Progress Report #1

2025/6/9

Akinobu Ono

Dictyostelium discoideum Genome Assembly

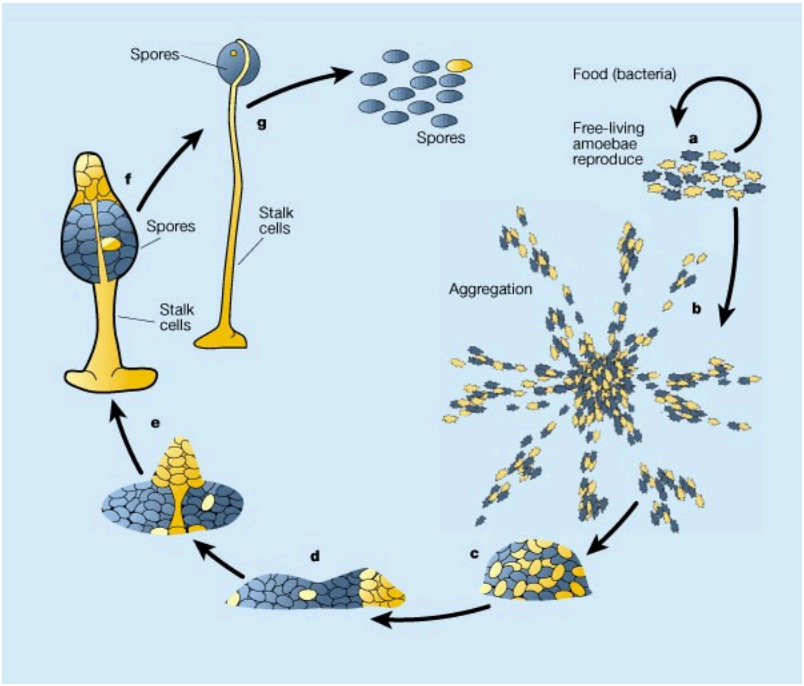
- **Objective**: Construction of high-quality genome sequence for model organism (chromoosome level) *D. discoideum*
- **Significance**: Centromere structure, chromosome segregation machinery, co-evolution analysis

Background: About *D. discoideum*

- Model organism of social amoeba
- Life cycle: Single-cell ⇄ Multicellular
- Genome size: ~34.2 Mb
- Chromosomes: 6 + extrachromosomal rDNA (~88 kb × ~100 copies) + mitochondria (~56 kb)

Genome Characteristics

- **AT-rich genome (77.6%)** – Challenges sequencing due to polymerase slippage, requiring long reads.
- **Numerous tRNA genes (≈ 390 copies)** – Similar clustered genes complicate assembly, needing long reads for accurate placement.
- **SSR-rich (> 11%)** – Dense repeats cause assembly ambiguities but can serve as markers.

