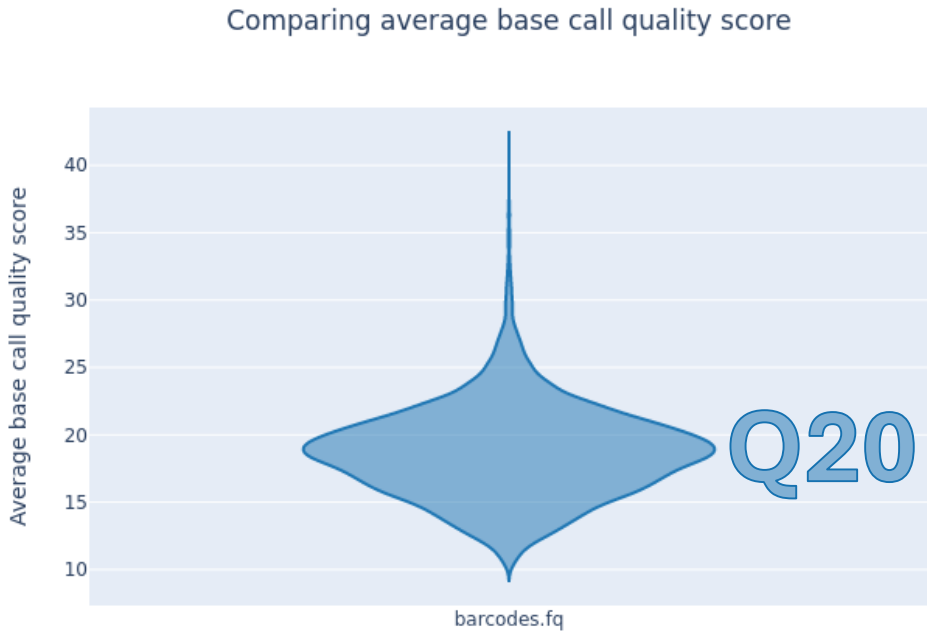
- * https://github.com/bbuchfink/diamond/wiki/3.-Command-line-options#sensitivity-modes
- # https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi

Alternative genetic codes

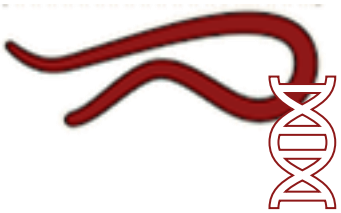
1. nucleotide	2. nucleotide: U	2. nucleotide: C	2. nucleotide: A	2. nucleotide: G	3. nucleotide
U	UUU F	UCU S	UAU Y	UGU C	U
	UUC F	UCC S	UAC Y	UGC C	C
	UUA L	UCA S	UAA St, Q_6	UGA St, $W_{1,2,3,4,5}$	A
	UUG L	UCG S	UAG St, Q_6	UGG W	G
C	CUU L, T_5	CCU P	CAU H	CGU R	U
	CUC L, T_5	CCC P	CAC H	CGC R	C
	CUA L, T_5	CCA P	CAA Q	CGA R	A
	CUG L, T_5, S_7	CCG P	CAG Q	CGG R	G
A	AUU I	ACU T	AAU N	AGU S	U
	AUC I	ACC T	AAC N	AGC S	C
	AUA I, $M_{2,3,4,5}$	ACA T	AAA K, N_1	AGA R, $S_{1,2}$, G_3, St_4	A
	AUG M	ACG T	AAG K	AGG R, $S_{1,2}$, G_3, St_4	G
G	GUU V	GCU A	GAU D	GGU G	U
	GUC V	GCC A	GAC D	GGC G	C
	GUA V	GCA A	GAA E	GGA G	A
	GUG V	GCG A	GAG E	GGG G	G

