Examining the *Acanthobothrium tortum* genome

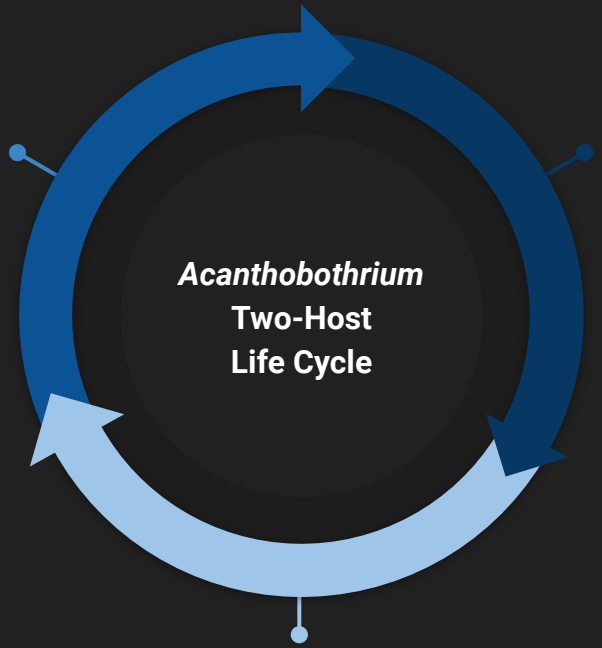
Khalia Cain & Kara Heilemann

Tapeworm Background

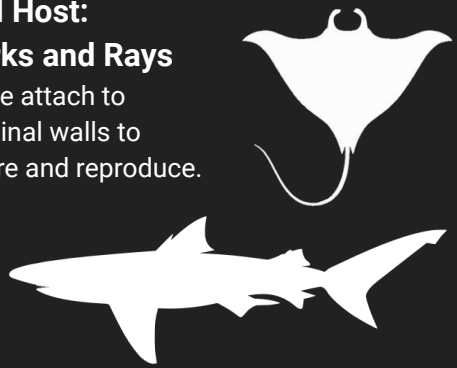
Tapeworms infect various organisms worldwide

Multi-host life cycle through the **food chain**

Intermediate Host:
Eggs develop into infective larvae



Final Host:
Sharks and Rays
Larvae attach to intestinal walls to mature and reproduce.



Environment:
Eggs pass from intestines to environment

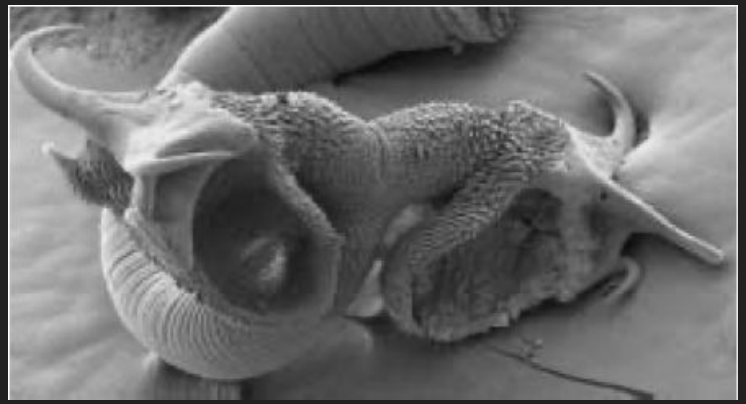
Tapeworm Background

Tapeworms have incredible morphological diversity!

Estimated **20,000 species** (¼ currently described!)

Acanthobothrium: genus comprised of 201 valid species parasitizing fresh and saltwater sharks and rays

Acanthobothrium, Fyler 2011



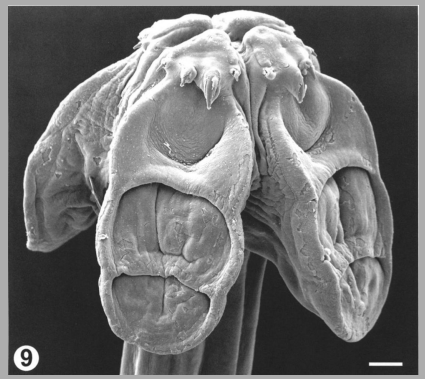
Yorkeria. Caira and Jensen

Paragrillotia, Caira and Jensen

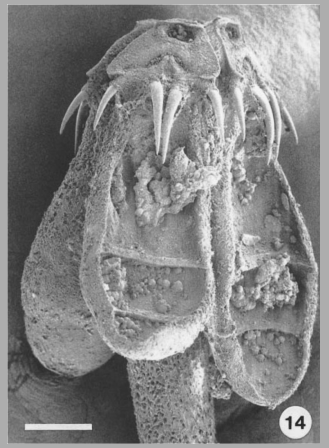
Echinobothrium, Caira

Motivation - *Acanthobothrium*

What is the evolutionary history of tapeworm freshwater and marine lineages?



