scientific data

Check for updates

OPEN

DATA DESCRIPTOR

The complete telomere-to-telomere genome assembly of *Lablab purpureus* (L.) Sweet

Heqin Liu[1,4], Xiuhui Liu[1,4], Guihua Zou[1], Xueqiang Zheng[2], Tao Zheng[1], Haomin Lyu[3] ✉, Qiaojun Lou[1] ✉ & Heyun Chen[2] ✉

Lablab purpureus **(L.) Sweet, a resilient tropical legume critical for sustainable agriculture, is prized for its adaptability, nutritional richness, and drought tolerance. Here, we present the first telomere-to-telomere (T2T) genome assembly of *Lablab purpureus* (L.) cv. Linghu integrating PacBio HiFi sequencing, ultra-long Oxford Nanopore Technologies (ONT) reads, and Hi-C scaffolding. The complete assembly spanning 460.76 Mb with contig N50 of 40.3 Mb resolved all the gaps, telomeres, and centromeres across the 11 chromosomes. We annotated 29,280 protein-coding genes, with functional roles assigned to 90.3% of genes. Repetitive elements constitute 38.38% of the genome, dominated by long terminal repeat retrotransposons (31.14%). Crucially, we resolved all centromeres (1.42–6.97 Mb) and telomeres, addressing pervasive gaps and mis-assemblies in prior assemblies. Rigorous validation using Merqury, CRAQ, and BUSCO confirmed exceptional assembly accuracy and continuity. This T2T genome provides an unparalleled resource for functional genomics in *L. purpureus*, enabling precise exploration of stress adaptation, nutrient biosynthesis, and yield-related traits. It further establishes a gold-standard reference for comparative legume genomics and molecular breeding in understudied crops.**

Background & Summary

Lablab purpureus (L.) Swee t (Fig. 1), commonly known as lablab or hyacinth bean, is a resilient tropical legume native to Africa and cultivated globally in tropical and subtropical regions for its exceptional adaptability to harsh climates and nutrient-poor soils[1–4]. Its remarkable drought and salinity tolerance establish it as a critical multipurpose crop for sustainable agriculture, contributing to food security, livestock feed, and soil conservation in marginal environments[5–7].

Early molecular studies on *L. purpureus* focused on population genetics, domestication history, and the genetic basis of agronomic traits such as flowering time and pod morphology[2,8–10]. As the sole member of the genus *Lablab* and phylogenetically distinct from cultivated legumes in the *Vigna* and *Phaseolus* genera, this species offers unique opportunities to study convergent evolution and domestication syndromes within the legume family[4,11,12]. The population genetic evidences suggest independent domestication events for two-seeded and four-seeded accessions of *L. purpureus*[1,4,9]. This makes it a compelling system to investigate genomic mechanisms underlying repeated evolution—a phenomenon that remains poorly understood and of enduring interest to evolutionary biologists[13,14]. Furthermore, its extreme drought tolerance positions it as a vital genetic resource for improving climate resilience in crops[2,8,15–18]. However, the lack of a complete genome assembly has hindered deeper exploration of its genomic architecture and adaptive traits.

Genomic research on *L. purpureus* has gradually increased in recent years[2,8]. A decade ago, the African Orphan Crops Consortium (AOCC) prioritized it for genomic sequencing due to its nutritional and agronomic value[19,20], yielding a quite fragmented draft genome (~395 Mb) with limited utility[21]. Subsequent efforts leveraging long-read sequencing and Hi-C scaffolding produced chromosome-scale assemblies for African (cv. Highworth, ~426 Mb) and Southeast Asian (cv. C204, ~367 Mb) cultivars, yet these references remain incomplete, with unresolved centromeres, telomeres, and discordant chromosomal synteny[4,12]. These inconsistencies

[1]Institute of Virology and Biotechnology, Zhejiang Academy of Agricultural Sciences, Hangzhou, China. [2]Institute of Crop and Nuclear Technology Utilization, Zhejiang Academy of Agricultural Sciences, Hangzhou, China. [3]HuaZhi Biotechnology Co., Ltd, Changsha, China. [4]These authors contributed equally: Heqin Liu, Xiuhui Liu. ✉e-mail: Haomin.Lyu@hotmail.com; lqj@sagc.org.cn; 2044670266@qq.com

