Genome Assembly and RNA-seq Annotation (partial) of the freshwater snake *Helicops angulatus*

Daniela Garcia Cobos

CG2 course final presentation

Helicops angulatus

Genome Assembly and RNA-seq Annotation

1. Genome assembly:

a) kmer analysis of raw reads: Quality check of the raw reads
b) Genome assembly using Hifiasm
c) Busco analysis to check for completeness of the genome
d) QUAST analysis to check for genome statistics

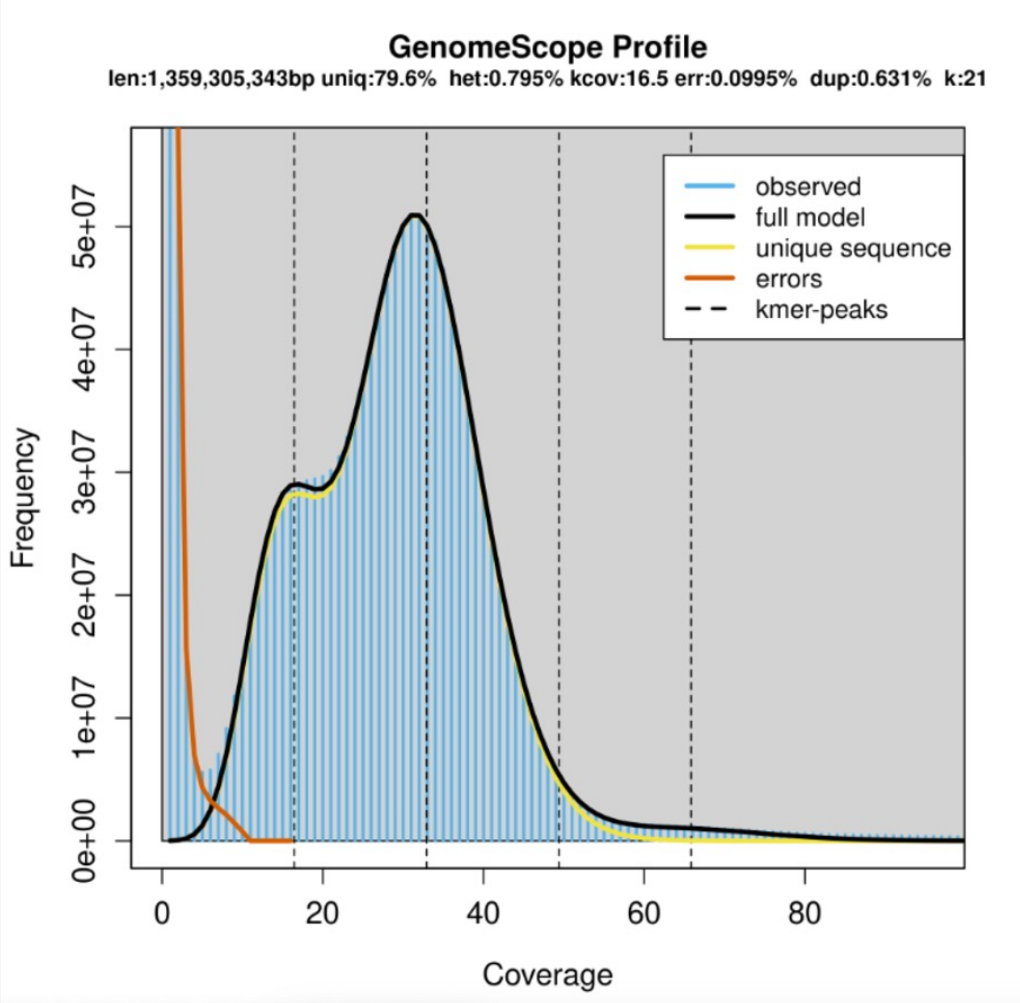
2. Genome annotation (partial- initial steps):

a) Mask genome using earlgrey
b) Assemble RNA-seq data using trinity
c) Quality check RNA-seq assembly with BUSCO scores

Genome assembly:

kmer analysis of raw reads: quality check of the raw reads

1.1 Results kmer analysis using jellyfish



- Blue bars represent the distribution of k-mer counts (frequencies) across the sequencing dataset.
- Sequencing error seems to be low, as shown by the orange line.
- Approximate genome length of 1.3 Gbps.
- Low heterozygosity
- ~ 17X Coverage

We have high-quality sequencing data with minimal errors, meaning we have good data to do a de novo genome assembly!