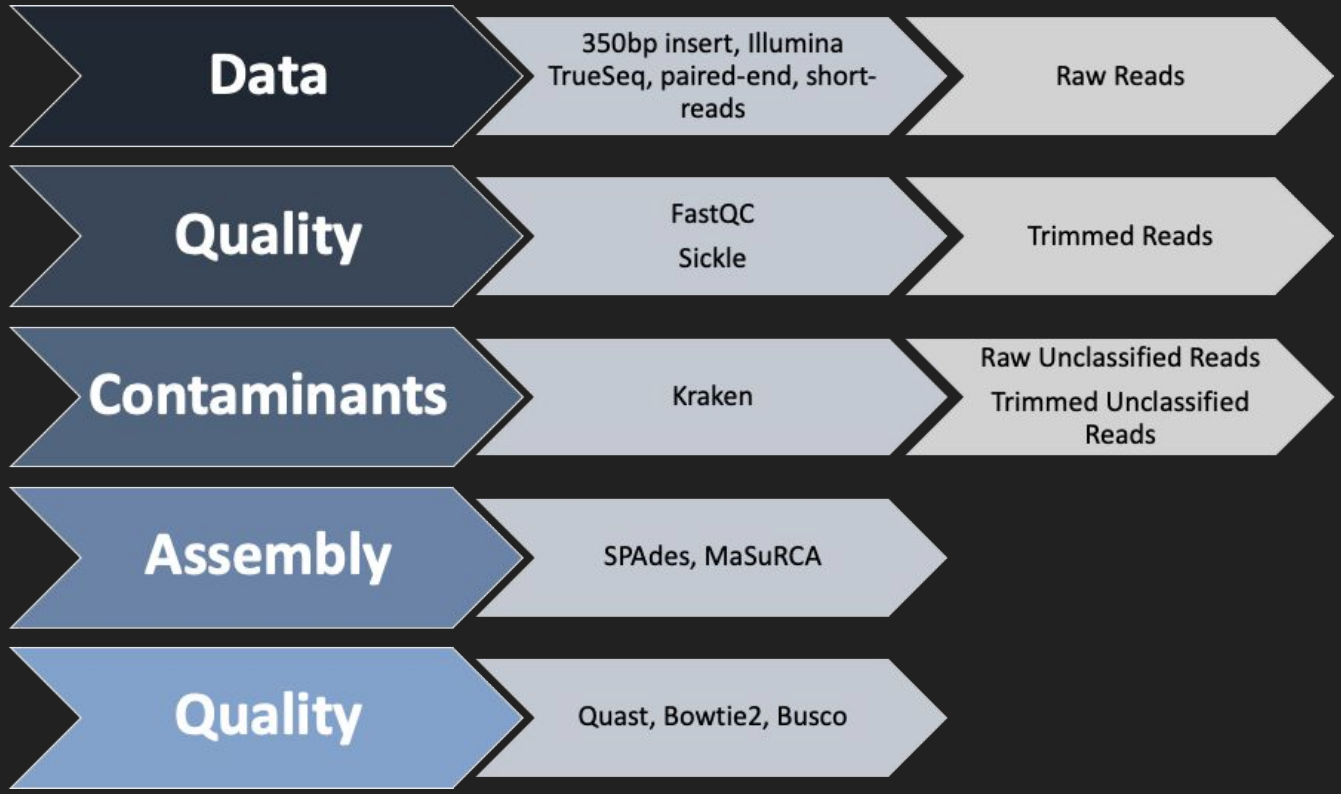
Freshwater



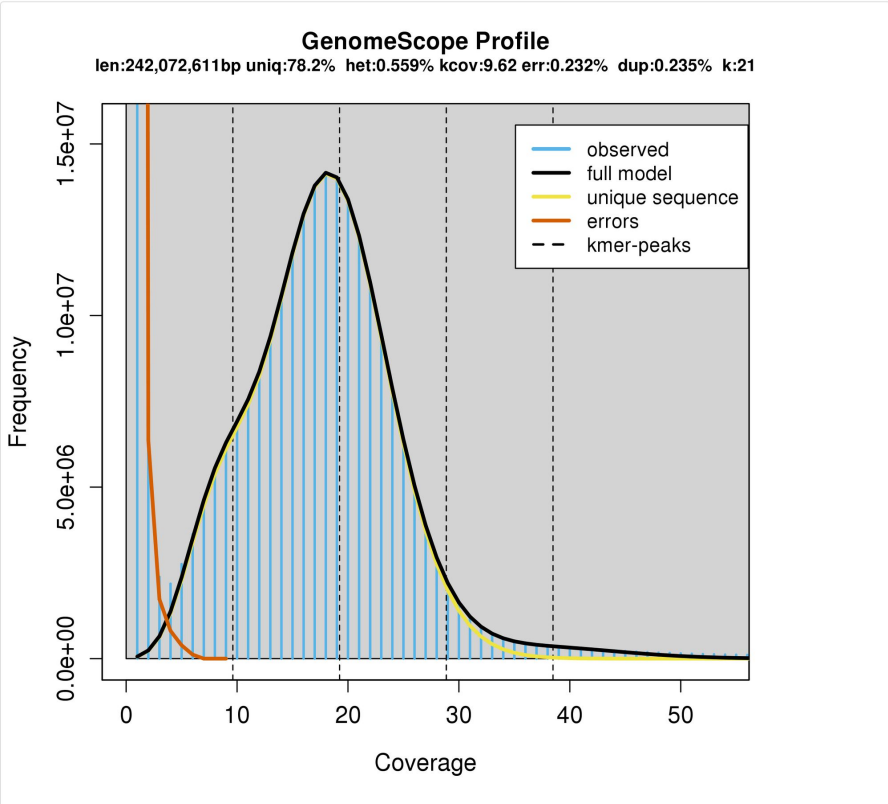
Marine

Research Objectives:

Exploration and genome assembly of *Acanthobothrium tortum*



Genome Size Estimate: Jellyfish and GenomeScope



Genome Length Estimate:
242,072,611 bp

Results

```
GenomeScope version 1.0
k = 21

property                min                 max
Heterozygosity          0.55575%            0.561302%
Genome Haploid Length   241,887,253 bp      242,072,611 bp