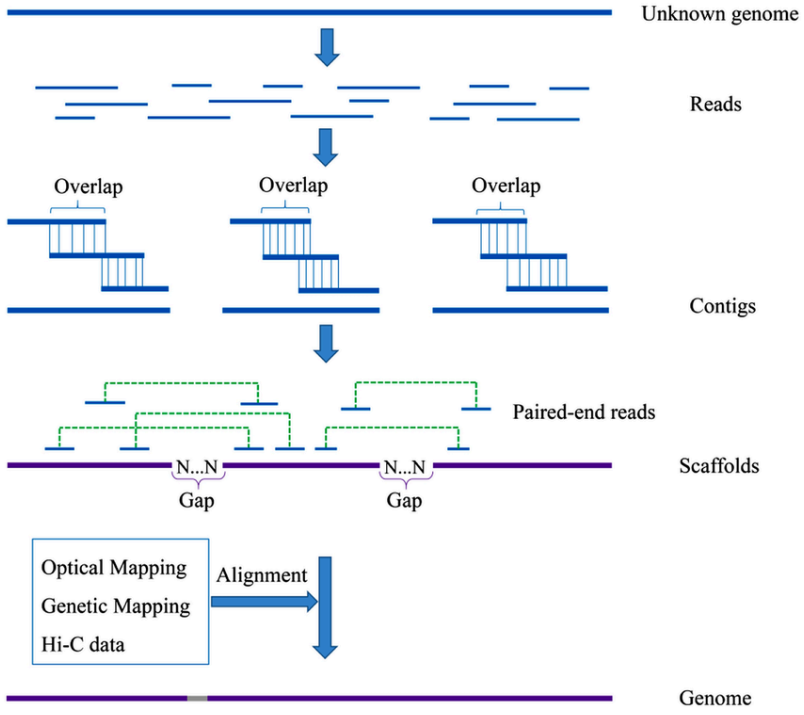


Comparison of ONT and Illumina Data

	ONT Long Reads	Illumina Short Reads
Read Length	Very long (up to ~139 kb)	Short (~150 bp)
Accuracy	Lower (high error rate)	Very high
Error Type	Indels, mismatches	Rare, mostly substitutions
Strengths	Resolves repeats, large SVs	Ideal for polishing
Weaknesses	Lower per-base accuracy	Cannot span long repeats