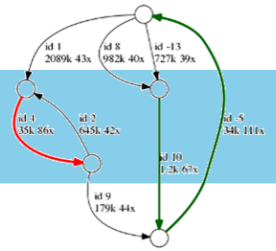
Read Length & Quality

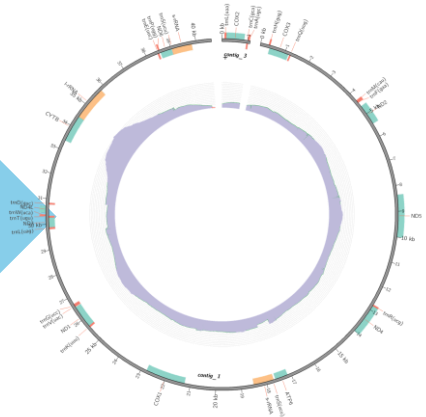




Pipeline







DNA extraction

ONT sequencing

Basecalling & Data QC

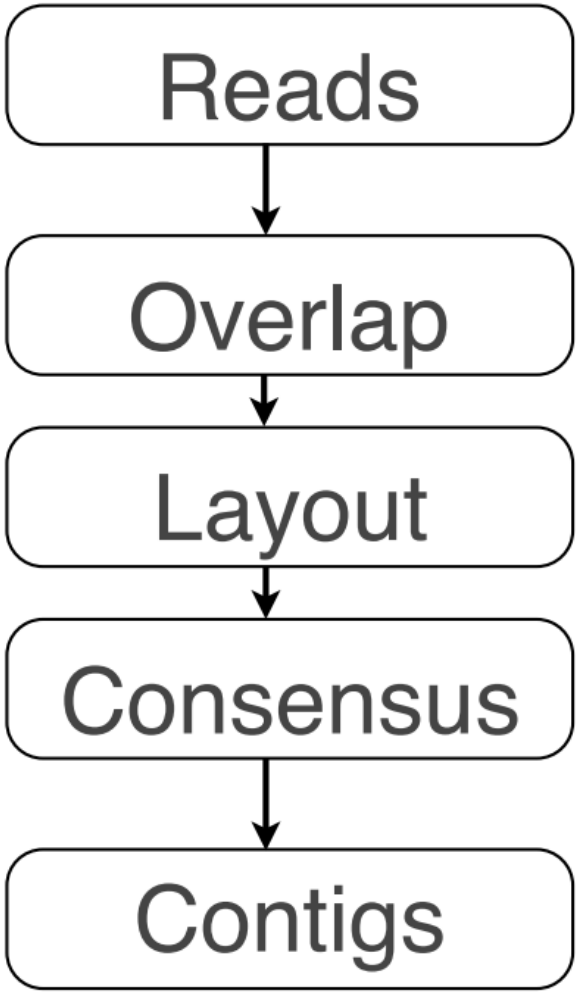
Diamond mtDNA reads selection

Flye assembly

Mitoz annotation

☑ ☑ ☑ ☑

Long Read Assembly Pipeline

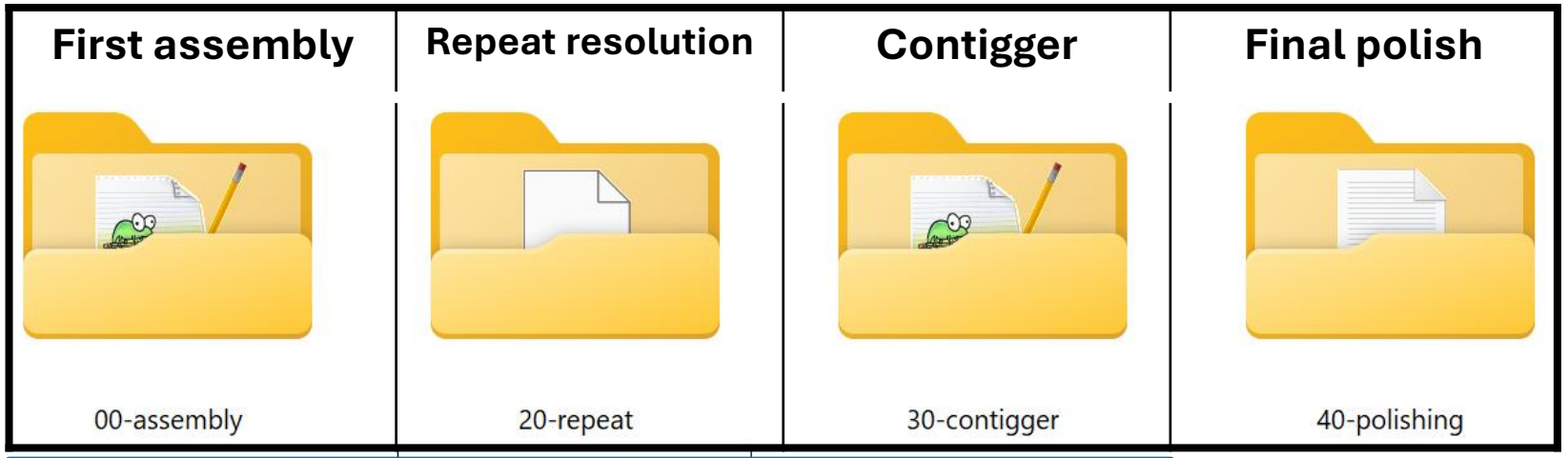


Build overlap graph

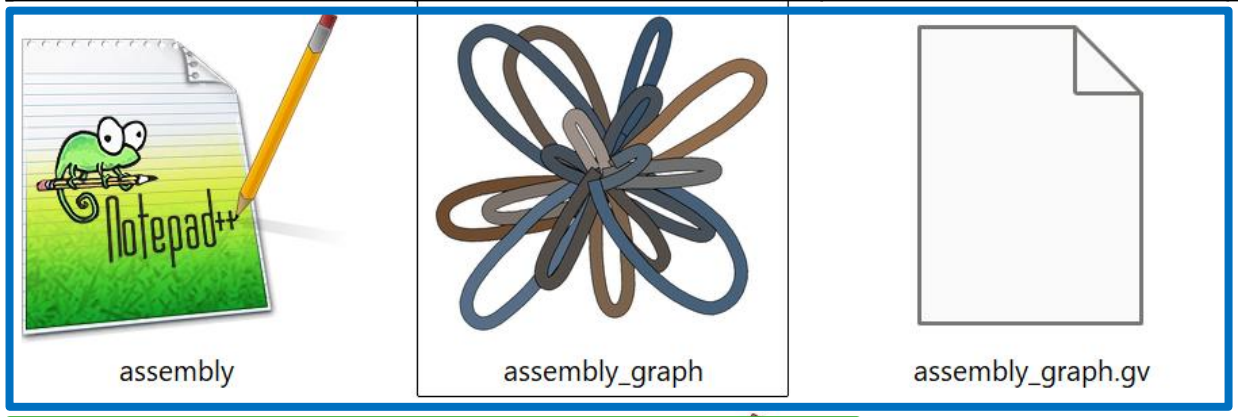
Bundle stretches of the overlap graph into contigs