Genome Repeat Length    52,807,787 bp       52,848,253 bp
Genome Unique Length    189,079,466 bp      189,224,358 bp
Model Fit               97.0743%            99.5966%
Read Error Rate         0.231986%           0.231986%
```

FastQC: Basic statistics

Raw Seq.

Basic Statistics

Measure	Value
Filename	BE8G1_R2.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	37507337
Sequences flagged as poor quality	0
Sequence length	101
%GC	46

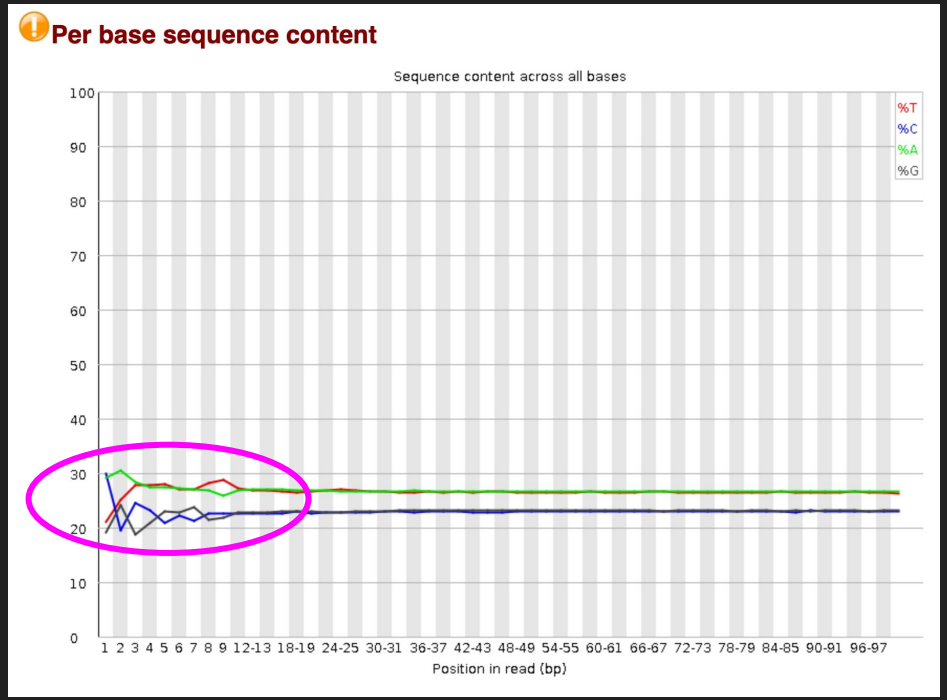
Sickle trimmed Seq.