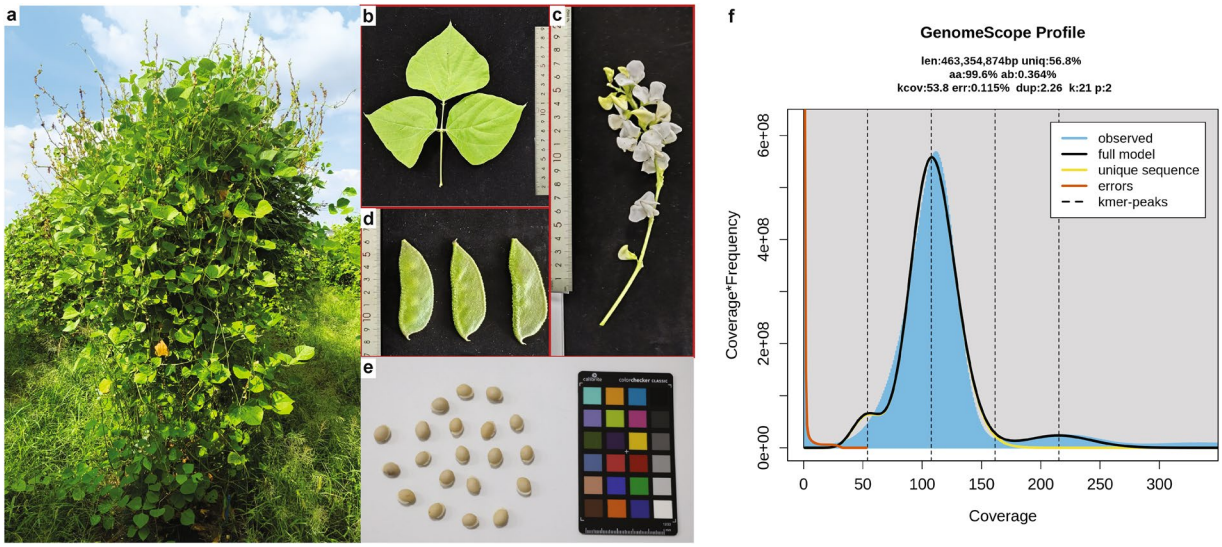

Fig. 1 Phenotypes and genomic survey of *Lablab purpureus* cv. Linghu including (1) Seedling, (**b**) leaf, (**c**) flower, (**d**) pod, and (**e**) seed. (**f**) The genome survey result using GenomeScope2.

underscore the need for a telomere-to-telomere (T2T) assembly to resolve structural ambiguities and enable precise genomic studies.

Here, we present the first complete T2T genome of *L. purpureus* cv. Linghu, a high-yielding cultivar adapted to the warm, humid riparian habitats of the Lake Tai region, where it is traditionally cultivated on pond banks using trees for support. This cultivar exhibits resilience to seasonal fluctuations and thrives in well-drained soils, reflecting its adaptation to local agricultural ecosystems. Combining PacBio HiFi sequencing, ultra-long Oxford Nanopore Technologies (ONT) reads, and Hi-C scaffolding, we generated a 460.76 Mb assembly spanning all 11 chromosomes, with centromeres, telomeres, and repetitive regions fully resolved. The assembly achieves 99.1% BUSCO completeness and encodes 29,280 protein-coding genes, 90.3% of which were functionally annotated. Transposable elements (TEs) dominate 38.38% of the genome, predominantly long terminal repeat retrotransposons (31.14%). This T2T resource resolves long-standing gaps and misassemblies in prior references, providing a foundation for advancing functional genomics, elucidating abiotic stress tolerance, and accelerating molecular breeding in *L. purpureus*. By establishing *L. purpureus* as a model for orphan crop genomics, this work bridges critical gaps in legume biology and sustainable crop development.