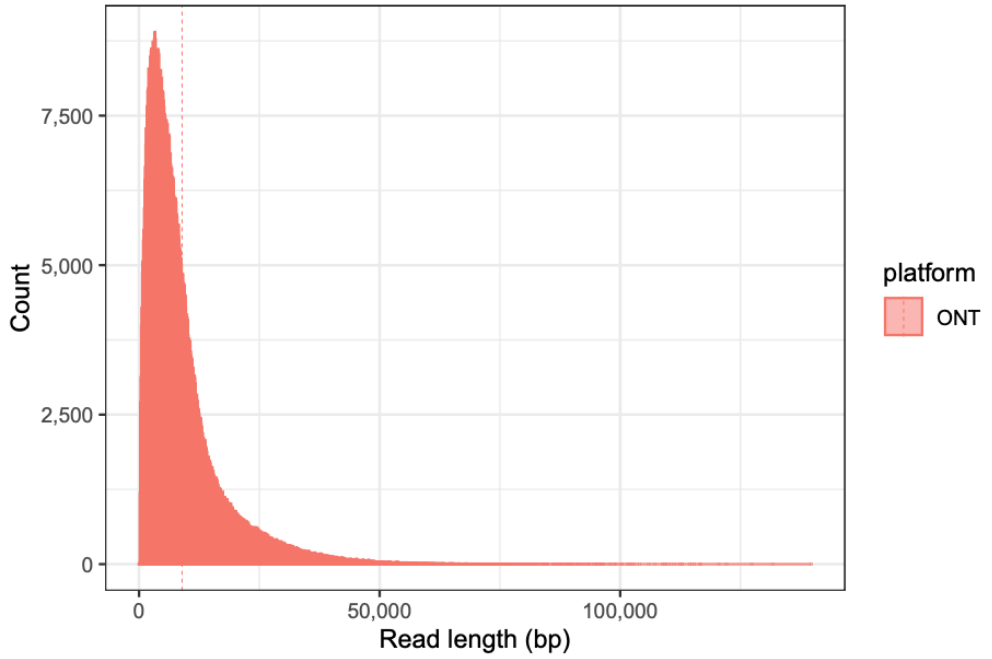
Genome Assembly Workflow

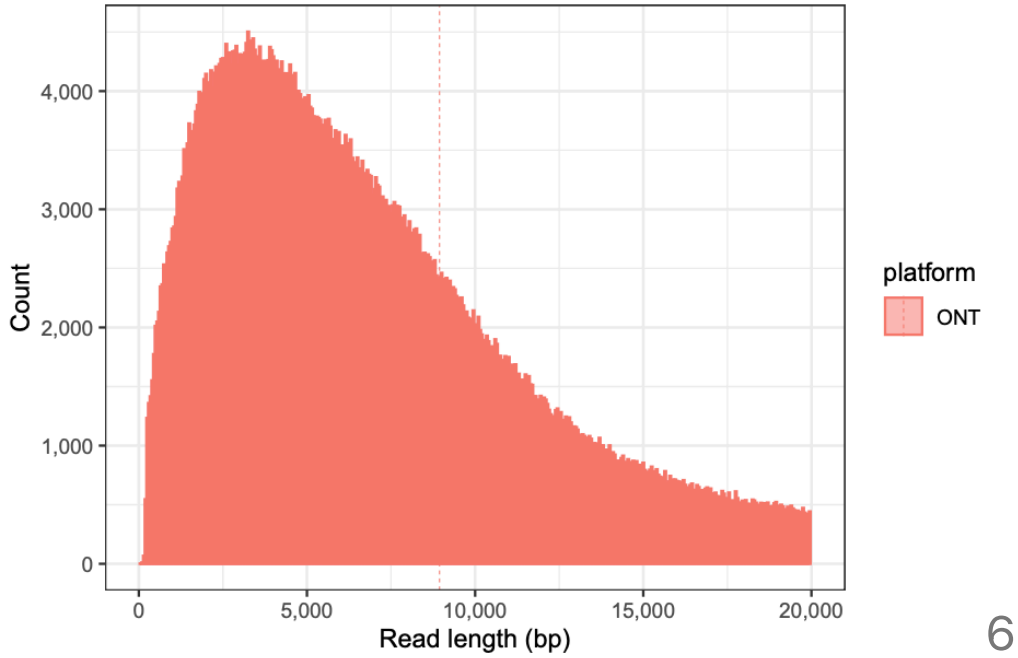
1. **Sequence Data Acquisition**
   - ONT (Long reads)
   - Illumina (Short reads)
2. **Quality Assessment & Preprocessing**
   - Read quality confirmation
3. **Assembly Execution**
   - Canu / Flye / Raven / Shasta
     → Comparison with QUAST
4. **Polishing (Error Correction)**
   - Pilon / Medaka
5. **Evaluation & Improvement**
   - Quality assessment with QUAST → Reassembly or Scaffolding as needed



Dictyostelium discoideum ONT Read Length Distribution

```
# Total length of bases in the sequence
sum = 8,359,638,019 bp

# Total number of reads
n = 934,886 reads

# Average read length
mean length = 8,941.88 bp

# Length of the longest read
max length = 139,714 bp

# Read length where 50% of total sequence
N50 = 12,777 bp
```