Basic Statistics

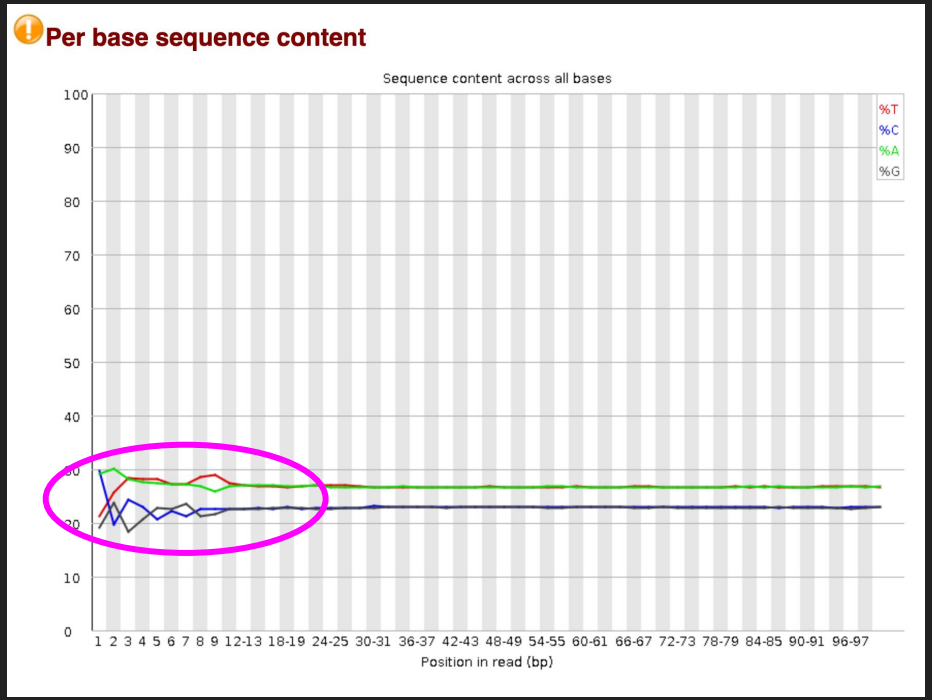
Measure	Value
Filename	trim_BE8G1_R1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	36413015
Sequences flagged as poor quality	0
Sequence length	20-101
%GC	46

FastQC: Per base sequence content

Raw Seq.

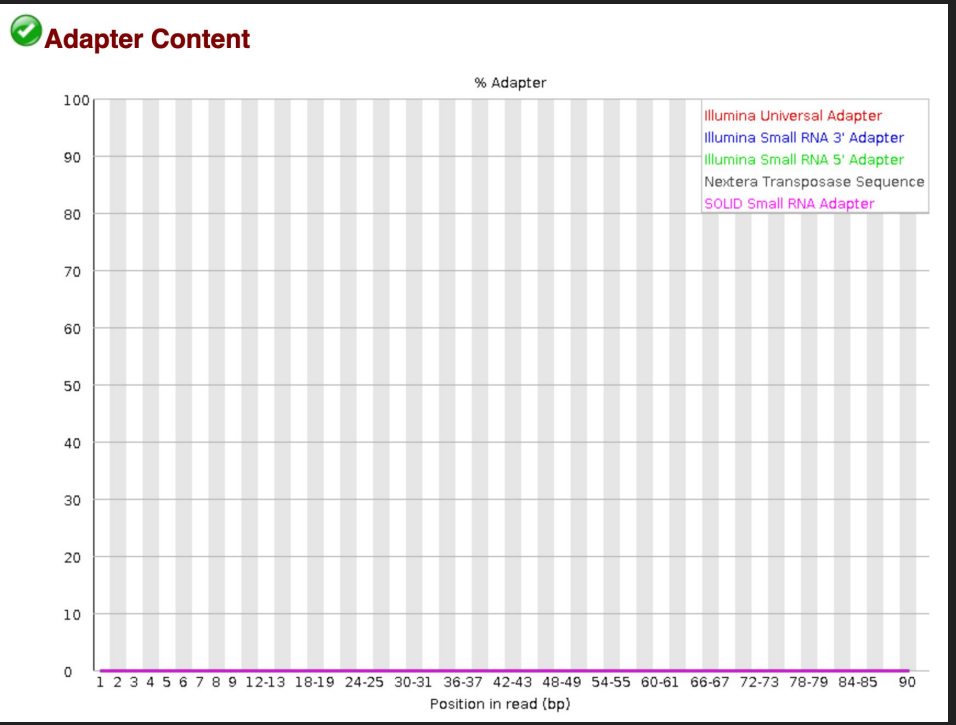


Sickle trimmed Seq.