Methods

Sample collection, library construction and sequencing. The plant material used in this study, *Lablab purpureus* cv. Linghu, was cultivated at the feild of Yangdu Innovation Base of Zhejiang Academy of Agricultural Sciences (120.42°E, 30.44°N). Fresh young leaves were collected for the preparation of high-fidelity (HiFi), ultra-long Oxford Nanopore Technologies (ONT), and Hi-C sequencing libraries.

High-molecular-weight genomic DNA (gDNA) was isolated and used to construct SMRTbell libraries with the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, CA, USA). Sequencing was performed on the PacBio Sequel IIe platform to generate HiFi reads. An ultra-long DNA library was prepared using the Ultra-Long Sequencing Kit (Oxford Nanopore Technologies, Oxford, UK) and sequenced on the PromethION platform to obtain ultra-long reads, enabling the resolution of complex genomic regions.

For Hi-C scaffolding, fresh leaves were cross-linked with 2% formaldehyde, flash-frozen in liquid nitrogen, and mechanically pulverized. The chromatin was digested with DpnII restriction enzyme (New England Biolabs, USA), and proximity-ligated DNA fragments were processed into Hi-C libraries. These libraries were sequenced on the DNBSEQ-T7 platform (MGI Tech, Shenzhen, China) to generate paired-end reads for chromatin interaction-based scaffolding.

To supplement assembly polishing, short-read genomic libraries were prepared and sequenced on the DNBSEQ-T7 platform. For transcriptomic annotation, total RNA was extracted from five tissues: leaf, root, stem, flower, and pod. The RNA-seq libraries were constructed for each tissue and sequenced on the DNBSEQ-T7 platform. Additionally, equal quantities of RNA from all five tissues were pooled to construct an Iso-Seq library, which was sequenced on the PacBio Sequel II platform to generate full-length transcript isoforms.

The sequencing efforts yielded 33.3 Gb (~69 × coverage) of PacBio HiFi reads (N50: 16 kb), 50.1 Gb (~111 × coverage) of ultra-long ONT reads (N50: 100 kb), and 65.2 Gb (~141 × coverage) of Hi-C paired-end data. Transcriptomic sequencing produced 66.9 Gb of tissue-specific RNA-seq data and 45.5 Gb of PacBio subreads, which were processed into 650 Mb of high-quality circular consensus sequencing (CCS) reads for isoform validation. We performed a genome survey by analyzing short-read genomic sequences from our L. purpureus sample with KMC and GenomeScope2[22,23]. This analysis estimated a genome size of approximately 463 Mb and a heterozygosity of 0.364% (Fig. 1f).