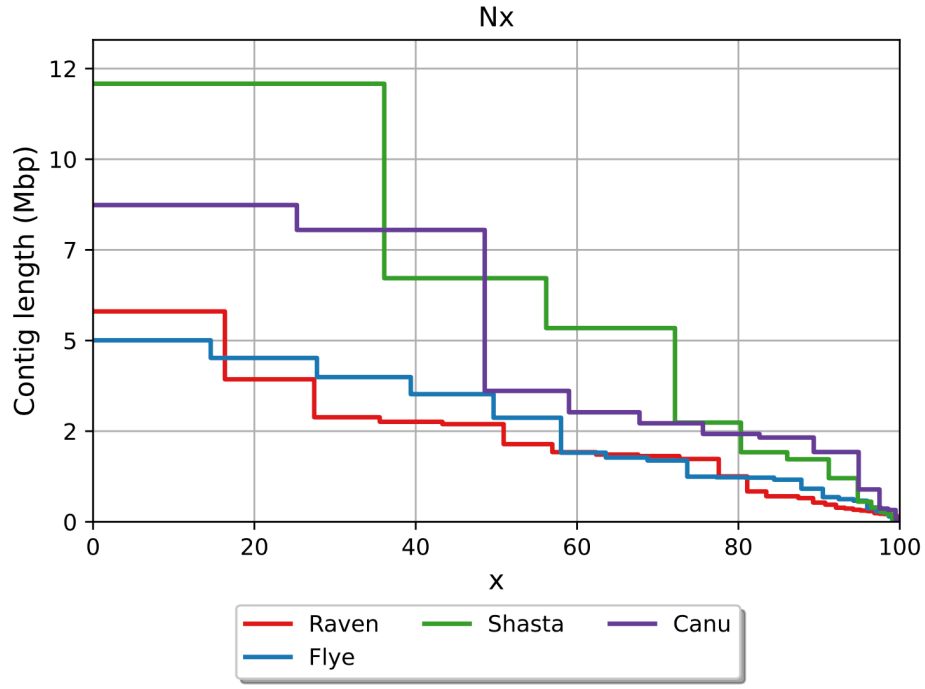


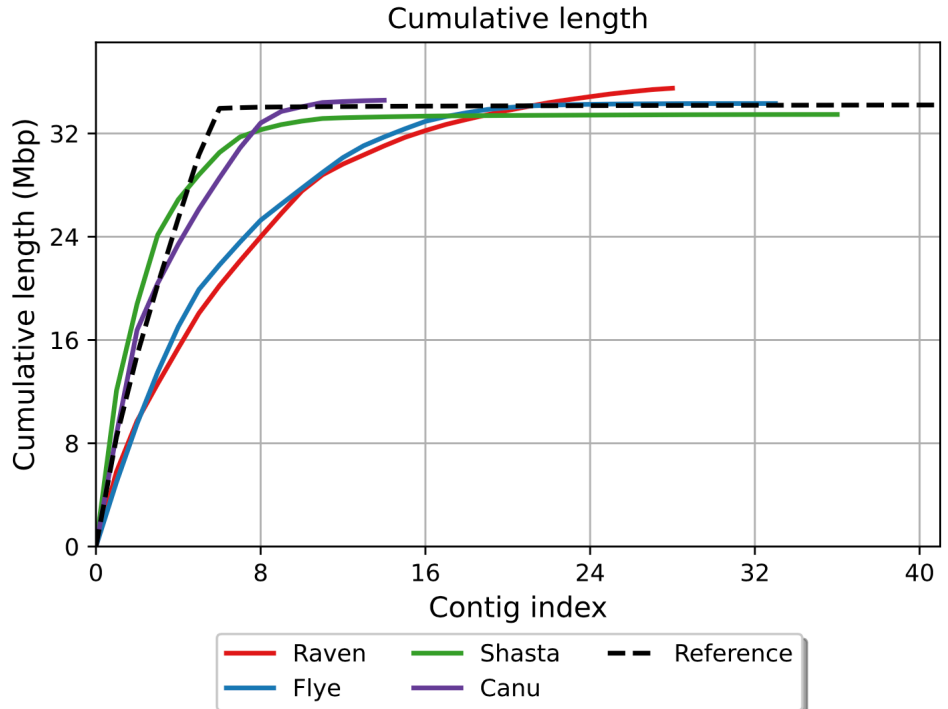
Assembly Experiment Overview

- **Data Used**: ~50% of ONT long reads (4.2 Gb)
    - Why 50%?...Excessive coverage can lead to increased computation time and reduced accuracy
    - Also testing other coverage levels (e.g., 25% and 75%), but 50% was most accurate
- **Assembly Tool Characteristics**:
    - Canu: Powerful error correction, longer computation time
    - Flye: Strong with repetitive sequences, good memory efficiency
    - Shasta: Ultra-fast but slightly lower accuracy
    - Raven: Low memory usage, high speed