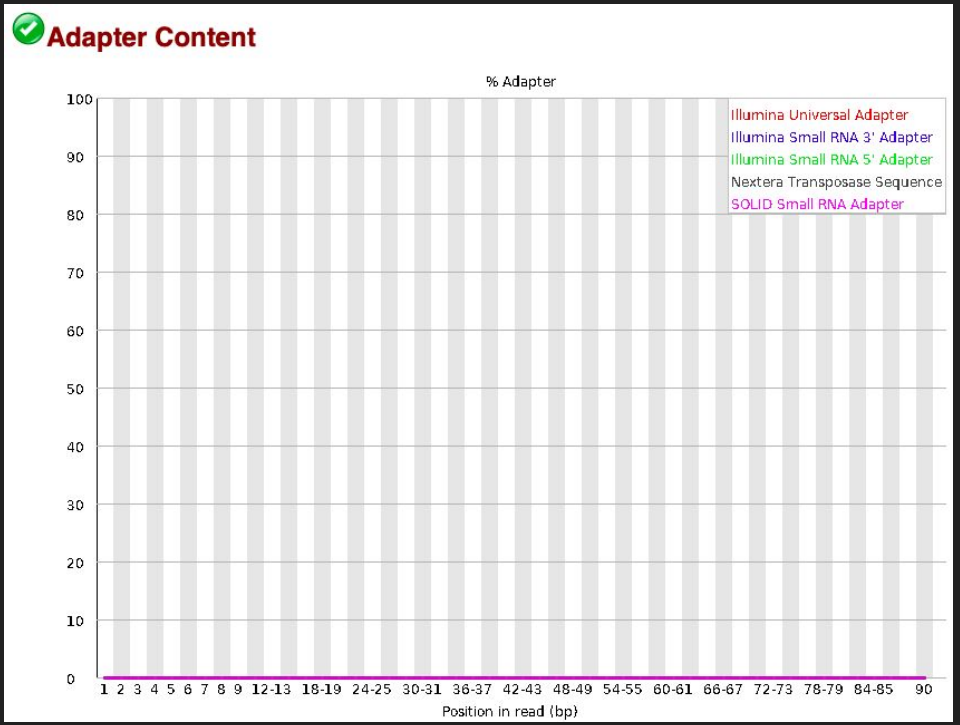


FastQC: Adapter Content

Raw Seq.

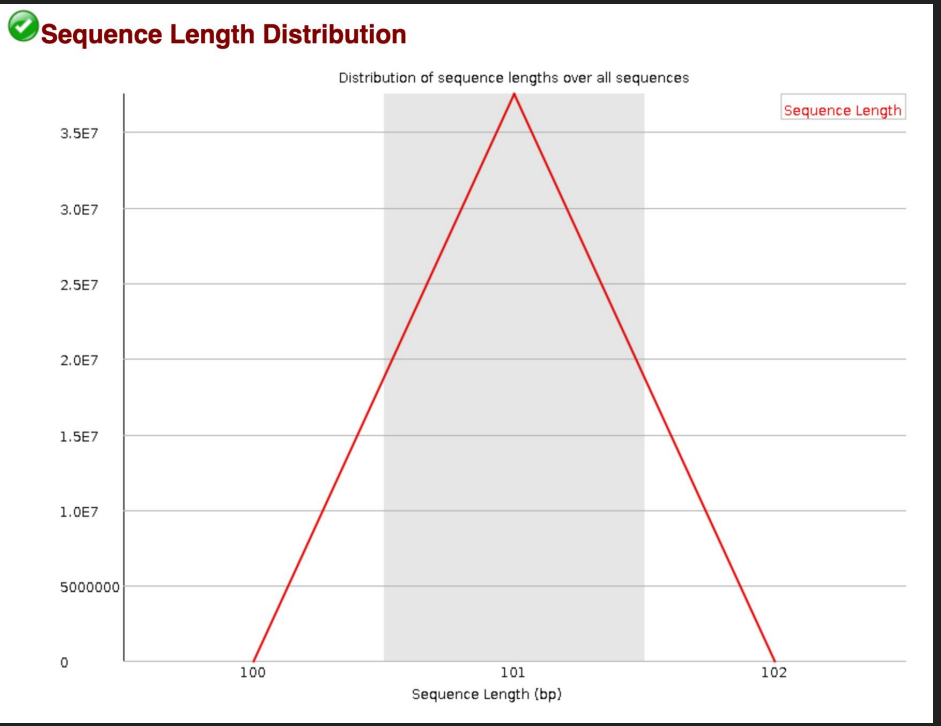


Sickle trimmed Seq.

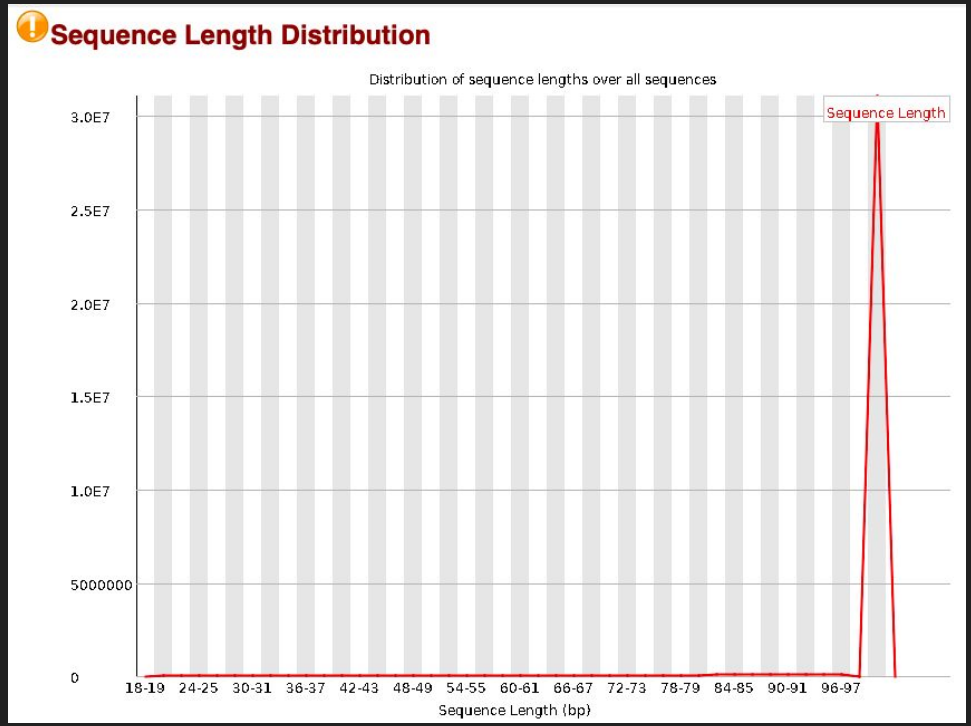


FastQC: Sequence length distribution

Raw Seq.