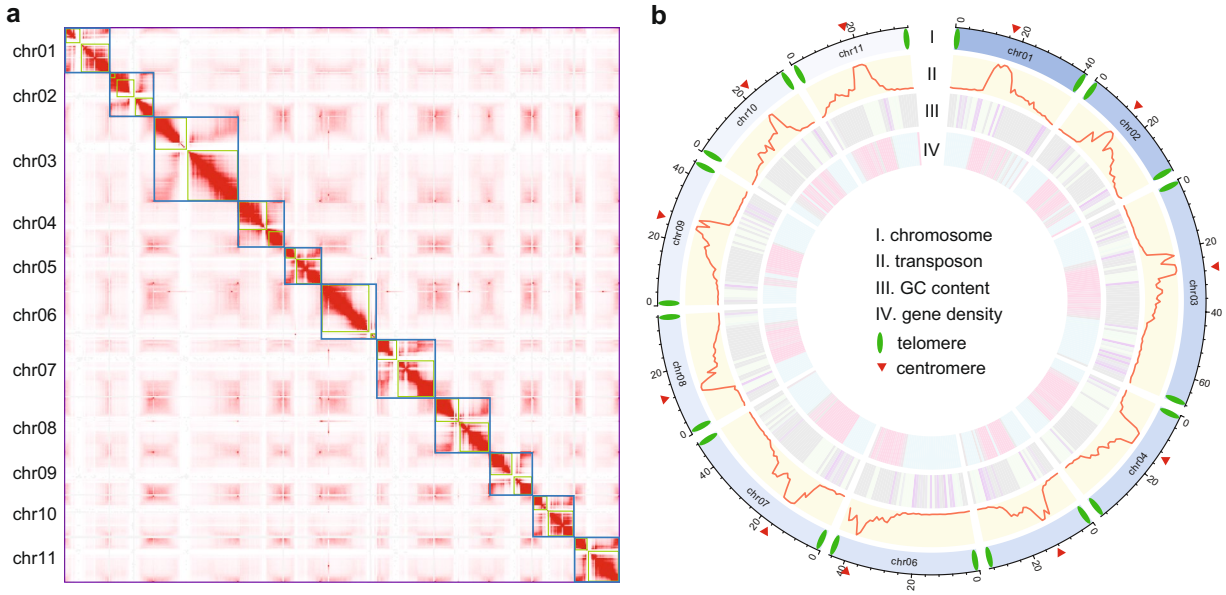

Fig. 2 Telomere-to-telomere (T2T) genome assembly and annotation of *Lablab purpureus* cv. Linghu. (**a**) Hi-C interaction heatmap. (**b**) Circos plot illustrating genome architecture. From outer to inner rings: (I) chromosome lengths, (II) transposable element density, (III) GC content, and (IV) protein-coding gene density. Centromeres (red rectangles) and telomeres (green circles) are annotated on all 11 chromosomes.

Assembly step	Total length (bp)	Number of contig/scaffold	Longest scaffold (bp)	N50 (bp)	L50	N90 (bp)	L90
HifiAsm	484207719	292	68475117	40295122	5	31536392	11
NextDenovo	442188036	57	40476310	19446080	9	9635057	21
Scaffolding	484210519	295	68475317	37964303	5	30389494	11
Gap filling	460985757	11	68475117	41320374	5	33944400	10
NextPolish	460764246	11	68461759	41296257	5	33898525	10

Table 1. Statistics of the *Lablab purpureus* cv. Linghu genome assembly and scaffolding.

Telomere-to telomere genome assembly. The *L. purpureus* cv. Linghu genome was assembled into a telomere-to-telomere (T2T) sequence through a four-stage process: (1) draft genome assembly, (2) Hi-C-assisted scaffolding, (3) gap filling and telomere resolution, and (4) error correction and polishing (Fig. 2a,b; Table 1).

To generate a high-quality initial assembly, PacBio HiFi reads, ultra-long Oxford Nanopore Technologies (ONT) reads, and Hi-C data were co-assembled using Hifiasm (v0.19.9-r616) with default parameters[24]. This produced a primary contig set spanning ~484 Mb (292 contigs, contig N50: 40.3 Mb) (Table 1). A secondary contig set was independently generated using the pipeline of NextDenovo (v2.5.2)[25] and NextPolish (v1.4.1)[26], yielding a ~442 Mb assembly (57 contigs, contig N50: 19.4 Mb) (Table 1).

Hi-C reads were aligned to the primary contig set using Chromap (v0.2.6-r490)[27], and chromosomal-scale scaffolds were constructed with YaHS (v1.2a.1)[28]. A Hi-C contact matrix was generated with Juicer (v1.1)[29], followed by manual corrections in Juicerbox (v1.11.08)[30] to resolve misjoins and optimize scaffold orientation. This step anchored 295 contigs (N50: 38.0 Mb) into 11 pseudochromosomes (total length: 460,762,660 bp), achieving near-chromosomal completeness (Fig. 2a; Table 1).

To close residual gaps and resolve telomeres, the secondary contig set, resulting from NextDenovo/NextPolish[25,26], was integrated into the chromosomal assembly using the quarTeT toolkit (v1.2.0)[31]. Ultra-long ONT reads were aligned to the genome with minimap2 (v2.2.6)[32] to identify and patch unresolved regions, resulting in a refined assembly of ~461 Mb across 11 chromosomes (Fig. 2a; Table 1). Finally, we obtained a complete genome of *L. purpureus* that all gaps have been closed.

The genome was polished iteratively using NextPolish (v1.4.1)[26] with PacBio HiFi reads and genomic short-read data to correct base-level errors. This final step resolved small indels and mismatches, producing a complete, contiguous T2T genome (Fig. 2b; Table 1).

