**MOOC Assembly Report Interpretation**

What is a contig?

A contig is a long piece of DNA sequence that scientists put together by combining smaller pieces of DNA.

Easy Example

Imagine you are solving a puzzle. Each puzzle piece is a small DNA fragment. When you join these pieces correctly, you get a full picture — this full picture is the contig.

In genetics

Scientists read small pieces of DNA, then use computers to assemble them into a longer sequence. This long sequence is called a contig.

One-line definition

Contig = a continuous stretch of DNA that is built by combining overlapping small DNA sequences.

Want me to give you an example from real research or show you how contigs help in genome assembly?

You said:

yes

ChatGPT said:

Perfect! Let’s go a bit deeper (but still easy to follow) — here’s an example of how contigs are used in real research, especially in genome assembly.

Real-life Example: Assembling a Plant Genome

Let’s say scientists want to figure out the full DNA sequence of a new rice plant species.

DNA is very long, so machines can’t read it all at once. Instead, they break the DNA into small fragments — maybe 100 to 300 letters (nucleotides) long.

Each fragment is sequenced (meaning the machine reads the order of A, T, G, C).

Many fragments overlap, meaning they share some of the same letters.

Scientists use special software to stitch these overlapping fragments together, creating contigs.

Each contig is a larger piece of the genome — kind of like completing part of the puzzle.

Simple Example to Visualize

Imagine 3 DNA fragments:

Fragment 1: ATGCGT

Fragment 2: CGTGAA

Fragment 3: GAATCC

These overlap (look at bold parts):

Fragment 1: ATGCGT

Fragment 2: CGTGAA

Fragment 3: GGAATCC

After assembly, the contig would be:

ATGCGTGAATCC

Why are contigs important?

They help scientists rebuild whole genomes.

They make it easier to find genes or detect mutations.

The more complete and accurate the contigs, the better the genome assembly.

Want me to show how contigs fit into the whole genome assembly process, including scaffolds and chromosomes?

You said:

verage depth (short reads): 67.82

average short read coverage: 67.817

good contigs file: contigs\_pilon\_2\_depth\_cov\_filtered.fasta

min\_contig\_coverage\_threshold: 5.0

min\_contig\_length\_threshold: 300

num contigs above thresholds: 96

num contigs below thresholds: 0

total length of good contigs: 5185866 interpret the result in sam eway

ChatGPT said:

Absolutely! Let’s break this result down in simple language, so you can understand what’s happening here — especially related to contigs and genome assembly.

Explanation (line-by-line)

1. Average depth (short reads): 67.82

This means, on average, each position (nucleotide) in the genome was covered (read) about 68 times.

Imagine reading the same page of a book 68 times to make sure you didn’t miss anything — that’s what "depth" means here.

Higher depth = more confidence in the sequence.

2. Average short read coverage: 67.817

This is basically the same thing as average depth, just showing it with more decimal precision (it’s 67.817 instead of rounding to 67.82).

Coverage tells you how many times each part of the genome was sequenced.

3. Good contigs file: contigs\_pilon\_2\_depth\_cov\_filtered.fasta

This is the actual file containing the "good" contigs that passed quality filters.

Scientists will use this file to study the genome further (like gene finding or mutation analysis).

4. min\_contig\_coverage\_threshold: 5.0

This means only contigs that were covered at least 5 times were kept.

If a contig had too little data (low confidence), it was discarded.

5. min\_contig\_length\_threshold: 300

This means only contigs that were at least 300 nucleotides long were kept.

Shorter contigs were removed because they are often incomplete or unreliable.

6. num contigs above thresholds: 96

This is the number of contigs that passed both filters (coverage ≥ 5 and length ≥ 300).

These 96 contigs are considered high-quality.

7. num contigs below thresholds: 0

This means there were no contigs that failed the filters, so your data quality is quite good.

8. total length of good contigs: 5,185,866

This is the total length of all 96 good contigs combined — around 5.2 million nucleotides.

This gives you an idea of how much of the genome you’ve assembled so far.

In summary

✅ You have 96 good contigs.

✅ They cover a total of 5.2 million bases.

✅ Each base was read about 68 times, giving you strong confidence.

✅ All of your contigs passed the coverage (5x) and length (300 bp) filters — meaning no really weak or tiny contigs were included.

✅ Your filtered contigs are saved in contigs\_pilon\_2\_depth\_cov\_filtered.fasta.

Easy Interpretation

This means you successfully assembled a high-quality draft genome made of 96 contigs. The coverage (68x) is very good, meaning your sequence is quite reliable. You can now move to the next steps — such as annotating genes, finding mutations, or comparing this genome to others.

