Comparative Genome Assembly of Indonesian Turmeric (Curcuma longa L.) Using Illumina Short Reads and Oxford Nanopore Long Reads

Fransiscus Jason Wiguna[1], Hadi Rayhan Hanifah Pawitan[1], Rizkita Rahmi Esyanti[1], Husna Nugrahapraja[1], Agung Karuniawan[2], Popi Septiani[1]*

[1] School of Life Sciences and Technology, Institut Teknologi Bandung, Indonesia
[2] Faculty of Agriculture, Universitas Padjadjaran, Indonesia
*Correspondence: popi.septiani@itb.ac.id

Abstract

Turmeric (Curcuma longa L.) is an economically important medicinal and culinary plant, but genomic resources representing Indonesian germplasm remain limited. This study compared draft genome assemblies generated using Illumina short-read and Oxford Nanopore Technologies (ONT) long-read sequencing to evaluate their quality and completeness. Illumina NovaSeq 6000 sequencing generated approximately 150 million paired-end reads (150 bp), totaling 22.5 Gbp at a targeted 10× coverage. The curated short-read assembly comprised 368.22 Mb across 9,217 scaffolds, with an N50 of 14.63 Mb and 74.5% BUSCO completeness. ONT PromethION sequencing generated 12.63 Gbp from 3,604,446 reads, with a read N50 of 5.89 kb. The curated long-read assembly comprised 368.89 Mb across 7,139 scaffolds, with an N50 of 12.81 Mb and 94.1% BUSCO completeness. Notably, the ONT assembly contained only 429,785 ambiguous bases, compared with 79.67 Mb in the Illumina assembly. Conversely, the Illumina assembly had a higher scaffold N50 and a longer largest scaffold. Repeat annotation identified only 2.24% repetitive sequences in the ONT assembly, suggesting incomplete repeat representation. Overall, ONT sequencing provided substantially better gene-space completeness and fewer ambiguous bases, while Illumina retained a higher N50. The resulting Indonesian turmeric draft genome provides a valuable resource for further genomic research and molecular breeding. However, additional sequencing, repeat-aware assembly refinement, and chromosome-scale scaffolding are needed to develop a comprehensive reference genome.

Keywords: Curcuma longa; genome assembly; short read sequencing; long-read sequencing; BUSCO

Introduction

Turmeric (Curcuma longa L.) is an important medicinal and culinary crop in Indonesia, valued for its rhizomes containing curcuminoids. Molecular breeding using genome-wide SNP markers offers opportunities to improve agronomic and biochemical traits, but reliable marker discovery requires a representative and sufficiently complete reference genome.

Turmeric genome assembly is challenging due to its complex and repetitive genomic regions. While Illumina short reads provide high sequencing accuracy, their limited length may hinder repeat resolution. In contrast, Oxford Nanopore Technologies (ONT) long reads can span longer genomic regions and potentially improve assembly completeness.

Although previous turmeric genome assemblies have been reported from Indian and Chinese accessions, genomic resources for Indonesian germplasm remain limited. This study compares Illumina- and ONT-based draft genome assemblies of Indonesian turmeric to evaluate assembly size, scaffold statistics, N50, ambiguous bases, GC content, and BUSCO completeness, providing insights for further reference genome improvement.

Objectives

To generate and compare draft genome assemblies of Indonesian turmeric (*Curcuma longa* L.) using Illumina short reads and Oxford Nanopore long reads.

Research Method

1 Plant Material & Genomic DNA



Curcuma longa L.

- **Material:** Selected genotypes from Ciparanje Experimental Field, Universitas Padjadjaran (±5-month-old leaves)
- **DNA extraction:** GenElute (Illumina); Promega Wizard + Geneaid (ONT)
- **Quality control:** Spectrophotometry, agarose gel, Qubit, TapeStation
- **Species confirmation:** ITS2 + Sanger → BLAST (*C. longa;* MK087904.1)

2 Genome Sequencing

Illumina (short reads)



NovaSeq 6000

- Paired-end (2 × 150 bp)
- ~150 million reads
- 22.5 Gbp (~10×)

ONT (long reads)



PromethION

- PromethION (FLO-PRO114M, R10.4.1)
- 3,604,446 reads
- 12.63 Gbp
- Read N50: 5.89 kb (longest: 92.39 kb)