Identification of centromeres and telomeres. Centromeric and telomeric regions in the *L. purpureus* cv. Linghu genome were systematically identified using the quarTeT toolkit (v1.2.0) with the "-c plant" option[31]. The TeloExplorer module detected telomeres by scanning for the plant telomeric repeat motif "AAACCCT/AGGGTTT" (complementary strand). A total of 22 telomeres were identified, corresponding to both ends of the 11 chromosomes (Fig. 2b; Table 2). Each telomeric region contained an average of 989 tandem repeats, confirming terminal completeness and validating the telomere-to-telomere (T2T) assembly (Fig. 2b; Table 2).

Chromosome	Length (bp)	Telomere repeat number		Centromere boundary		
		Left	Right	Start	End	TR coverage
chr01	41296257	261	344	17078432	19489564	91.17%
chr02	35129353	476	94	13003447	18430950	72.88%
chr03	68461759	126	513	25129572	29279532	85.55%
chr04	37989361	243	3653	9625967	16599331	79.89%
chr05	33410887	188	404	11808340	15038905	70.08%
chr06	44927925	233	118	37972245	40627976	58.91%
chr07	48462640	9157	256	14100758	19907194	83.22%
chr08	36462010	229	115	10527874	14594931	77.88%
chr09	44718085	103	249	28004045	29420091	89.55%
chr10	33898525	2507	1858	21426455	24173786	81.38%
chr11	36007444	119	515	15815077	22334675	77.02%

Table 2. Statistics of centromeres and telomeres for the *Lablab purpureus* cv. Linghu genome.

Class	Count	Masked (bp)	Masked (%)
LINE	—	—	—
L1	2876	1257680	0.27%
RTE	819	232608	0.05%
LTR	—	—	—
Copia	59950	64451349	13.99%
Gypsy	34013	77314187	16.78%
unknown	4818	1696874	0.37%
SINE	—	—	—
unknown	460	208064	0.05%
TIR	—	—	—
CACTA	19518	9913048	2.15%
Mutator	17124	6787584	1.47%
PIF_Harbinger	2483	697819	0.15%
Tc1_Mariner	4453	932001	0.20%
hAT	15387	5780472	1.25%
nonTIR	—	—	—
helitron	19657	6934107	1.50%
repeat_fragment	3011	642925	0.14%
Total	184569	176848718	38.38%

Table 3. Information on the repeat annotation in the *Lablab purpureus* cv. Linghu genome assembly.

The CentroMiner module predicted centromeric regions by analyzing chromosome-scale composition of tandem repeat, leading to all the 11 centromeres ranging from 1.42 Mb to 6.97 Mb (Fig. 2b; Table 2). These regions displayed characteristic satellite repeats and retrotransposon enrichment, further supporting their repetitive annotation and assembly continuity.

Annotation of repetitive sequences. Repetitive elements in the *L. purpureus* cv. Linghu genome were annotated using the EDTA pipeline (v2.2.1) using default settings[33]. Three distinct methodologies, including homology-based, structural-based, and *de novo* approaches, were integrated in this pipeline. Our analysis uncovered a total of 176.85 Mb of repetitive sequences, accounting for 38.38% of the T2T genome assembly (Fig. 2b; Table 3). Specifically, the long terminal repeat retrotransposons (LTR-RTs) were found to be the dominant class, representing 31.14% of the genome sequence (Table 3).

Protein-coding gene prediction and functional annotation. Prior to gene prediction, repetitive sequences in the *L. purpureus* cv. Linghu T2T genome were masked using RepeatMasker (v4.1.5)[34] with the EDTA-derived repeat library to reduce false-positive gene annotation. For comprehensive protein-coding gene annotation, we implemented an integrated approach combining three complementary methodologies: homology-based prediction, transcriptome-based evidence, and ab initio gene prediction.

For homology-based prediction, protein sequences from six closely related and well-annotated plant genomes (*Arabidopsis thaliana* Araport11, *Glycine max* Wm82, *Mimulus guttatus* TOLv5.0, *Phaseolus vulgaris* GCA_000499845.2, *Solanum lycopersicum* ITAG5.0, and *Vigna unguiculata* GCA_003958685.2) were collected from Phytozome and NCBI repositories. The homology-driven gene modeling was predicted using GeMoMa (v1.9) with standard parameters[35].