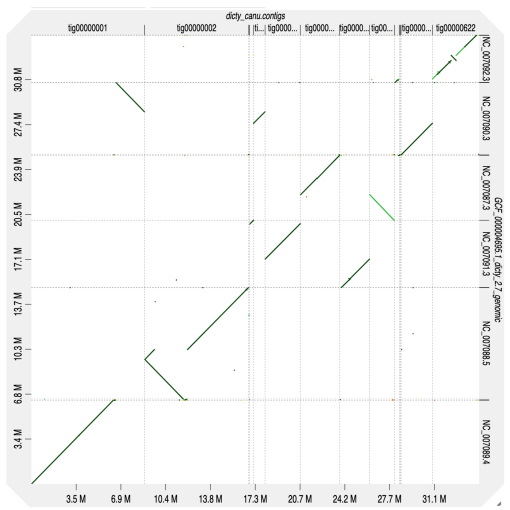
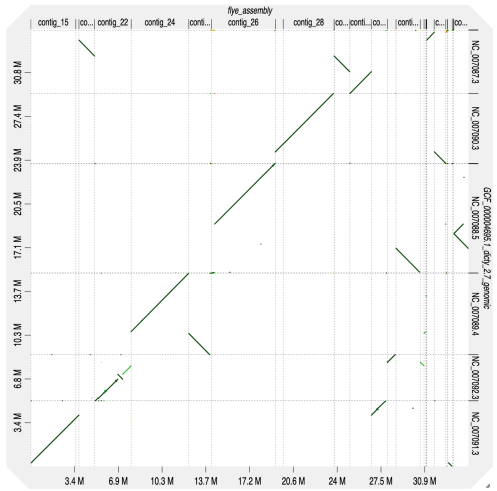
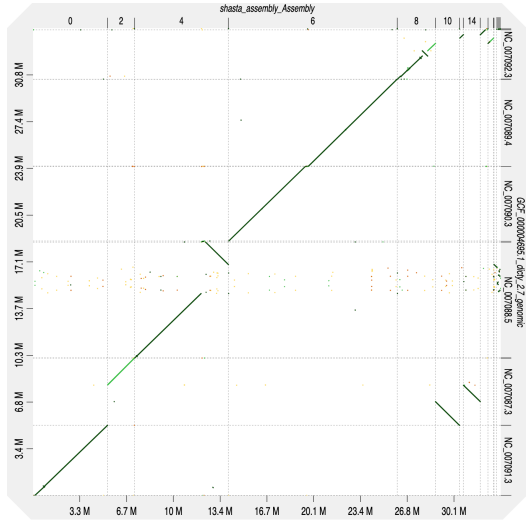
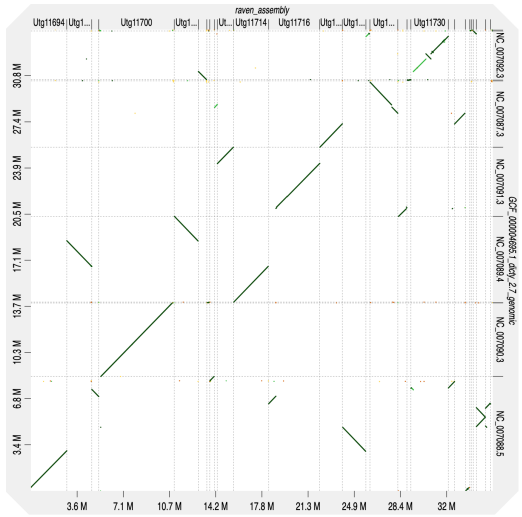
Assembly Result Comparison

Metric	Raven	Flye	Shasta	Canu
contigs	28	33	36	14
Largest contig	5.8 Mb	5.0 Mb	12.0 Mb	8.7 Mb
Total length (Expected longer than ~34.2 Mb)	35.5 Mb	34.3 Mb	33.5 Mb	34.6 Mb
N50	2.7 Mb	2.8 Mb	6.7 Mb	3.6 Mb





Evaluate Assembly Accuracy

Canu	Flye	Shasta	Raven
			

BUSCO Score

Metric	Canu	Shasta	Raven	Flye	Description & Ideal
Complete (C)	94.9 %	91.4 %	94.9 %	94.9 %	Fraction of expected genes found completely. Higher is better (ideally > 95 %).
• Single-copy (S)	236 (92.5 %)	229 (89.8 %)	235 (92.2 %)	236 (92.5 %)	Single-copy orthologs without duplication. Should be high to show low redundancy (ideally > 90 %).
• Duplicated (D)	6 (2.4 %)	4 (1.6 %)	7 (2.7 %)	6 (2.4 %)	Orthologs found more than once. Low duplicated count is good (ideally < 5 %).
Fragmented (F)	3 (1.2 %)	3 (1.2 %)	3 (1.2 %)	3 (1.2 %)	Partial matches of expected genes. Lower is better (ideally < 2 %).
Missing (M)	10 (3.9 %)	19 (7.5 %)	10 (3.9 %)	10 (3.9 %)	Genes not detected. Fewer missing is better (ideally < 5 %).
Stop-codon errors (E)	2 (0.8 %)	1 (0.4 %)	2 (0.8 %)	3 (1.2 %)	Complete genes containing internal stops. Few errors are acceptable (ideally < 1 %).
Total BUSCOs (n)	255	255	255	255	Number of BUSCO groups searched. Always constant for the chosen lineage.