Pick most likely nucleotide sequence for each contig

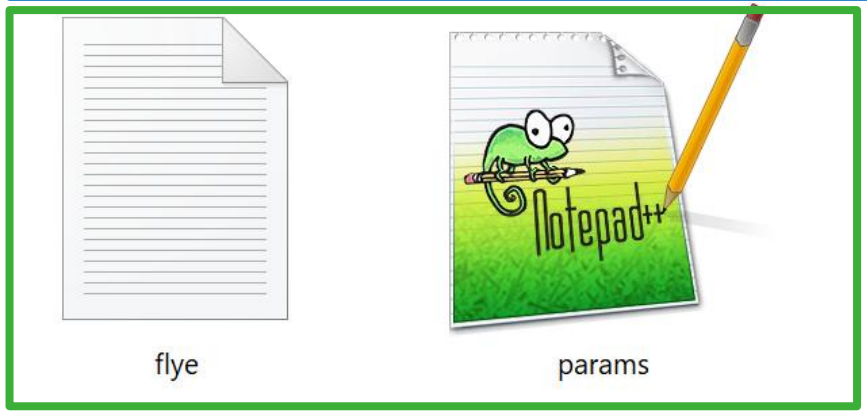
Flye



Final Assembly files

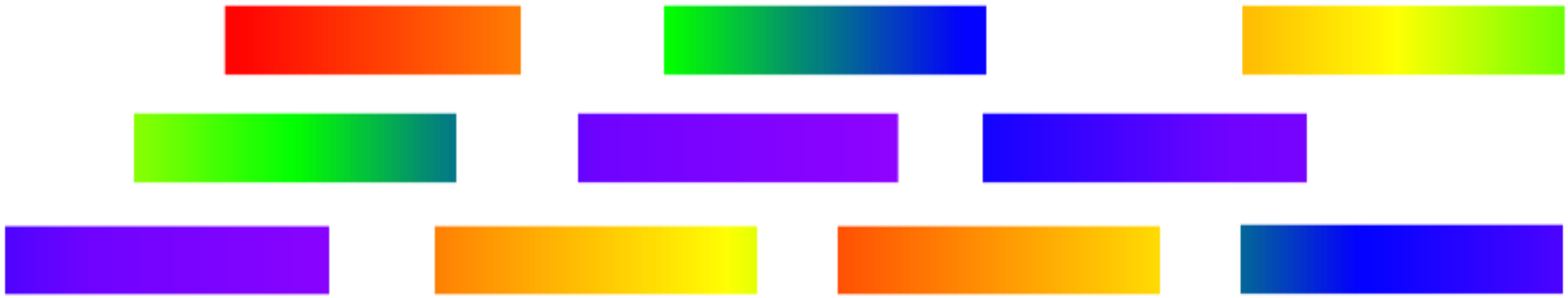


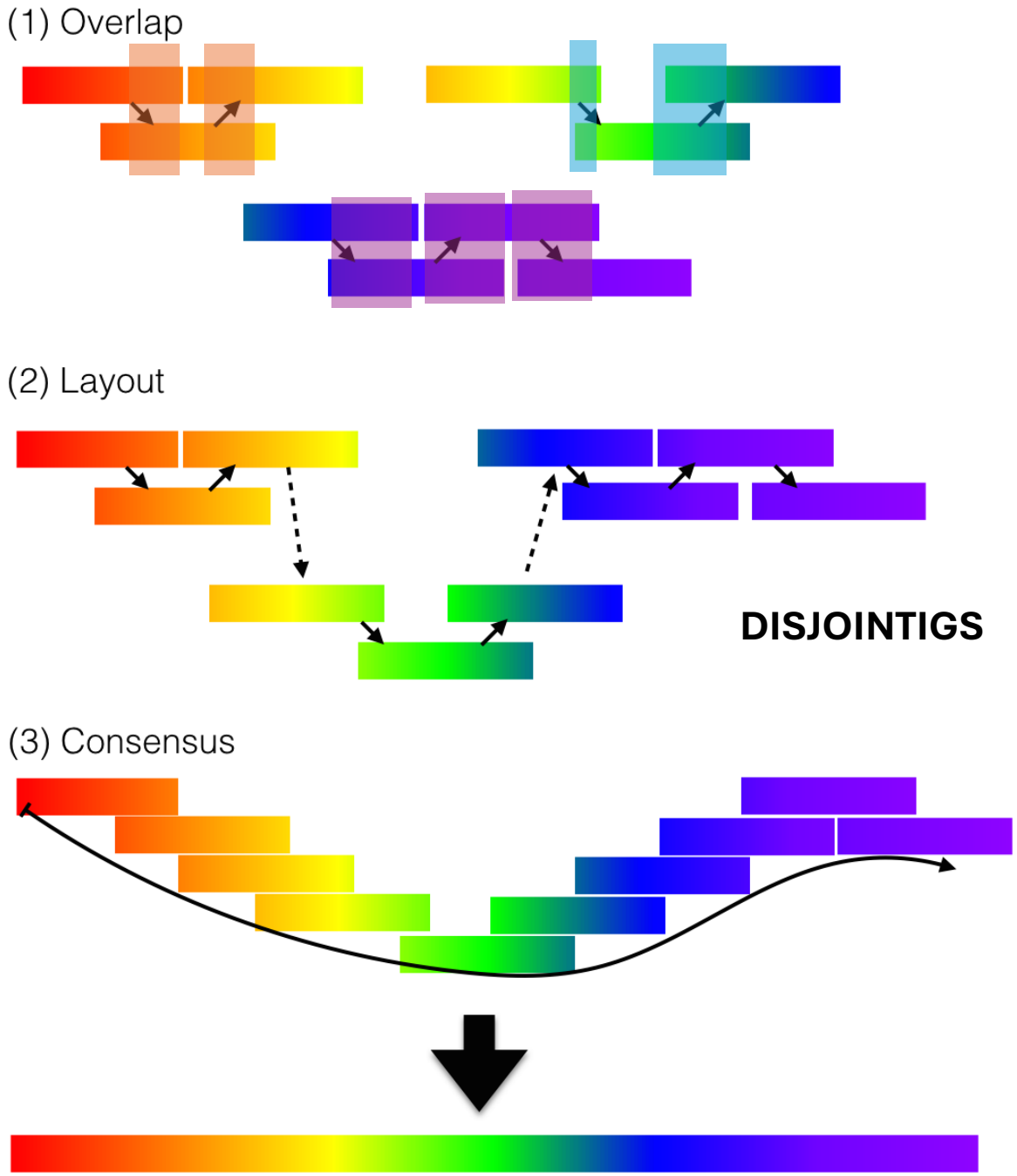
Flye log

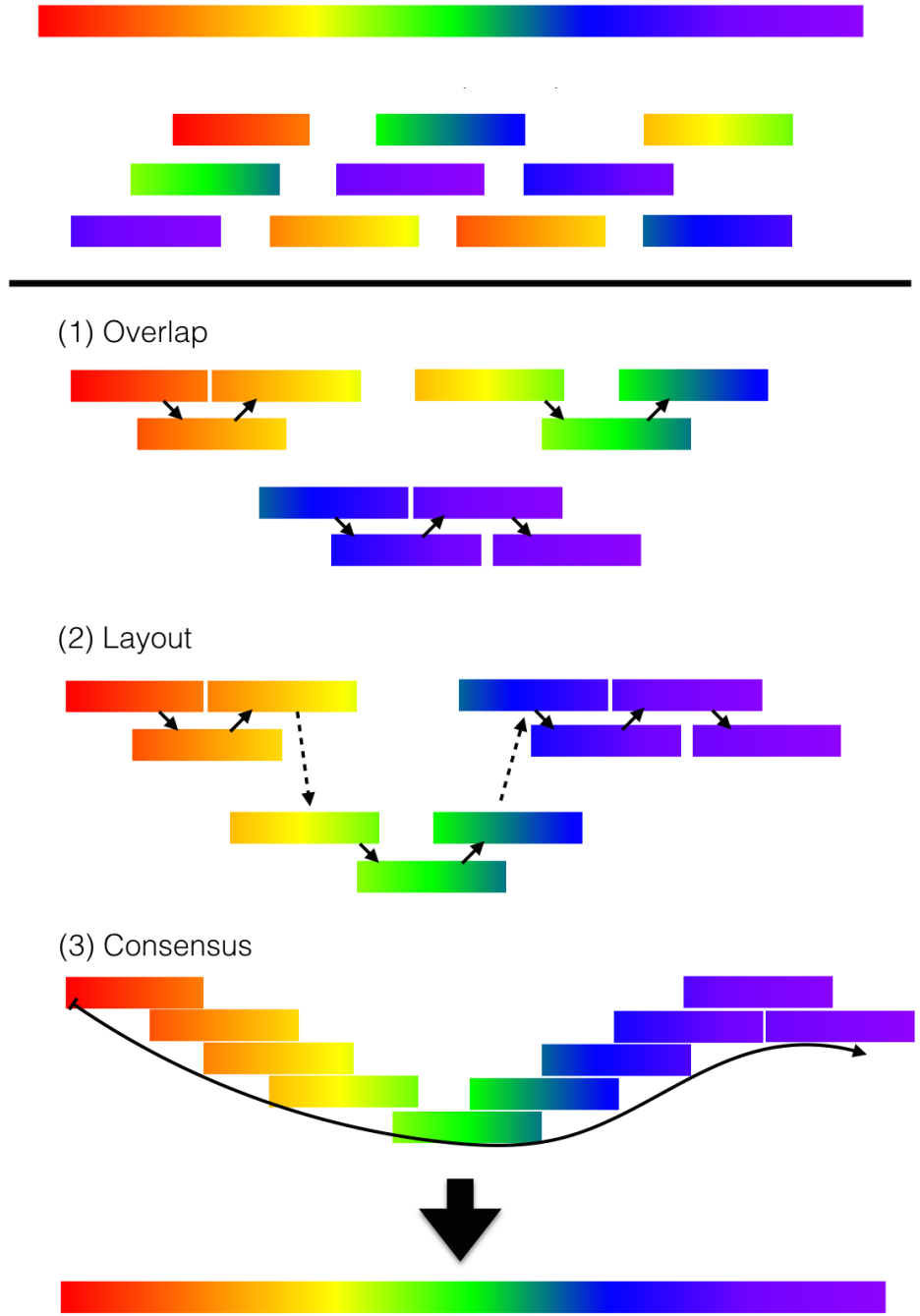


GENOME

READS




(1) Overlap
(2) Layout
DISJOINTIGS
(3) Consensus


(1) Overlap
(2) Layout
(3) Consensus

Assembly with Flye

```
flye
--nano-raw drenale.fq
-t 4
--meta
--keep-haplotypes
-o ./flye
```

Program
Input file (**sequenced**)
Threads
For metagenomes or **with variable coverage**
Maintain haplotypes (do not collapse alternative haplotypes)
Name **Output folder**

Sequencing methodology-dependent options:

OUR CASE

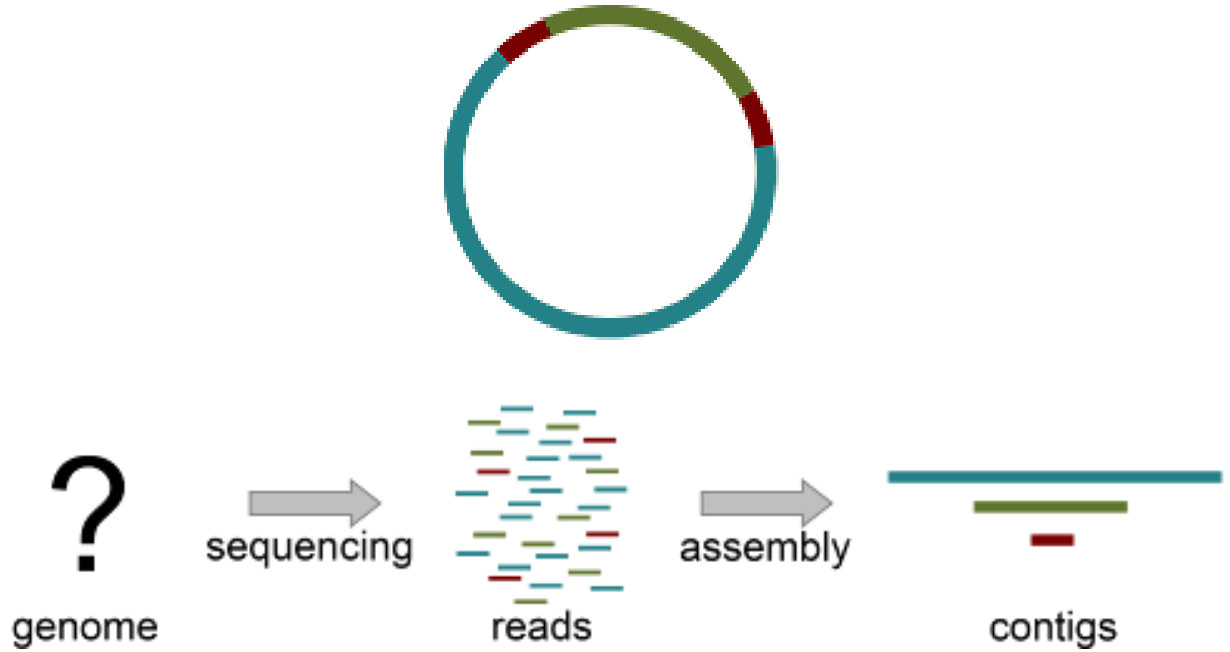
--nano-raw Nanopore regular reads (<20% error)

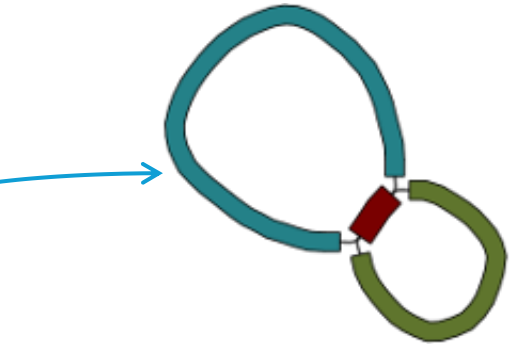
--nano-hq Nanopore high-quality reads: SUP or **Q20** (<3% error)

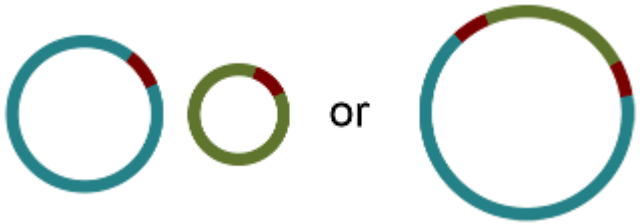
Similar options are designed for Pacbio:
(https://github.com/mikolmogorov/Flye/blob/flye/docs/USAGE.md#inputdata)

Assembly QC

- `assembly_info.txt` :
  - Number of Contigs
  - Total length of assembly
- Visualize GFA in Bandage





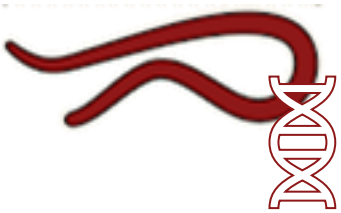


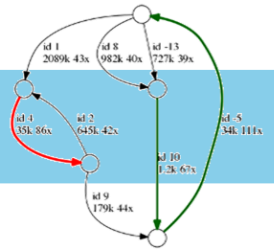
Any questions?