Class	Statistics
Final assembly size (bp)	460764246
Gap number	0
GC content (%)	30.06
Gene number	29280
Average CDS length (bp)	1191.44
Gene annotated by KEGG	12404
Gene annotated by GO	13719
Annotated proportion (%)	90.30

Table 4. Statistics of genome assembly and annotation.

For transcriptome-based evidence, we incorporated by short-read and full-length RNA-seq datasets. The short-read RNA-seq data of five tissues (leaf, root, stem, flower, pod) were aligned to the genome using Hisat2 (v2.2.1) with default parameters[36]. The transcript isoforms subsequently reconstructed through TACO (v0.7.3)[37]. For PacBio Iso-Seq data, we followed the Iso-Seq3 pipeline including circular consensus sequence (CCS) generation, transcript clustering, and polishing. Putative coding sequences were then extracted using TransDecoder (v5.7.0) (https://github.com/TransDecoder/TransDecoder).

For *de novo* gene prediction, we employed Braker2 (v2.1.2)[38] which integrates Augustus (v3.3.2)[39] and GeneMark-ET (v4.59)[40]. The pipeline was trained using RNA-seq alignments (BAM files from Hisat2 alignments) and protein sequences from Swiss-Prot database to enhance prediction accuracy. GeneMark-ET was first used to generate initial gene predictions which were then used to train Augustus. The trained Augustus model was subsequently applied for comprehensive gene prediction.

Finally, predictions from these three strategies were integrated using EVidenceModeler (v2.0.0)[41] with evidence weights assigned as follows: transcript evidence (weight: 8), protein homology evidence (weight: 6), and *ab initio* predictions (weight: 6). This weighting scheme prioritized direct experimental evidence while incorporating supportive computational predictions. The integrated gene models were then filtered to remove truncated genes and those without coding potential, resulting in a unified, high-confidence annotation set for downstream analysis. Finally, we obtained a unified, high-confidence annotation set. Predicted genes were functionally characterized using eggNOG-mapper (v2.1.12)[42] against the eggNOG database (v5.0.2)[43].

Finally, we identified a total of 29,280 protein-coding genes, of which 26,439 (90.3%) were functionally annotated by eggNOG framework (Fig. 2b; Table 4). The annotation demonstrated 13,719 (46.9%) gene products associated with Gene Ontology (GO) descriptors and 12,404 (42.4%) entities mapped to KEGG Orthology (KO) identifiers (Table 4).

Assembly quality. To evaluate the continuity and accuracy of our *L. purpureus* cv. Linghu T2T assembly, we performed whole-genome alignments against two previously published assemblies (cv. Highworth and cv. C204) using SyRI[44]. Our T2T genome presents large-scale genomic mismatches with formerly published assemblies (Fig. 3a,b). This observations have been supported by the D-Genie web service[45], a tool optimized for detecting and visualizing large-scale structural variations (SVs) and conserved syntenic blocks (Fig. 3c). These comparative analysis highlighted large-scale chromosomal rearrangements and alignment discontinuities, which were further investigated in the context of transposable element (TE) dynamics and centromeric organization.

TE annotation was conducted across all three assemblies using the EDTA pipeline (v2.2.1) with default settings[33]. TE density profiles were generated to assess their distribution along chromosomes (Fig. 3c). Breakpoints - regions where genome alignments fragmented - were identified and cross-referenced with local TE density valleys (low-density regions) and peaks (high-density regions), as well as the annotated centromeres of the cv. Linghu assembly (Fig. 3c).

In the cv. Linghu T2T assembly, centromeric regions consistently aligned with TE density peaks, reflecting the inherent repeat-rich nature of functional centromeres (Fig. 3c). Strikingly, in the cv. Highworth and cv. C204 assemblies, four and three chromosomes, respectively, exhibited TE density valleys at putative centromeric regions. These TE-depleted valleys coincided with alignment breakpoints, suggesting mis-assembly between long and short chromosome arms or fragmentation of centromeres in the earlier assemblies (Fig. 3c).