Genome assembly using Hifiasm

Assembly and Busco

- Hifiasm is a fast haplotype-resolved de novo assembler for PacBio HiFi reads.

- Hifiasm produces primary/alternate assemblies or partially phased assemblies only with HiFi reads.

- BUSCO estimates the completeness and redundancy of processed genomic data based on universal single-copy orthologs.

```
***** Results: *****

C:93.0%[S:91.4%,D:1.6%],F:1.3%,M:5.7%,n:7480
6958    Complete BUSCOs (C)
6837    Complete and single-copy BUSCOs (S)
121     Complete and duplicated BUSCOs (D)
98      Fragmented BUSCOs (F)
424     Missing BUSCOs (M)
76      Total BUSCO groups searched
```

- Good Busco scores!
- Completeness score could be better (ideal > 95%)
- Very low percentage of duplicates 1.6%.

Genome assembly using Hifiasm

QUAST (genome statistics) results

•QUAST stands for QUality ASsessment Tool. It evaluates genome/metagenome assemblies by computing various metrics.

Genome statistics	Results
Total size	2.1 Gbp
N50	44.57 Mbp
L50	9
Busco (completeness)	93%
Busco (duplicates)	1.6%

Report

	Helicops_angulatus_NP4.asm.bp.p_ctg
# contigs (>= 0 bp)	1074
# contigs (>= 1000 bp)	1074
# contigs (>= 5000 bp)	1073
# contigs (>= 10000 bp)	1027
# contigs (>= 25000 bp)	872
# contigs (>= 50000 bp)	651
Total length (>= 0 bp)	2145892085
Total length (>= 1000 bp)	2145892085
Total length (>= 5000 bp)	2145888057
Total length (>= 10000 bp)	2145526026
Total length (>= 25000 bp)	2142817626
Total length (>= 50000 bp)	2134986056
# contigs	1074
Largest contig	197493652
Total length	2145892085
GC (%)	41.13
N50	44570384
N75	15821402
L50	9
L75	30
# N's per 100 kbp	0.00

Genome Assembly and RNA-seq Annotation

1. Genome assembly:

a) kmer analysis of raw reads: Quality check of the raw reads
b) Genome assembly using Hifiasm
c) Busco analysis to check for completeness of the genome
d) QUAST analysis to check for genome statistics

2. Genome annotation (partial- initial steps):

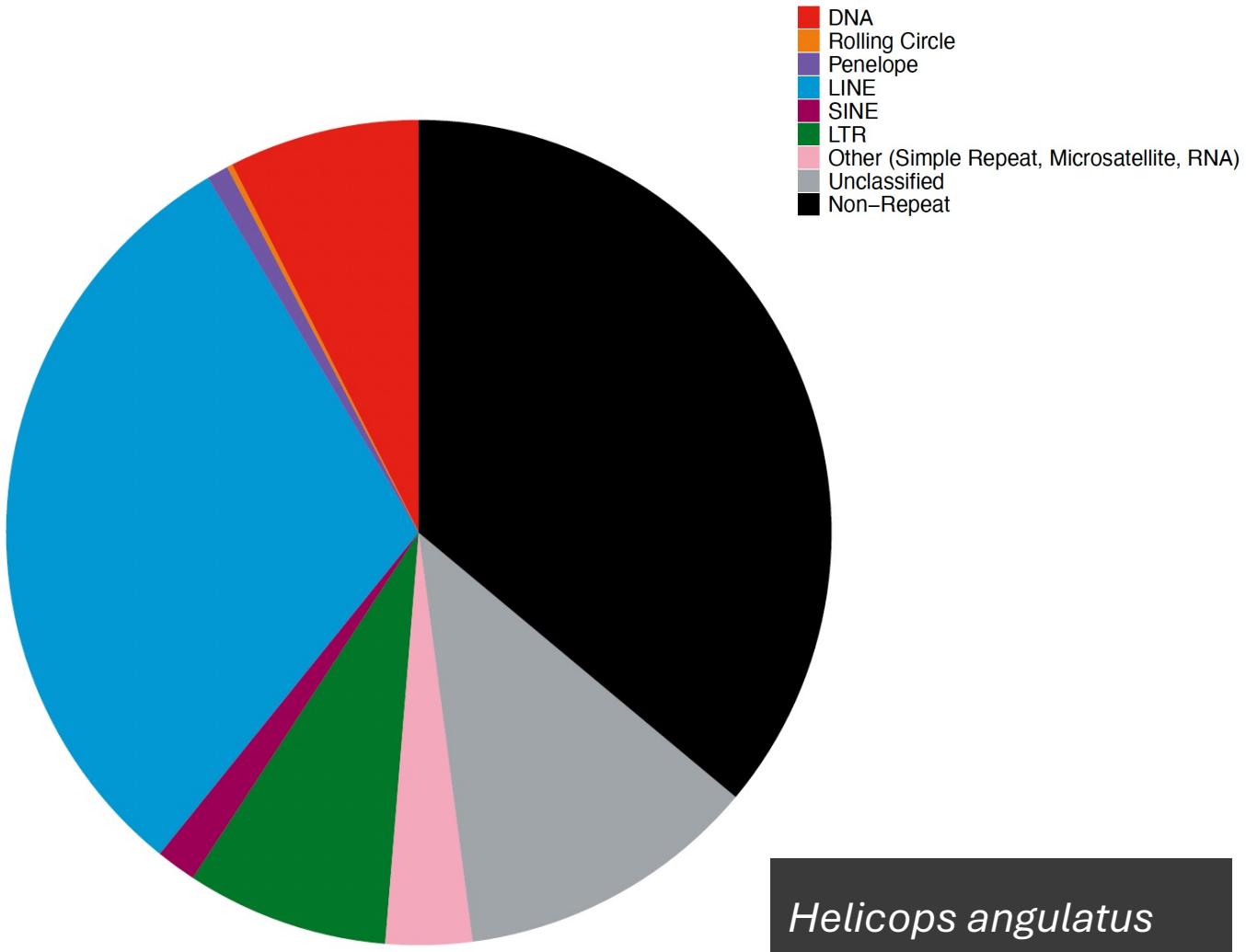
a) Mask genome using earlgrey
b) Assemble RNA-seq data using trinity
c) Quality check RNA-seq assembly with BUSCO scores