Please contact wellcomeconnectingscience.org for more information.

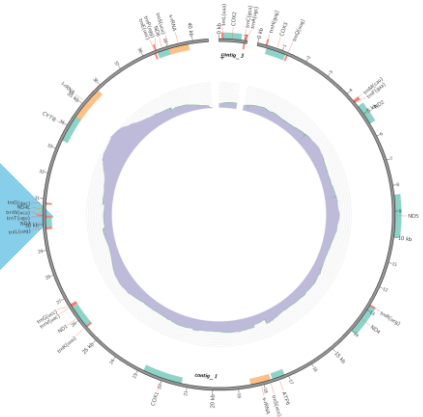
May 2025

Pipeline





MITO



DNA extraction

ONT sequencing

Basecalling & Data QC

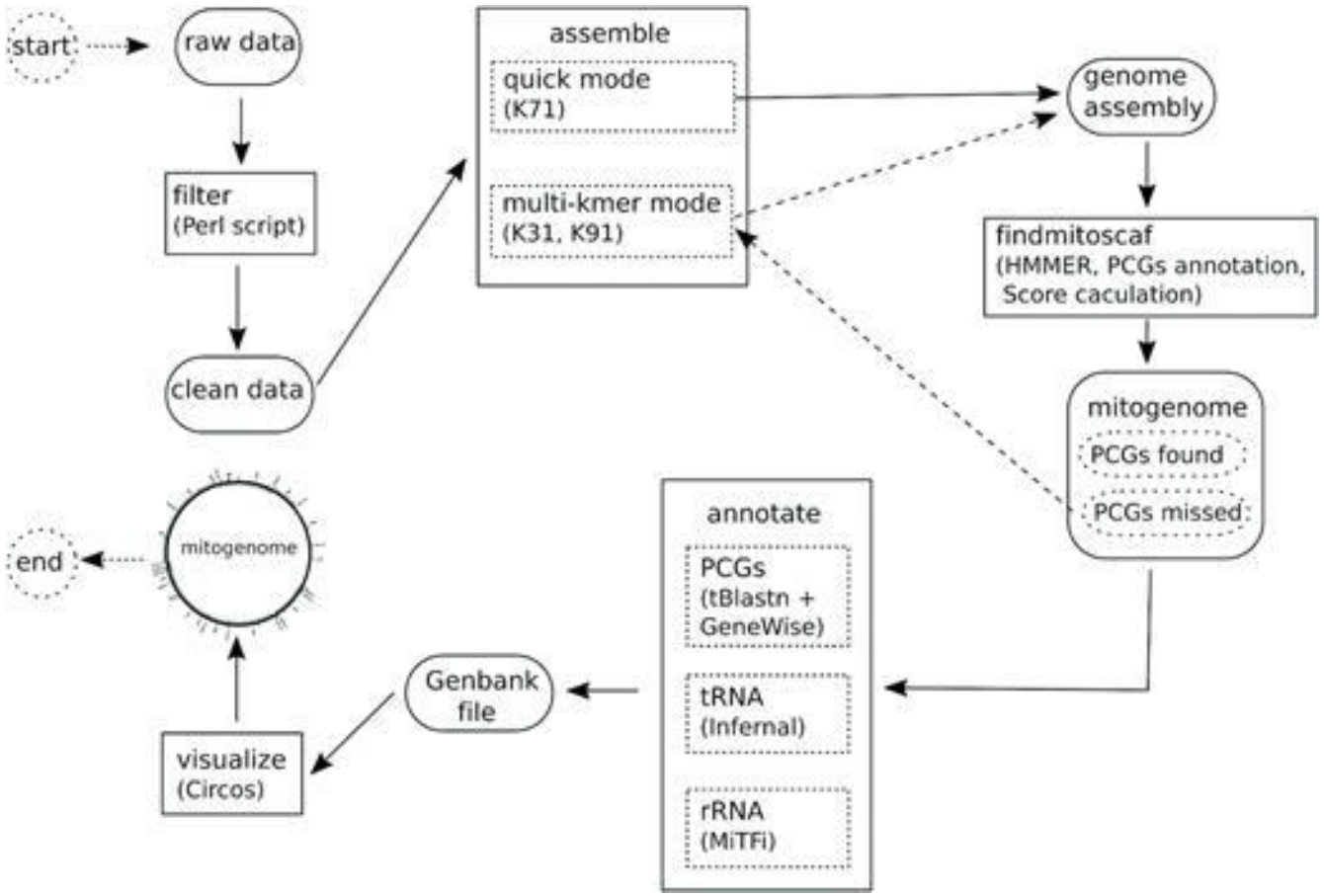
Diamond mtDNA reads selection

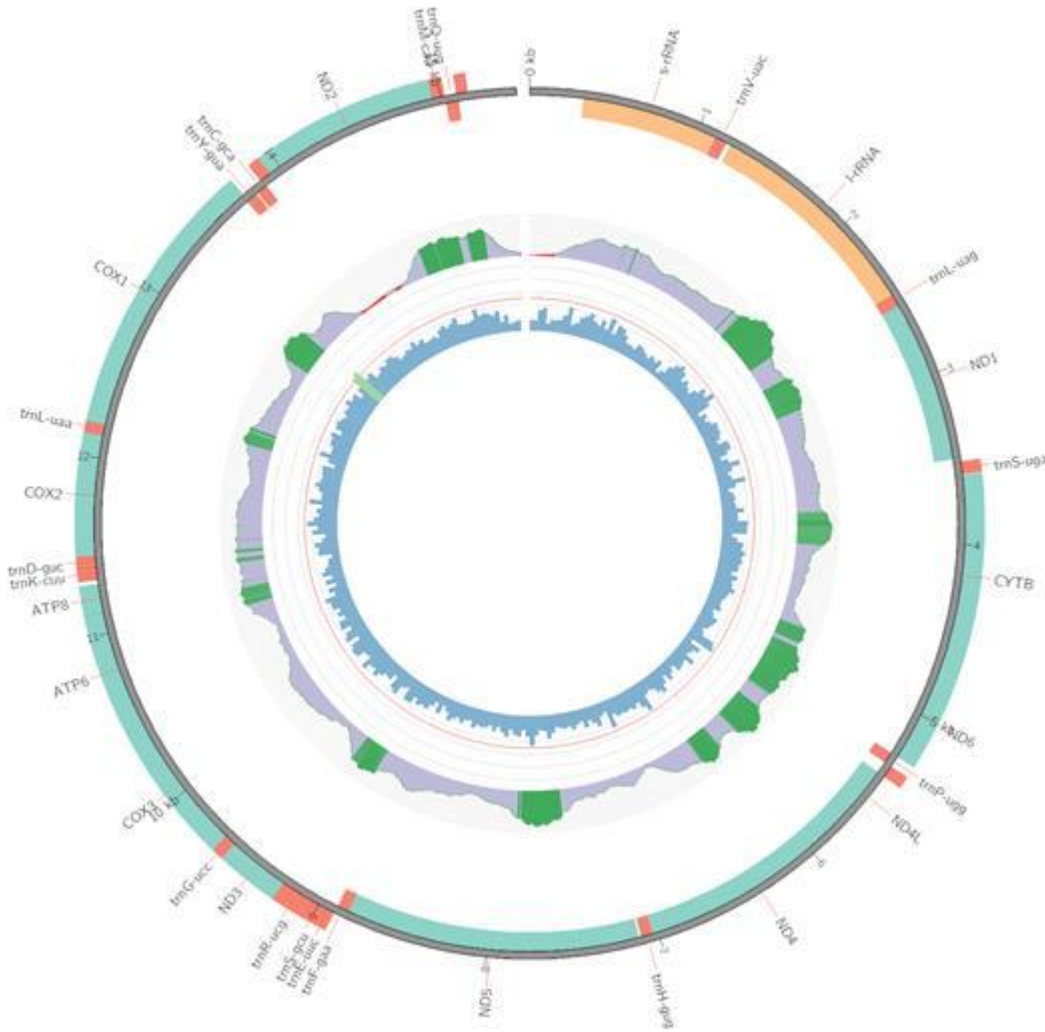
Flye assembly

MitoZ annotation

☑ ☑ ☑ ☑ ☑

MITOZ


start
raw data
filter
(Perl script)
clean data
assemble
quick mode
(K71)
multi-kmer mode
(K31, K91)
genome
assembly
findmitoscaf
(HMMER, PCGs annotation,
Score caculation)
mitogenome
PCGs found
PCGs missed
annotate
PCGs
(tBlastn +
GeneWise)
tRNA
(Infernal)
rRNA
(MiTFi)
Genbank
file
visualize
(Circos)
mitogenome
end


ND2
COX1
COX2
ATP8
ATP6
COX3
ND3
ND5
ND4
ND4L
ND6
CYTB
ND1
s-rRNA
l-rRNA

MITOZ

```
mitoz annotate
--workdir ./
--fastafiles cox1.fasta
--clade Nematoda
--genetic_code 5
--outprefix mitoz_drenale
```

Program
Working directory name
Fasta file with mitogenome assembly
Clade
Clade specific genetic code
Name **Output folder**

wellcome connecting science

thanks!