The congruence between TE density peaks, centromere annotations, and unbroken alignments in the cv. Linghu genome underscores its superior continuity compared to previous assemblies. This work resolves critical gaps and mis-assemblies in prior references, particularly in complex centromeric regions where repetitive sequences likely confounded short-read or lower-coverage long-read strategies.

Data Records

All raw sequencing data generated in this study – including PacBio HiFi reads, ultra-long ONT reads, Hi-C data, genomic resequencing data, and tissue-specific RNA-seq (short-read and Iso-Seq) datasets – are publicly available at the NCBI Sequence Read Archive (SRA) with the accession SRP583761[46]. The complete telomere-to-telomere (T2T) genome assembly of *Lablab purpureus* cv. Linghu is deposited in NCBI GenBank under the accession number GCA_051403735.1[47]. Supplementary materials including genome assembly metrics, gene structural and functional annotations, and repetitive element predictions have been archived in the Figshare repository[48].

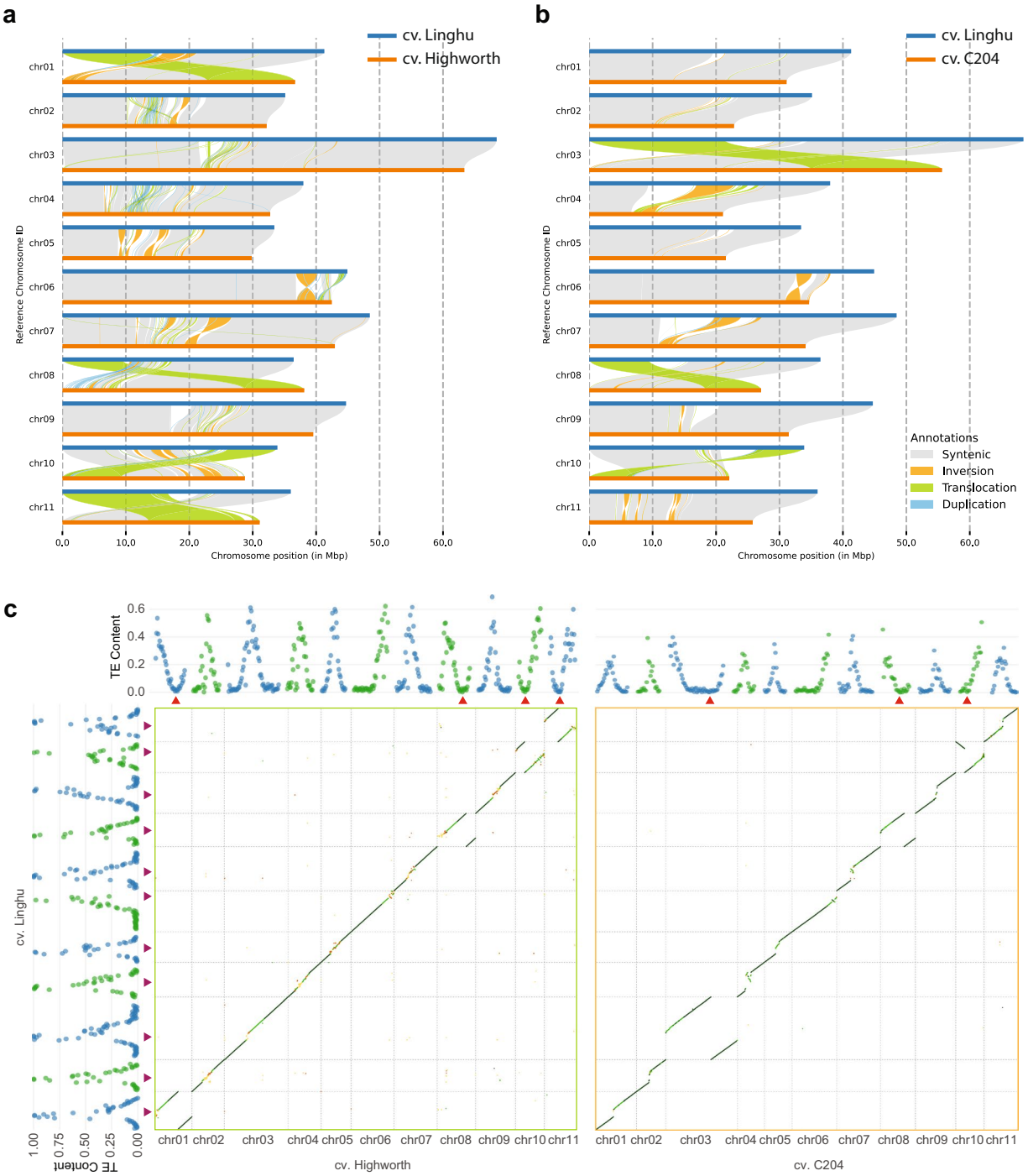