Want me to explain what you can do with these contigs next? Or how to check if the assembly is complete?

You said:

All statistics are based on contigs of size >= 300 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Statistics without reference contigs\_pilon\_2\_depth\_cov\_fil…

# contigs 96

# contigs (>= 0 bp) 96

# contigs (>= 1000 bp) 77

# contigs (>= 5000 bp) 49

# contigs (>= 10000 bp) 46

# contigs (>= 25000 bp) 40

# contigs (>= 50000 bp) 29

Largest contig 597230

Total length 5185866

Total length (>= 0 bp) 5185866

Total length (>= 1000 bp) 5175823

Total length (>= 5000 bp) 5110003

Total length (>= 10000 bp) 5092725

Total length (>= 25000 bp) 4987743

Total length (>= 50000 bp) 4590808

N50 170302

N90 44988

auN 228827

L50 9

L90 31

GC (%) 50.33

Mismatches

# N's per 100 kbp 0

# N's 0

ChatGPT said:

This is great data — I can walk you through how to interpret these statistics in simple language. These are assembly statistics, meaning they describe the quality and structure of your assembled genome (your set of contigs). Let’s break it down:

🔬 What this report is telling you

1️⃣ Number and size of contigs

# contigs = 96

➡️ You have 96 contigs in total after filtering.

# contigs (>= 1000 bp) = 77

➡️ Out of these 96, 77 are at least 1000 base pairs (bp) long — this is a good sign because longer contigs are more useful.

# contigs (>= 5000 bp) = 49

➡️ 49 contigs are at least 5000 bp long.

# contigs (>= 10000 bp) = 46

➡️ 46 contigs are at least 10,000 bp long.

# contigs (>= 25000 bp) = 40

➡️ 40 contigs are at least 25,000 bp long.

# contigs (>= 50000 bp) = 29

➡️ 29 contigs are at least 50,000 bp long.

🧩 What this means

The more long contigs you have, the better your assembly. Short contigs mean your assembly is fragmented (like a broken puzzle), while long contigs mean you successfully stitched big chunks together.

2️⃣ Total length of the assembly

Total length = 5,185,866 bp

➡️ Your total assembled genome is about 5.2 million base pairs. This is typical for bacteria or smaller genomes.

Total length (>= 1000 bp) = 5,175,823 bp

➡️ Almost all your contigs are over 1000 bp, which shows most of your assembly is made up of useful, reliable sequences.

Total length (>= 5000 bp) = 5,110,003 bp

➡️ Even better — a large portion of your assembly is made up of long contigs (5000 bp+).

3️⃣ Key quality metrics: N50, N90, auN

N50 = 170,302 bp

➡️ This is a famous assembly quality metric. It means:

If you sort all your contigs from longest to shortest, and start adding their lengths, once you reach 50% of the total genome length, the contig at that point is 170,302 bp long.

Bigger N50 = better assembly.

For your data, 170 kb is quite good, meaning you have some very long contigs.

N90 = 44,988 bp

➡️ This is the same idea, but for the point where you reach 90% of the genome.

N90 is lower than N50 (expected), but still healthy — almost 45 kb.

auN = 228,827

➡️ This is a more modern version of N50, but weighted differently. It confirms that you have relatively large contigs overall.

4️⃣ L50 and L90

L50 = 9

➡️ This means: just 9 contigs make up half the genome length.

L90 = 31

➡️ Just 31 contigs make up 90% of the genome.

➡️ Smaller L50 and L90 are better, because it means fewer, bigger contigs cover the genome. This is a very good result — your assembly is not too fragmented.

5️⃣ GC content

GC% = 50.33%

➡️ This is the proportion of Guanine (G) and Cytosine (C) bases in your genome. Different organisms have different GC content — 50% is fairly balanced, which is typical for many species.

6️⃣ Gaps and errors

# N’s per 100 kbp = 0

# N’s = 0

➡️ No gaps ("N" bases) in your assembly. This is excellent — it means all contigs are fully resolved sequences.

✅ Overall interpretation (easy summary)

Metric Interpretation

Number of contigs 96 is good for smaller genomes, not too fragmented

Contig length Many contigs over 50kb, showing strong assembly

Total genome size ~5.2 million bp — reasonable size (likely bacterial-sized genome)

N50 170 kb — good quality

GC% Balanced (50%) — looks typical

Gaps None (perfect)

🌱 In short:

✅ This is a high-quality draft genome assembly.

✅ You have good long contigs and high coverage, which means this genome can be used for downstream analysis like:

Finding genes

Comparing to other species

Looking for mutations or interesting features

Functional annotation (what each gene does)

Want me to help you write a summary report or explain how to improve the assembly further if needed?