Please contact wellcomeconnectingscience.org for more information.

May 2025

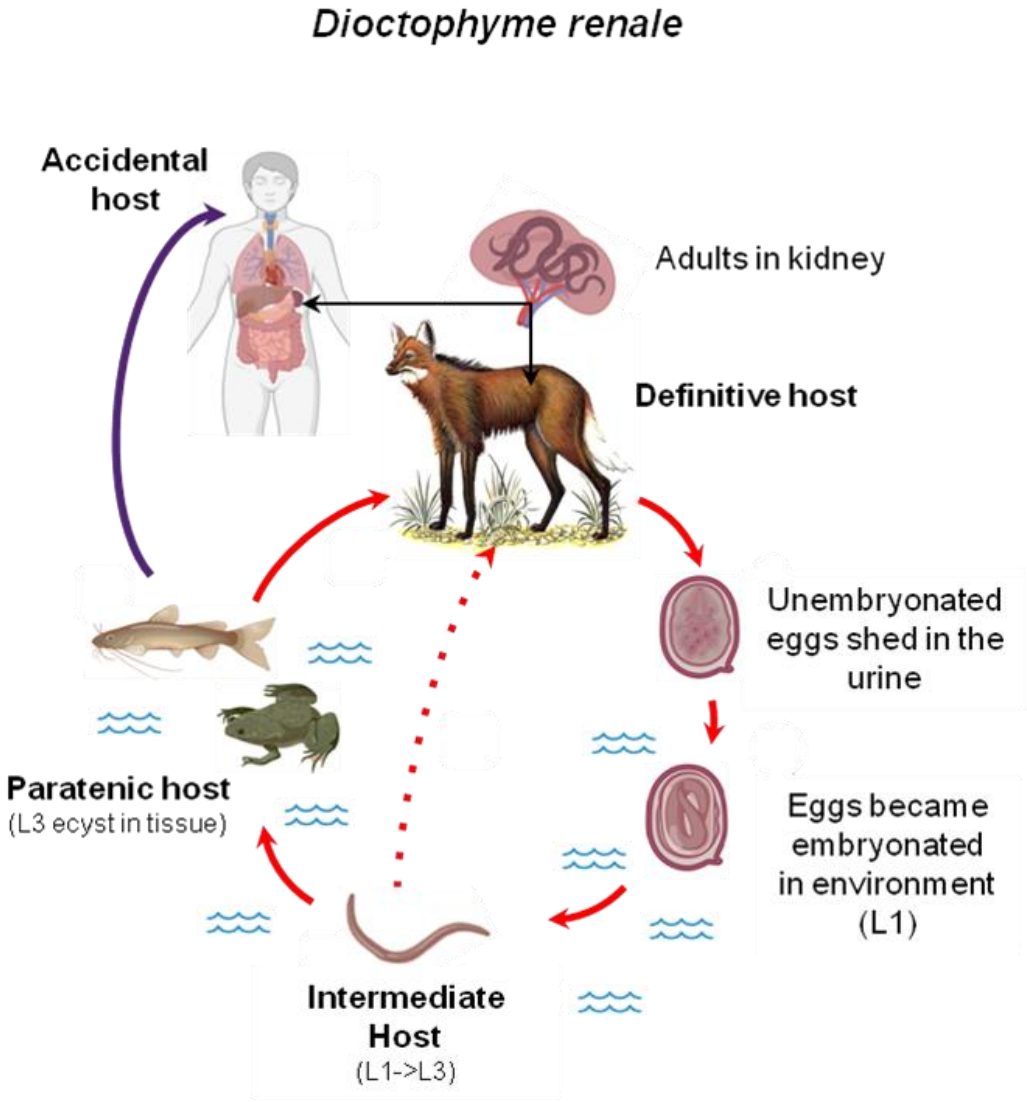
Dioctophyme renale

Caracterización mitogenómica

Dioctophyme renale

"Giant kidney worm"

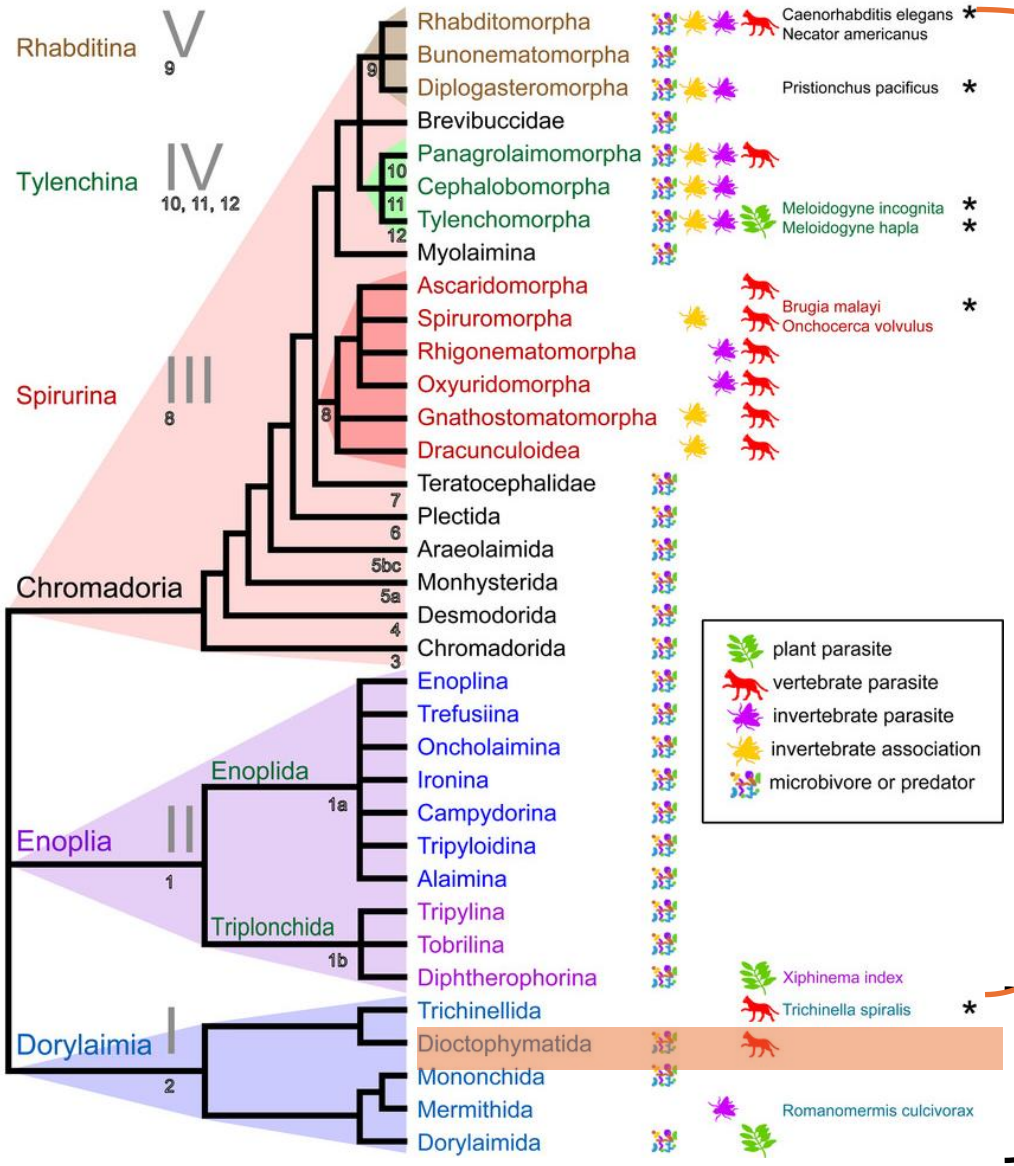
- **Nematode parasite**
- Considered the **largest parasitic** nematode of terrestrial vertebrates
- Health problem in **dogs and threatened wildlife** living near aquatic environments
- High risk of causing **infections in human** populations in riparian areas
- **Hard to sample**. surgeries or roadkill wildlife (*degraded DNA*)
- Little molecular information on this organism (no **genome** nor **transcriptome**)



Phylogenetic Relationships - Nematodes

mtDNA has valuable characteristics :

- high mutation and substitution rates
- infrequent genetic recombination
- maternal transmission
- high copy number (!)
- easy accessibility.