Genome annotation (partial- initial steps):

Mask genome using earlgrey

- Given an input genome, Earl Grey will run through numerous steps to identify, curate, and annotate transposable elements (TEs)

- It identifies and processes repetitive sequences to aid genome annotation or other downstream analyses



Helicops angulatus has a very repetitive genome (~60%) ☹

Genome annotation (partial- initial steps):

Assemble RNA-seq data using trinity

- RNA seq for 2 tissues: liver and kidney (CG1)

Steps performed:

1) Fastqc

2) Kmer correction of reads with rCorrector

3) Trimmomatic

4) Assemble genome with Trinity

Genome annotation (partial- initial steps):

Assemble RNA-seq data using trinity

- RNA seq for 2 tissues: liver and kidney (CG1)

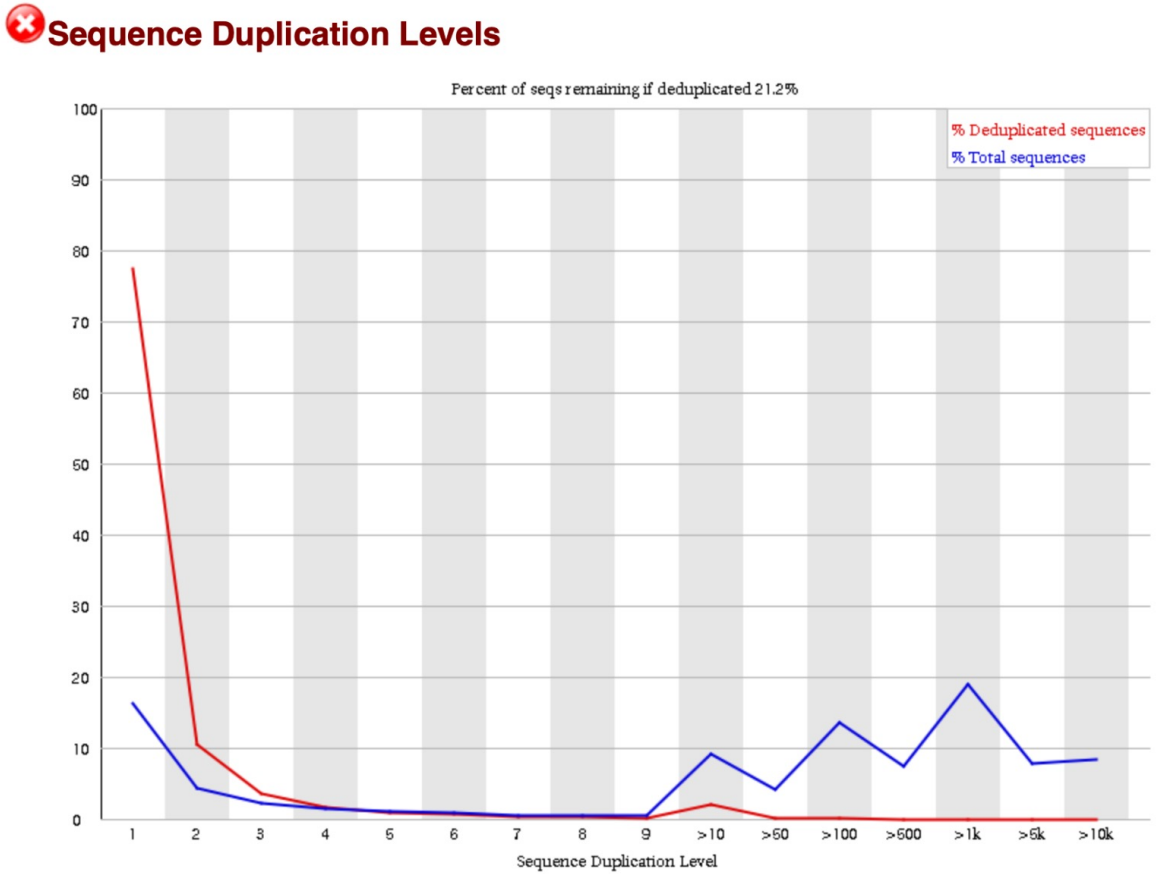
Steps performed:

1) Fastqc

2) Kmer correction of reads with rCorrector

3) Trimmomatic

4) Assemble genome with Trinity



Is this % of duplicates something I should worry about?

Genome annotation (partial- initial steps):

Assemble RNA-seq data using trinity

- RNA seq for 2 tissues: liver and kidney (CG1)

Steps performed:

Specimen	Tissue	Total PE reads	Removed PE reads	Retained PE reads	R1 corrected	R2 corrected	Pairs corrected	R1 unfixable	R2 unfixable	both reads unfixable
IAvH-CT-36861	Kidney	20150326	3082212	17068114	3831621	3814239	5935385	863403	1266925	951884
IAvH-CT-36861	Liver	21631298	3170764	18460534	4105252	4108408	6299004	727882	1218620	1224262

1) Fastqc

2) Kmer correction of reads with rCorrector

3) Trimmomatic

4) Assemble genome with Trinity

Around 15% of the reads were purged after correcting or eliminating bad quality reads

Genome annotation (partial- initial steps):

Assemble RNA-seq data using trinity

- RNA seq for 2 tissues: liver and kidney (CG1)

Steps performed:

1) Fastqc
2) Kmer correction of reads with rCorrector
3) **Trimmomatic**
4) Assemble genome with Trinity

This step trims the adaptors from library prep. After trimming adaptors I am left with paired reads of ~ 1 Gb (began with 1.5 Gb).

Genome annotation (partial- initial steps):

Assemble RNA-seq data using trinity

- RNA seq for 2 tissues: liver and kidney (CG1)

Steps performed:

1) Fastqc
2) Kmer correction of reads with rCorrector
3) Trimmomatic
4) **Assemble genome with Trinity**

Obtained two files per tissue as results: 1) name_Trinity.fasta and 2) name_Trinity.fasta.gene_trans_map

Genome annotation (partial- initial steps):

Quality check RNA-seq assembly with BUSCO scores

- Results for kidney:

```
# BUSCO was run in mode: euk_tran

        ***** Results: *****

        C:58.8%[S:34.8%,D:24.0%],F:7.4%,M:33.8%,n:7480
        4401    Complete BUSCOs (C)
        2604    Complete and single-copy BUSCOs (S)
        1797    Complete and duplicated BUSCOs (D)
        553     Fragmented BUSCOs (F)
        2526    Missing BUSCOs (M)
        7480    Total BUSCO groups searched
```

- Results for liver:

```
# BUSCO was run in mode: euk_tran

        ***** Results: *****

        C:42.7%[S:30.4%,D:12.3%],F:8.4%,M:48.9%,n:7480
        3196    Complete BUSCOs (C)
        2274    Complete and single-copy BUSCOs (S)
        922     Complete and duplicated BUSCOs (D)
        630     Fragmented BUSCOs (F)
        3654    Missing BUSCOs (M)
        7480    Total BUSCO groups searched
```

58% completeness in kidney

43% completeness in liver


Questions