Sickle trimmed Seq.

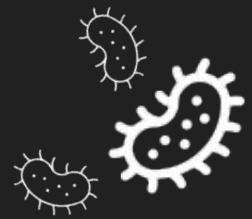


Contaminant Screening: Kraken

Contaminant	Raw	Trimmed
Bacteria	690865	647092
Eukaryota	595949	394742
Viruses	10630	9920
Archaea	9751	8949

Statistic	Raw	Trimmed
Classified	3.82%	3.25%
Unclassified	96.18%	96.75%

Greatest Contaminants: Human, Burkholderiaceae, *Ralstonia solanacearum*, *Staphylococcus, Photobacterium*



Quality Checks Conclusion

Good overall QC stats raw and trimmed

- Didn’t get better/worse with *Sickle* trimming

de novo assemblies on raw, raw unclassified and trimmed and trimmed, unclassified reads

- *SPAdes 3.15.0*
- *MaSuRCA 4.0.1*

Trimmed, Unclassified SPAdes Assembly: Quality and Alignment

Quast

Assembly	scaffolds
# contigs (>= 0 bp)	1560905
# contigs (>= 1000 bp)	63180
# contigs (>= 5000 bp)	4556
# contigs (>= 10000 bp)	784
# contigs (>= 25000 bp)	17
# contigs (>= 50000 bp)	0
Total length (>= 0 bp)	347175831
Total length (>= 1000 bp)	144598218
Total length (>= 5000 bp)	35771431
Total length (>= 10000 bp)	10615747
Total length (>= 25000 bp)	519003
Total length (>= 50000 bp)	0
# contigs	129688
Largest contig	47938
Total length	192506970
GC (%)	46.24
N50	1795
N75	1002
L50	26340
L75	62963
# N's per 100 kbp	440.04

Bowtie2

```
35231013 reads; of these:
  35231013 (100.00%) were paired; of these:
    17780032 (50.47%) aligned concordantly 0 times
    17190167 (48.79%) aligned concordantly exactly 1 time
    260814 (0.74%) aligned concordantly >1 times
    ----
    17780032 pairs aligned concordantly 0 times; of these:
      5764456 (32.42%) aligned discordantly 1 time
    ----
    12015576 pairs aligned 0 times concordantly or discordantly; of these:
      24031152 mates make up the pairs; of these:
        7343006 (30.56%) aligned 0 times
        7861580 (32.71%) aligned exactly 1 time
        8826566 (36.73%) aligned >1 times
89.58% overall alignment rate
```

Trimmed, Unclassified SPAdes Assembly: Completeness

BUSCO

```
C:35.3%[S:34.7%,D:0.6%],F:24.7%,M:40.0%,n:954
337     Complete BUSCOs (C)
331     Complete and single-copy BUSCOs (S)
6       Complete and duplicated BUSCOs (D)
236     Fragmented BUSCOs (F)
381     Missing BUSCOs (M)
954     Total BUSCO groups searched
```

Trimmed Unclassified MaSuRCA Assembly: Quality and Alignment

Quast

Assembly	final.genome.scf
# contigs (>= 0 bp)	114521
# contigs (>= 1000 bp)	66857
# contigs (>= 5000 bp)	9398
# contigs (>= 10000 bp)	1034
# contigs (>= 25000 bp)	3
# contigs (>= 50000 bp)	0
Total length (>= 0 bp)	228223418
Total length (>= 1000 bp)	200786011
Total length (>= 5000 bp)	68246039
Total length (>= 10000 bp)	12885610
Total length (>= 25000 bp)	83875
Total length (>= 50000 bp)	0
# contigs	93691
Largest contig	31981
Total length	220216713
GC (%)	46.25
N50	3402
N75	1881
L50	19626
L75	41308
# N's per 100 kbp	0.00

Bowtie

```
75014674 reads; of these:
  75014674 (100.00%) were unpaired; of these:
    8841063 (11.79%) aligned 0 times
    47280714 (63.03%) aligned exactly 1 time
    18892897 (25.19%) aligned >1 times
88.21% overall alignment rate
```