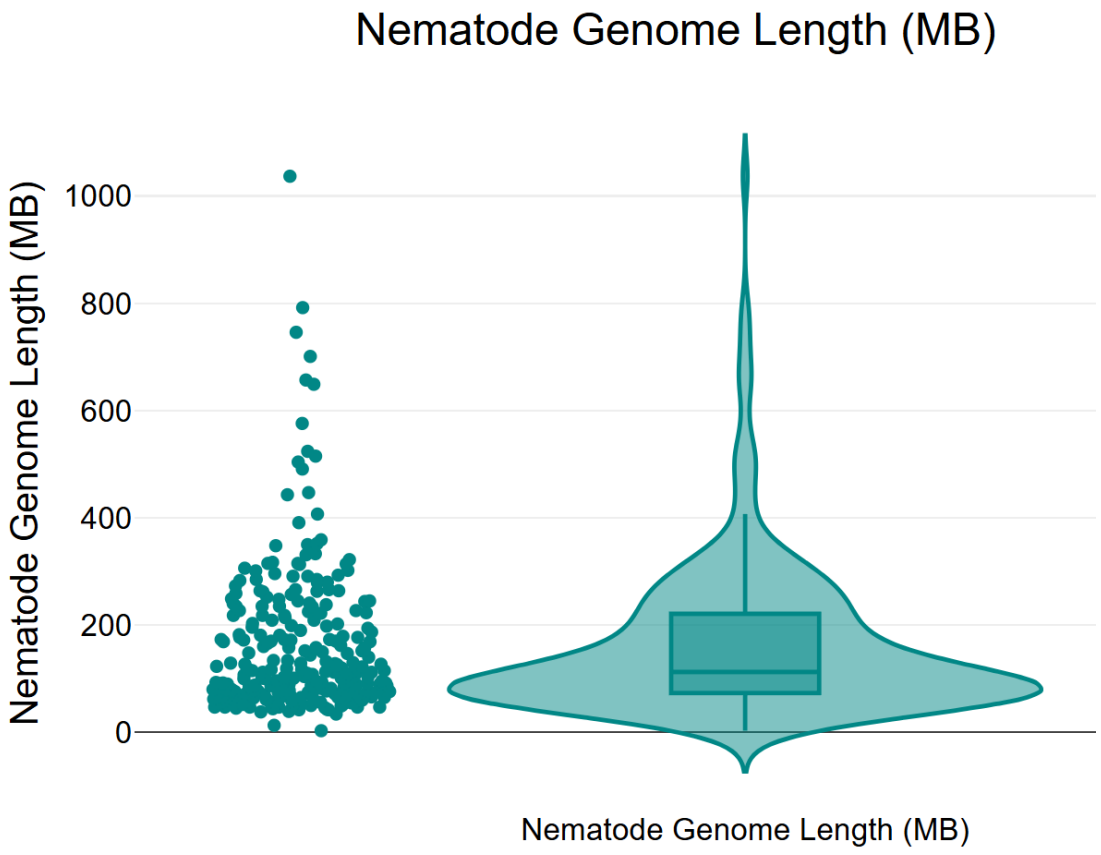


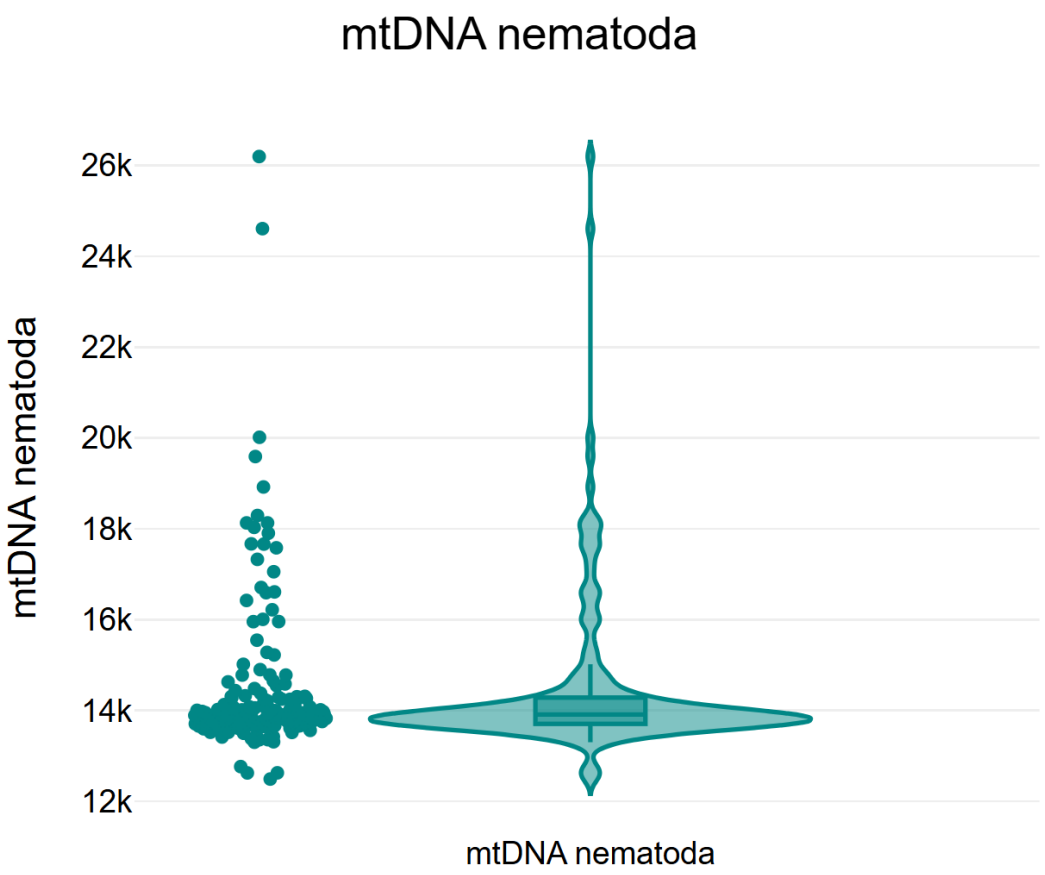
C. elegans

Trichinella sp.
Trichuris sp.