Overall Assembly Evaluation

Tool	Evaluation	Comments
Canu	★ High accuracy, low fragmentation	- Fewest contigs (14), good N50 (3.6 Mb), max contig 8.7 Mb - Consistently strong in Nx/cumulative plots
Shasta	🟢 Good for structure	- Longest contig (12 Mb), top N50 (6.7 Mb) - Nx/cumulative plots: covers most with few contigs - Highest contig count (36)
Flye	△ Balanced	- Similar contig/N50 to Raven - Max contig smaller (5 Mb), total length moderate (34.3 Mb)
Raven	△ Fast & practical	- Longest total length (35.5 Mb), max contig 5.8 Mb - N50/Nx lower than Shasta/Canu

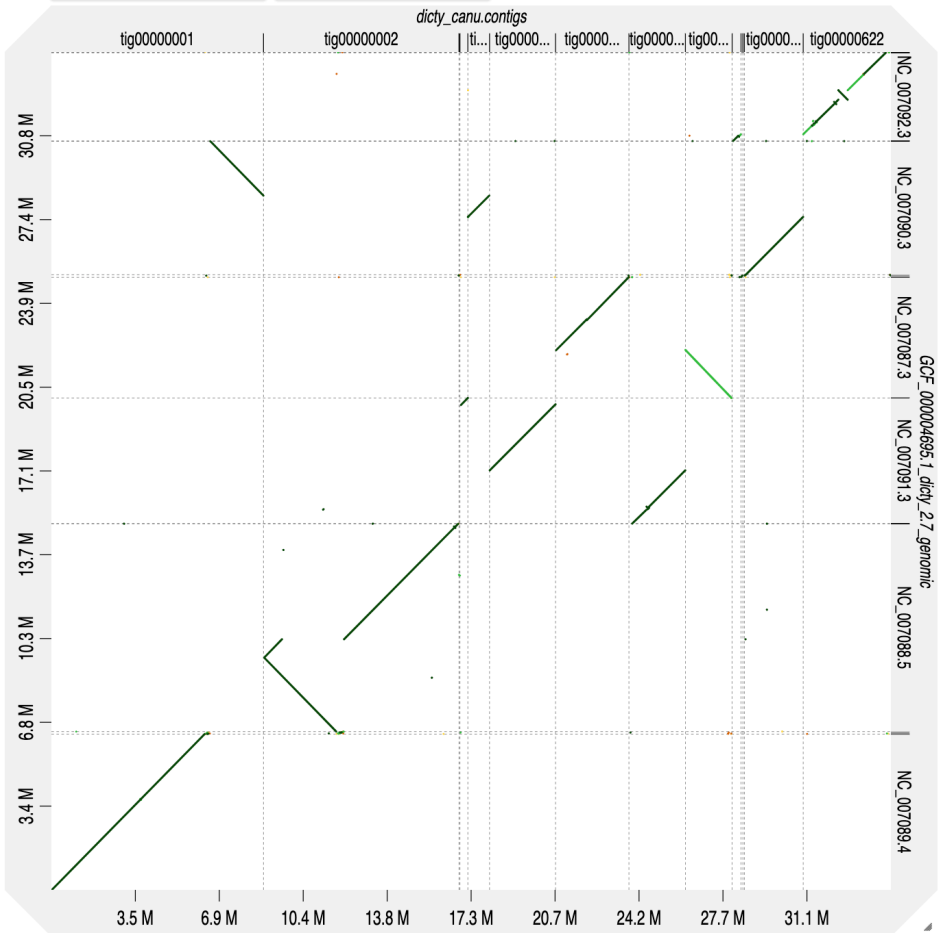
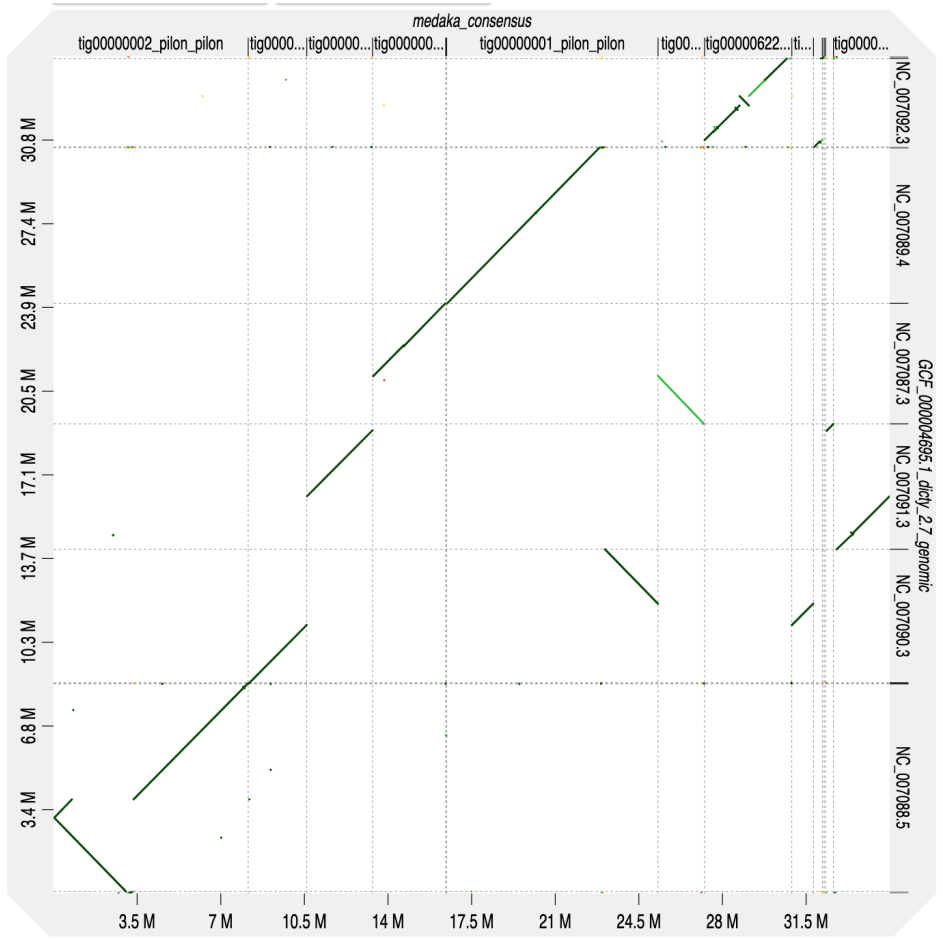
Polishing Experiment Overview