Trimmed Unclassified MaSuRCA Assembly: Completeness

BUSCO

```
---------------------------------------------------
|Results from dataset metazoa_odb10               |
---------------------------------------------------
|C:45.8%[S:45.5%,D:0.3%],F:20.4%,M:33.8%,n:954    |
|437    Complete BUSCOs (C)                       |
|434    Complete and single-copy BUSCOs (S)       |
|3      Complete and duplicated BUSCOs (D)        |
|195    Fragmented BUSCOs (F)                     |
|322    Missing BUSCOs (M)                        |
|954    Total BUSCO groups searched               |
---------------------------------------------------
```

Trimmed MaSuRCA Assembly: Quality and Alignment

Quast

Assembly	final.genome.scf
# contigs (>= 0 bp)	103925
# contigs (>= 1000 bp)	61608
# contigs (>= 5000 bp)	11564
# contigs (>= 10000 bp)	1922
# contigs (>= 25000 bp)	24
# contigs (>= 50000 bp)	1
Total length (>= 0 bp)	232921872
Total length (>= 1000 bp)	208969021
Total length (>= 5000 bp)	90399398
Total length (>= 10000 bp)	25234411
Total length (>= 25000 bp)	721985
Total length (>= 50000 bp)	55186
# contigs	84459
Largest contig	55186
Total length	225456149
GC (%)	46.26
N50	4073
N75	2184
L50	16521
L75	35320
# N's per 100 kbp	0.00

Bowtie2

```
75014674 reads; of these:
  75014674 (100.00%) were unpaired; of these:
    8097210 (10.79%) aligned 0 times
    47790174 (63.71%) aligned exactly 1 time
    19127290 (25.50%) aligned >1 times
89.21% overall alignment rate
```

Trimmed MaSuRCA Assembly: Completeness

BUSCO

```
---------------------------------------------------
|Results from dataset metazoa_odb10               |
---------------------------------------------------
|C:49.8%[S:49.4%,D:0.4%],F:17.8%,M:32.4%,n:954    |
|475    Complete BUSCOs (C)                       |
|471    Complete and single-copy BUSCOs (S)       |
|4      Complete and duplicated BUSCOs (D)        |
|170    Fragmented BUSCOs (F)                     |
|309    Missing BUSCOs (M)                        |
|954    Total BUSCO groups searched               |
---------------------------------------------------
```

Raw, Unclassified MaSuRCA Assembly: Quality and Alignment

Quast

Assembly	final.genome.scf
# contigs (>= 0 bp)	106793
# contigs (>= 1000 bp)	63506
# contigs (>= 5000 bp)	10946
# contigs (>= 10000 bp)	1560
# contigs (>= 25000 bp)	6
# contigs (>= 50000 bp)	0
Total length (>= 0 bp)	231836438
Total length (>= 1000 bp)	207192427
Total length (>= 5000 bp)	82792061
Total length (>= 10000 bp)	19849338
Total length (>= 25000 bp)	186414
Total length (>= 50000 bp)	0
# contigs	87162
Largest contig	35020
Total length	224308771
GC (%)	46.32
N50	3836
N75	2106
L50	17669
L75	37288
# N's per 100 kbp	0.00

Bowtie2

```
75014674 reads; of these:
  75014674 (100.00%) were unpaired; of these:
    8841063 (11.79%) aligned 0 times
    47280714 (63.03%) aligned exactly 1 time
    18892897 (25.19%) aligned >1 times
88.21% overall alignment rate
```

Raw, Unclassified MaSuRCA Assembly: Completeness

```
|Results from dataset metazoa_odb10                                |
|C:46.2%[S:45.8%,D:0.4%],F:19.8%,M:34.0%,n:954                     |
|441    Complete BUSCOs (C)                                        |
|437    Complete and single-copy BUSCOs (S)                        |
|4      Complete and duplicated BUSCOs (D)                         |
|189    Fragmented BUSCOs (F)                                      |
|324    Missing BUSCOs (M)                                         |
|954    Total BUSCO groups searched                                |
```

Raw Masurca Assembly: Quality and Alignment

Quast

Assembly	final.genome.scf
# contigs (>= 0 bp)	96148
# contigs (>= 1000 bp)	57280
# contigs (>= 5000 bp)	13099
# contigs (>= 10000 bp)	2750
# contigs (>= 25000 bp)	50
# contigs (>= 50000 bp)	2
Total length (>= 0 bp)	236491644
Total length (>= 1000 bp)	214798689
Total length (>= 5000 bp)	107787492
Total length (>= 10000 bp)	37182510
Total length (>= 25000 bp)	1528129
Total length (>= 50000 bp)	112142
# contigs	77538
Largest contig	56936
Total length	229374137
GC (%)	46.35
N50	4683
N75	2505
L50	14526
L75	31184
# N's per 100 kbp	0.00

Bowtie2

```
75014674 reads; of these:
  75014674 (100.00%) were unpaired; of these:
    7382102 (9.84%) aligned 0 times
    48255234 (64.33%) aligned exactly 1 time
    19377338 (25.83%) aligned >1 times
90.16% overall alignment rate
```