Sub-represented genomic data

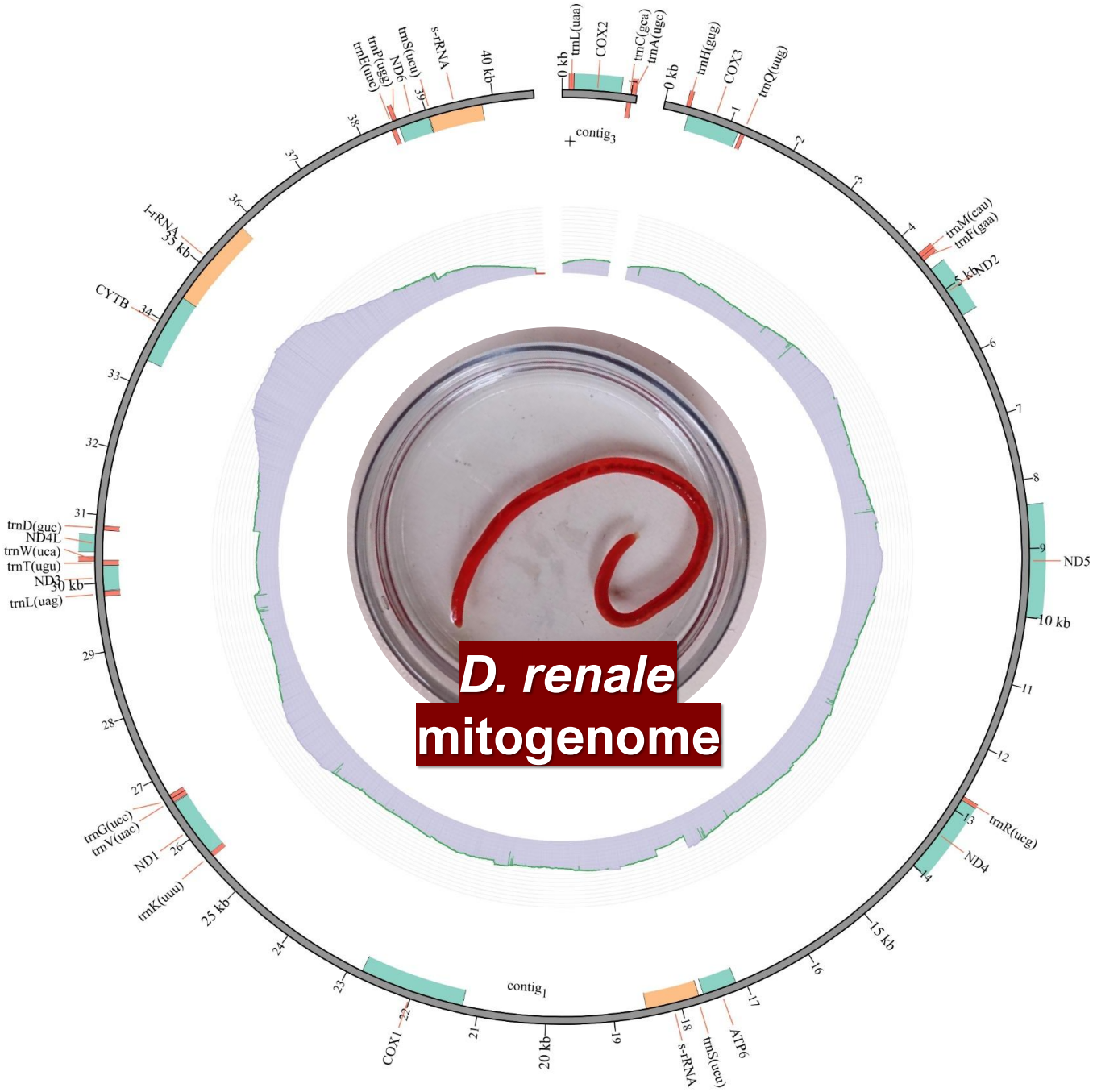
(Mito)Genomes lengths in Nematodes

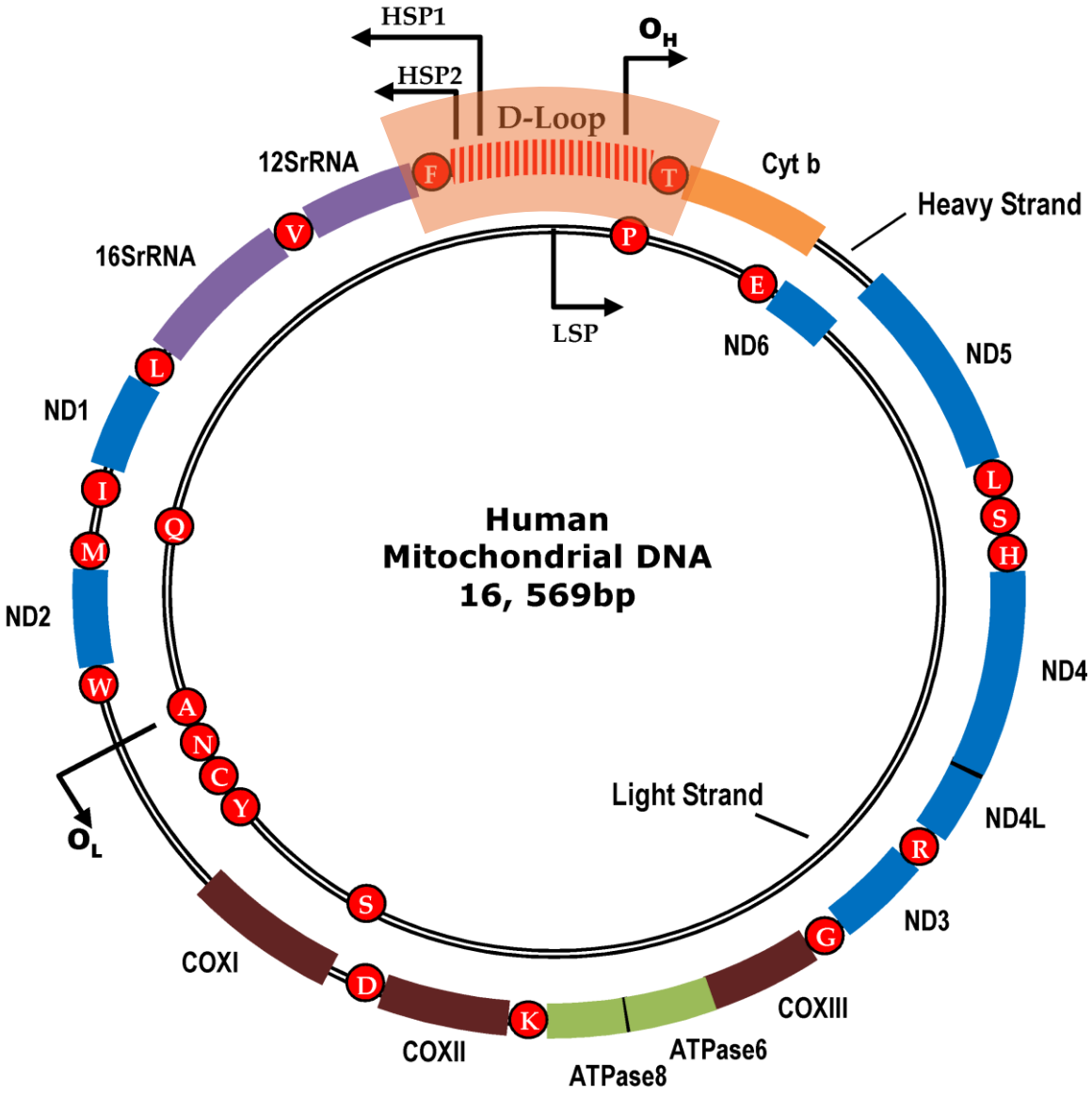


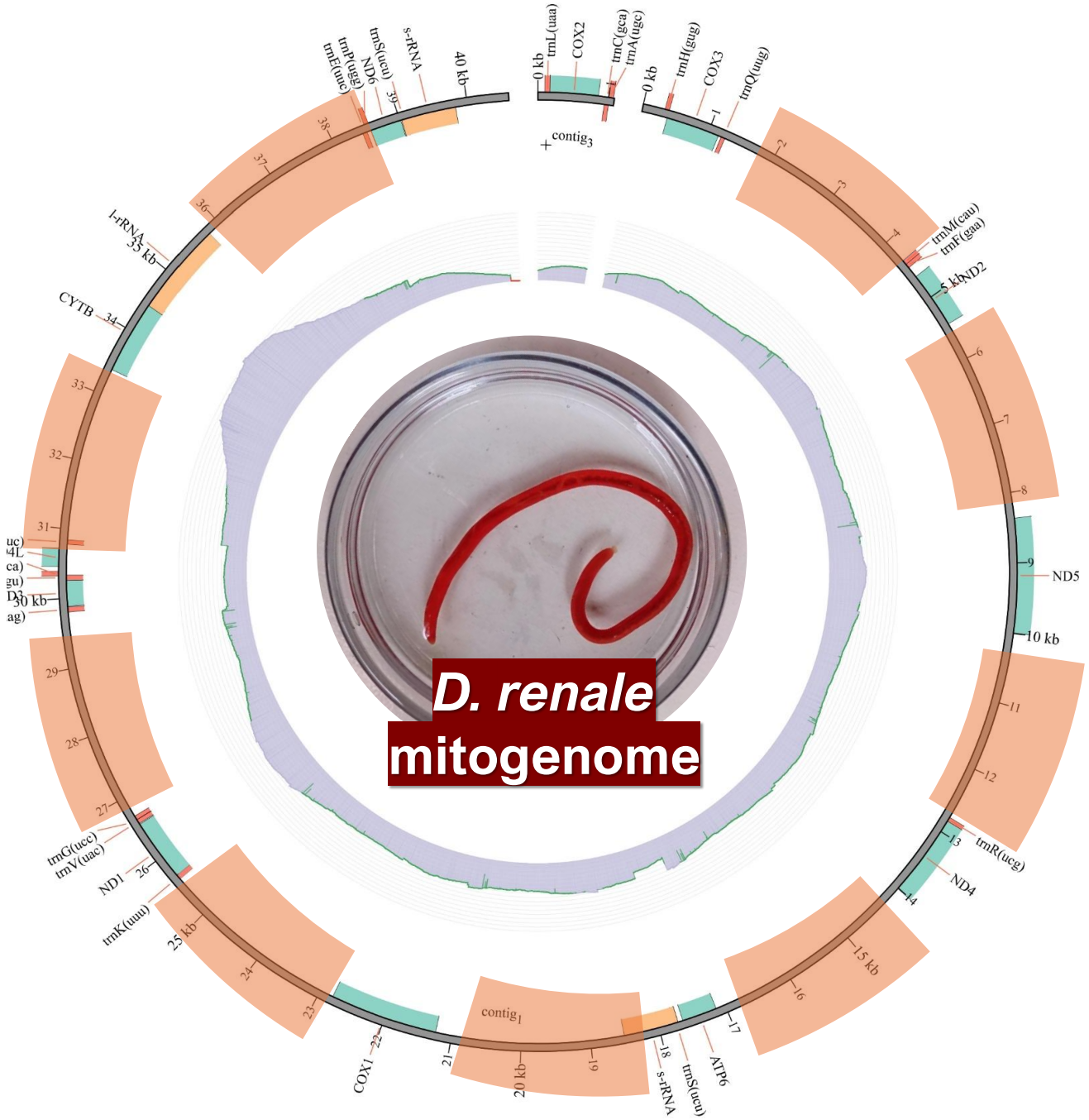


ATP6
COX1
COX2
COX3
CYTB
NAD1
NAD2
NAD3
NAD4
NAD4L
NAD5
NAD6

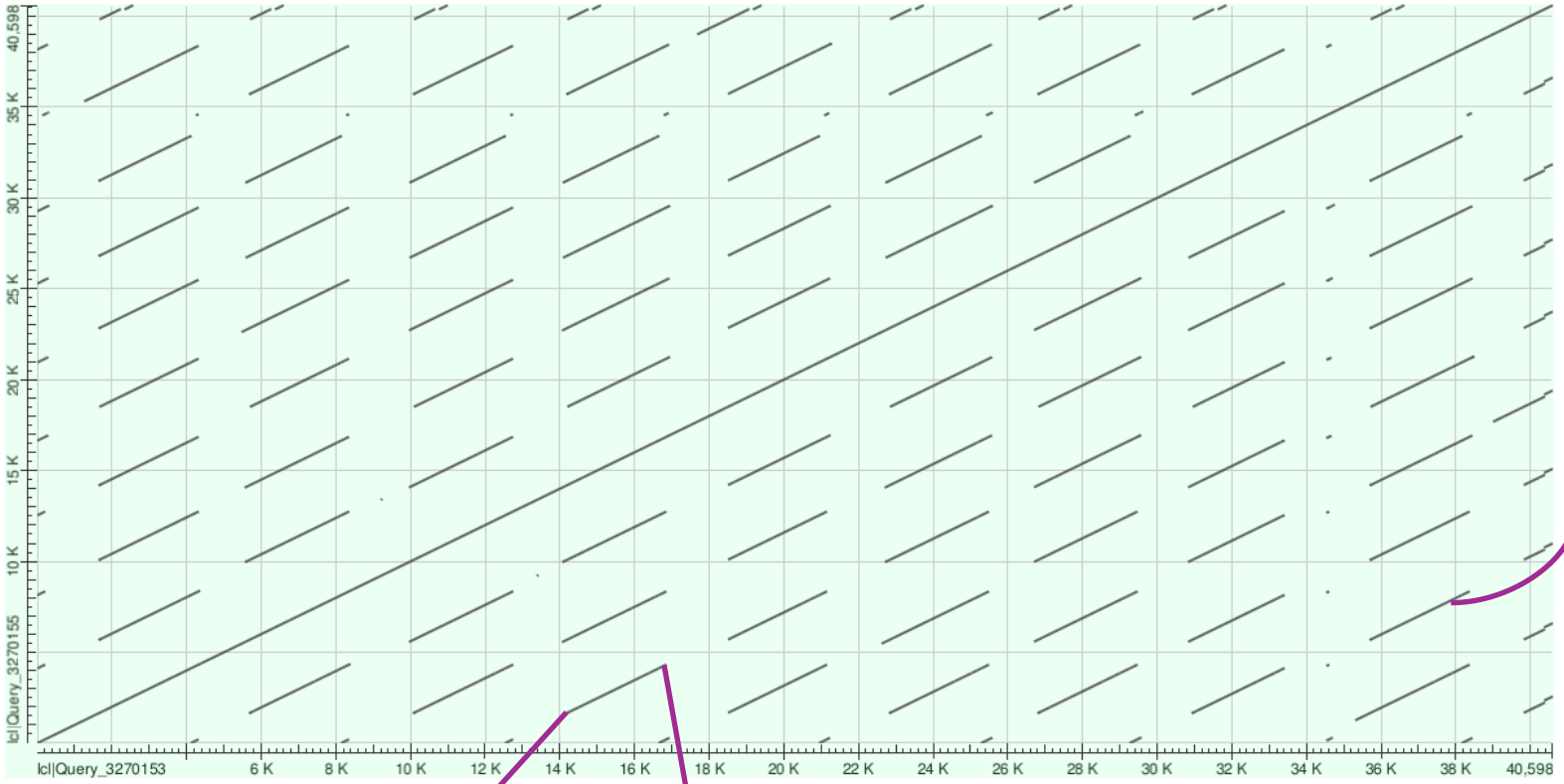
rRNA-L
rRNA-S
22 x tRNA




HSP1
HSP2
D-Loop
O_H
12SrRNA
Cyt b
Heavy Strand
16SrRNA
LSP
ND6
ND5
ND1
ND2
ND4
ND4L
ND3
COXI
COXII
COXIII
ATPase6
ATPase8
O_L
Light Strand
Human
Mitochondrial DNA
16, 569bp

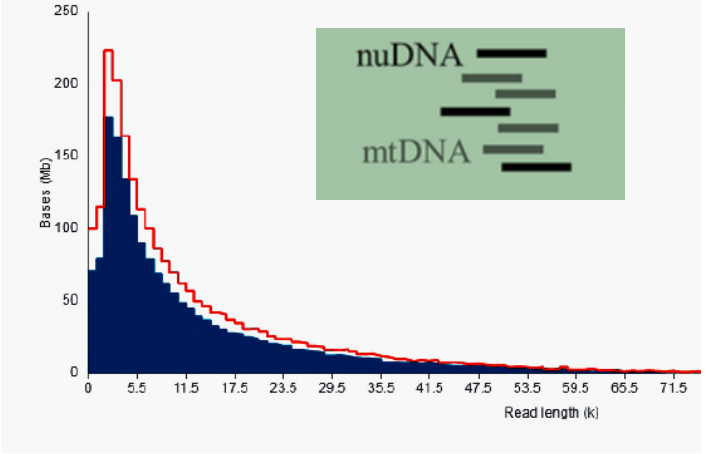

D. renale
mitogenome
contig3
contig1
COX2
COX3
ND2
ND5
ND4
ND1
COX1
ATP6
CYTB
ND6
s-rRNA
l-rRNA
ND3
ND4L

Non-Coding regions (NCR)
Each CDS is flanked by an NCR



The assembly is complicated by very similar repeated NCRs

Conclusions



- ❖ Genome skimming bioinformatics allowed us to **assemble and annotate** a new mitochondrial genome
- ✓ **Without a reference genome**
- ✓ **No mitochondrial separation required**
- ✓ **No PCR** amplification needed