Polishing...Process of correcting errors in the assembled genome sequence to improve its accuracy.

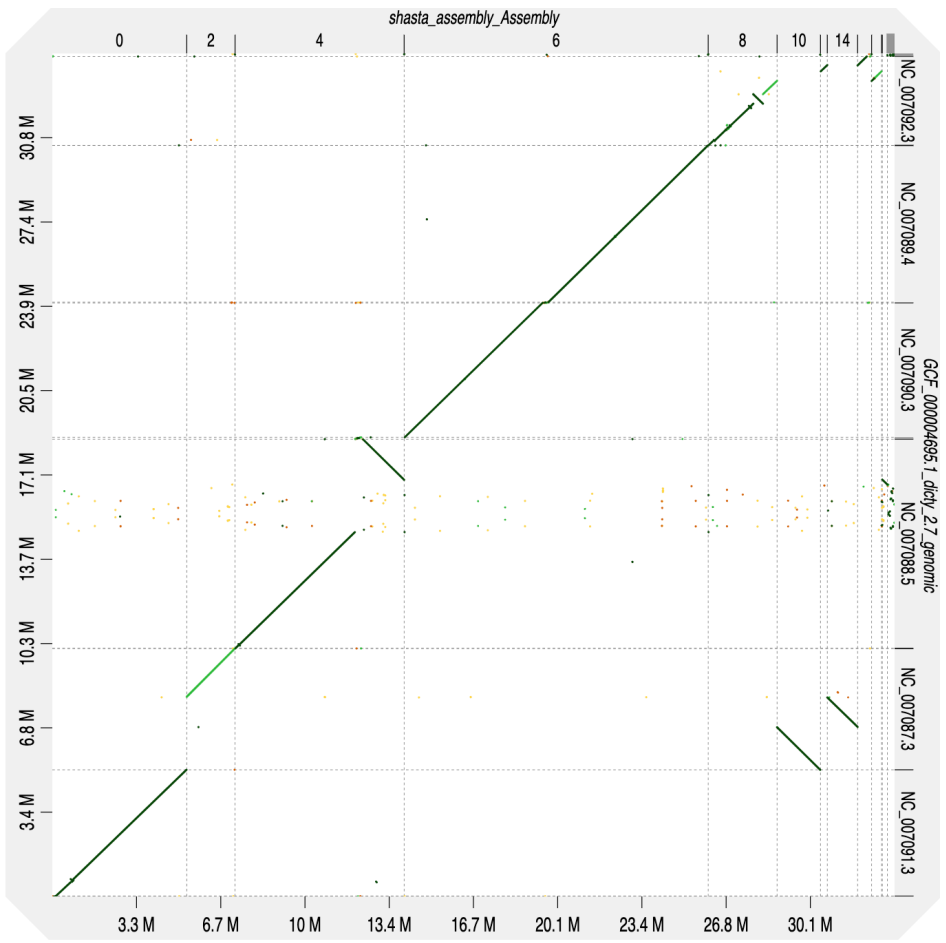
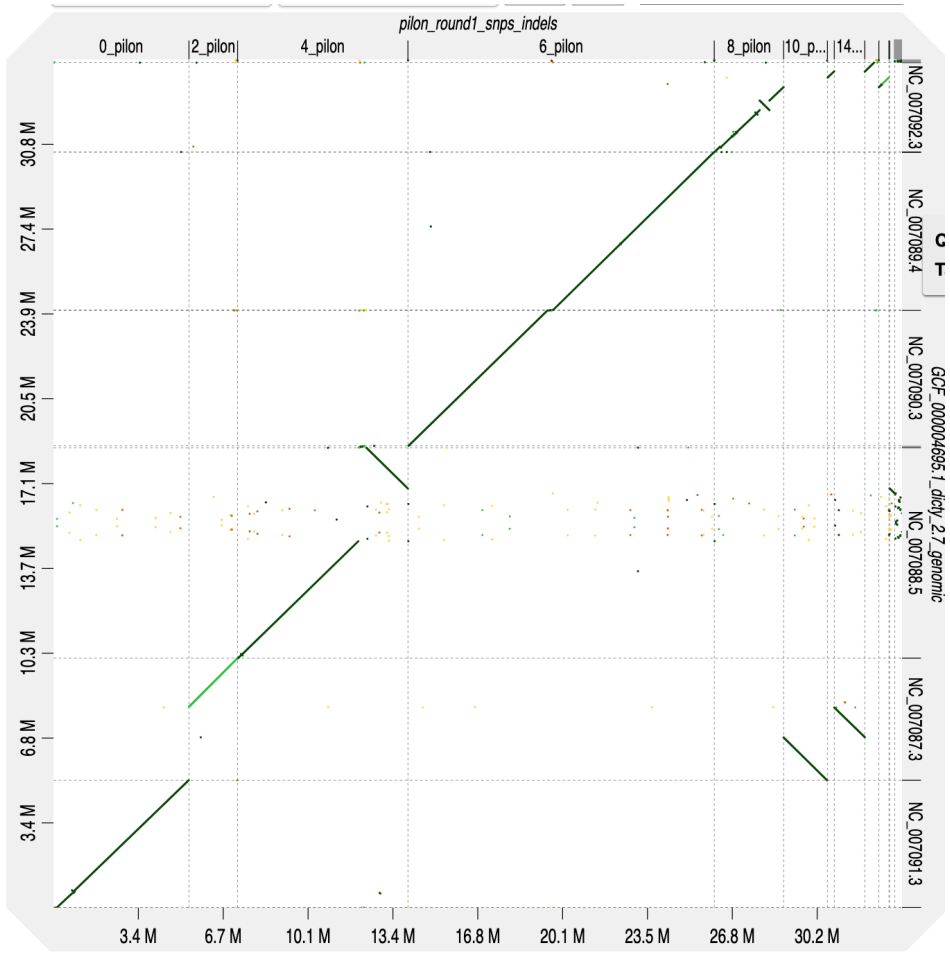
- **Procedure**:
  i. Pilon
    - Using Illumina reads and ONT long reads
    - Effective for base substitution and indel corrections

  ii. Medaka
    - Using ONT reads
    - Pre-trained on ONT-specific error patterns
    - Strong in homopolymer region correction

Evaluate Canu Accuracy Improvement

Canu Before Polishing	Canu Pilon+Medaka
	

Evaluate Shasta Accuracy Improvement

Shasta Before Polishing	Shasta Pilon
	

Future Directions

1. **Compare Polishing Result**
    - Select better performing assembly for further improvement
2. **Advanced Polishing**
    - Homopolish (for homopolymer regions)
    - NextPolish (alternative approach)
3. **Introduce Scaffolding**
    - Apply scaffolding to polished assembly
4. **Multi-assembly Integration**
    - Combine best features from both assemblies