Raw Masurca Assembly: Completeness

BUSCO

```
---------------------------------------------------
|Results from dataset metazoa_odb10               |
---------------------------------------------------
|C:50.6%[S:50.1%,D:0.5%],F:17.3%,M:32.1%,n:954    |
|483    Complete BUSCOs (C)                       |
|478    Complete and single-copy BUSCOs (S)       |
|5      Complete and duplicated BUSCOs (D)        |
|165    Fragmented BUSCOs (F)                     |
|306    Missing BUSCOs (M)                        |
|954    Total BUSCO groups searched               |
---------------------------------------------------
```

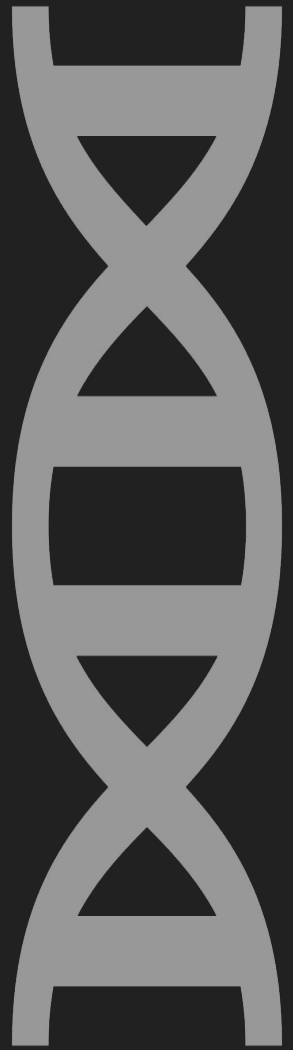
Assembly Comparison Summary

Statistics	MaSuRCA				SPAdes
	Raw	Raw, Unclassified	Trimmed	Trimmed, Unclassified	Trimmed, Unclassified
N50	4683	3836	4073	3402	1795
L50	14526	17669	16521	19626	26340
Number of Contigs	77538	87162	84459	93691	129688
Largest Contig	56936	35020	55186	31981	47938
Total Length	229,374,137	224,308,771	225,456,149	228,223,418	192,506,970
Overall Alignment Rate	90.16%	88.21%	89.21%	87.16%	89.58%
Complete BUSCOs	50.60%	46.2%	49.80%	45.80%	35.30%

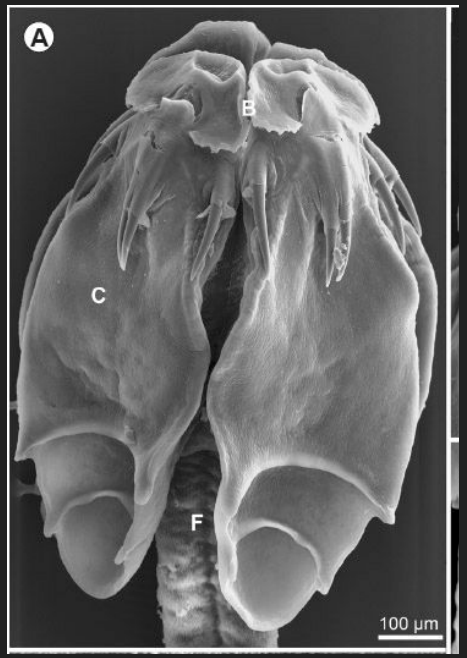
Future Work

1.) Analyze SPAdes results once completed
   - Quast, Bowtie2, BUSCO

2.) Filter out contigs <1000kb
   - awk command
   - Python removal

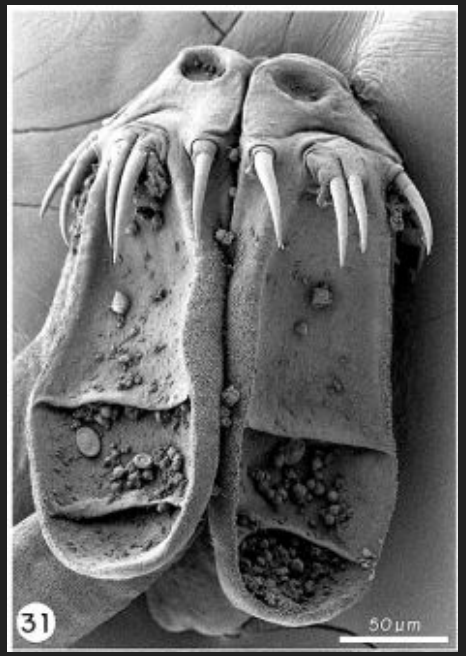


Questions?



Acanthobothrium dominage Franzese & Ivanov, 2020

Acanthobothrium carolinae Franzese & Ivanov, 2020



Acanthobothrium larsoni Reyda & Caira, 2006

Why tapeworms?

- Application to human and vet medicine
- Measure of intactness and healthiness of ecosystem
- Tapeworm life histories: better understandings of the evolution of parasitism
- “A survey of tapeworms from vertebrate bowels of the earth”