- ✓ All mitochondrial proteins were annotated with **82-99% AA identity** to other phylogenetically related nematodes
- ✓ All mitochondrial **rRNA** and **tRNA** were found
- ✓ Two mitochondrial **molecular markers** were designed and implemented in phylogeography studies
- ✓ Unique nematode mitochondrial **features** were identified

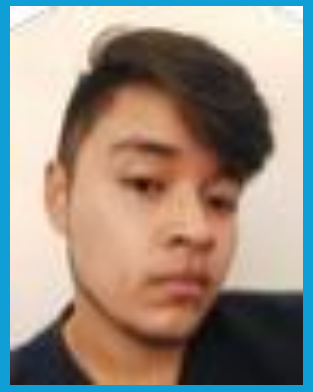
Genomics and Bioinformatics LAB10. iB3-UBA-CONICET

Natalia Macchiaroli
Researcher

Lucas Arce
Fellowship

Prof. Gisela Franchini
INBIOLP-UNLP-CONICET

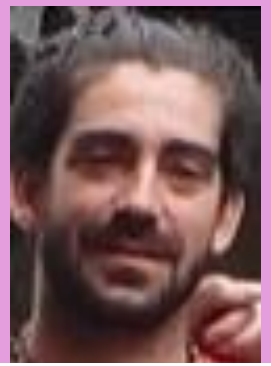
Prof. Laura Kamenetzky
Principal Investigator



Kevin Calupiña
Fellowship

Ines Sananez
Fellowship

Marina Ingravidi Fellowship



Juan Arrabal
Fellowship

Melisa Magallanes
Fellowship

Prof. Enrique Caviedes-Vidal
Principal Investigator

iB3 Instituto de Biociencias, Biotecnología y Biología traslacional

INIBIOLP
Instituto de Investigaciones Bioquímicas de La Plata
"Profesor Doctor Rodolfo R. Brenner"

CONICET UNSL
IMIBIO-SL