3 Genome Assembly & Processing

Illumina workflow

- QC & trimming (FastQC, Trimmomatic)
- Assembly (SOAPdenovo2, k=33–83)
- Scaffolding (RagTag; Chinese reference)
- Curation & evaluation (BBMap <150 bp, QUAST, BUSCO, BWA-MEM2)

ONT workflow

- QC & trimming (LongQC, Porechop)
- De novo assembly (Flye v2.9.3)
- Polishing & scaffolding (Racon → RagTag → Racon; flye_r_rag_r1)
- Curation & evaluation (remove contaminant, <1,000 bp, terminal Ns; QUAST, BUSCO)

4 Genome Profile & Repeat Analysis

Genome profile (from Illumina k-mer)

- **Tool:** GenomeScope v2.0
- **Outputs:** genome size (641.3–661.6 Mb), homozygous fraction (97.37–97.55%), repetitive length (~402.4–415.1 Mb)
- **Ploidy:** Smudgeplot v0.2.2 (diploid pattern)

Repetitive sequence (from ONT assembly)

- **Tools:** RepeatModeler v2.0.5 & RepeatMasker v4.1.6
- **Libraries:** RepBase (20181026) + *de novo* library
- **Output:** repeat classes and proportions (total repeat: 2.24%)

5 Comparative Assessment

- **Data:** curated assemblies (JBEHDU000000000: Illumina; JBFDAI000000000: ONT)
- **Evaluation:** QUAST (assembly length, scaffold number & size, GC content, N50, ambiguous bases) and BUSCO v5.4.7 (viridiplantae_odb10)
- **Note:** comparison used curated assemblies (after removal of contaminants, short sequences, and terminal Ns).

Results

Sequencing Output

Illumina (short reads)



NovaSeq 6000

- 150 million paired-end reads (2 × 150 bp)
- 22.5 Gbp (~10× coverage)

ONT (long reads)



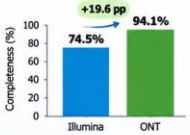
PromethION

- 3,604,446 reads
- 12.63 Gbp
- Read N50: **5.89 kb**
- Longest read: **92.39 kb**

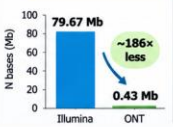
Genome Assembly Comparison (curated assemblies)

Assembly metric	Illumina (short reads)	ONT (long reads)
Total assembly length (Mb)	368.22	368.89
Number of scaffolds	9,217	7,139
N50 (Mb)	**14.63**	12.81
Longest scaffold (Mb)	**34.44**	32.25
GC content (%)	37.45	37.76
Ambiguous bases, N (Mb)	79.67	**0.43**
BUSCO completeness (%)	74.5	**94.1**

BUSCO Completeness



Ambiguous Bases (N)



Repeat Annotation (ONT assembly)

2.24%

repetitive sequences detected (likely incomplete)

Discussion

Gene-space completeness
- ONT achieved higher BUSCO completeness (94.1% vs. 74.5%; +19.6 pp).
- Indicates better recovery of conserved plant genes.

Assembly gaps
- Markedly fewer ambiguous bases in ONT (0.43 Mb vs. 79.67 Mb; ~186× less).
- Long reads span repetitive/complex regions more effectively, reducing gaps.

Contiguity (scaffold size)
- Illumina had a higher N50 (14.63 vs. 12.81 Mb) and a longer largest scaffold (34.44 vs. 32.25 Mb).
- Likely reflects differences in assembly strategy (reference-guided scaffolding) and data characteristics.

Repetitive sequences
- Only 2.24% repeats detected in the ONT assembly.
- Suggests that repeat content is incompletely represented, and further sequencing and repeat-aware assembly are needed.

Conclusion



- Illumina and ONT produced draft assemblies of similar total length (~368.2–368.9 Mb) but with different assembly quality.
- ONT long reads improved gene-space completeness (BUSCO 94.1%) and substantially reduced ambiguous bases (0.43 Mb), whereas the Illumina assembly retained a higher N50 (14.63 Mb).

Implication

- Provides a valuable genomic resource for SNP discovery and genomic breeding.
- Supports genetic diversity analysis and marker-assisted selection.
- Informs strategies for further genome improvement (e.g., additional sequencing, repeat-aware assembly, and chromosome-scale scaffolding).

Main Takeaway

- **ONT long reads:** better gene-space representation and fewer gaps.
- **Illumina short reads:** higher assembly contiguity (N50).