Fig. 3 Comparative genome alignment and structural variation analysis of the cv. Linghu T2T genome. (**a**,**b**) Large-scale structural variations between the cv. Linghu T2T assembly and the previously published cv. Highworth (**a**) and cv. C204 (**b**) genomes using SyRI. (**c**) Whole-genome alignments between the cv. Linghu T2T assembly and two previously published genomes (cv. Highworth and cv. C204) reveal assembly discrepancies. Left panel: TE density profile and centromere positions (purple triangles) for the cv. Linghu genome. Top panels: TE density profiles for cv. Highworth (left) and cv. C204 (right), with red triangles marking alignment breakpoints (structural variation sites) that coincide with TE-depleted valleys. Mis-assembled centromeres in prior genomes correlate with abrupt TE content drops and alignment fragmentation.

Technical Validation

The quality of the T2T genome assembly was rigorously validated using multiple metrics. Merqury (v1.1)[49] assessed base-level accuracy using PacBio HiFi reads, yielding chromosomal quality values (QV) of 37~54 (Fig. 4a). The Merqury analysis also indicated a completeness of 99.48%. The continuity of genome assembly was also evaluated using the LAI indicator, of which an estimation of 19.18 support a good quality of our genome

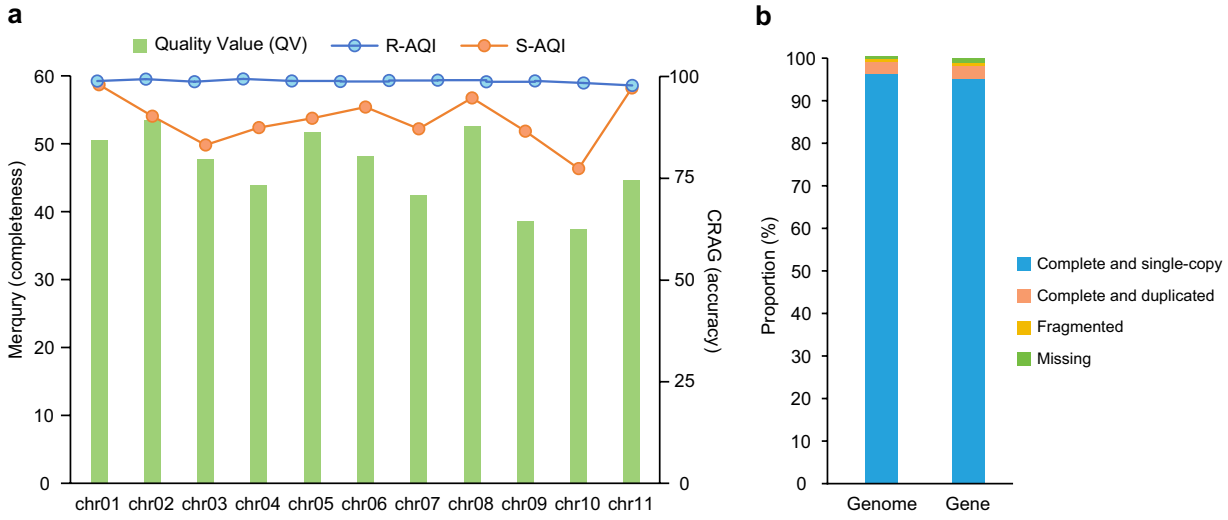

Fig. 4 Quality assessment of the *Lablab purpureus* cv. Linghu T2T genome assembly and annotation. (**a**) The completeness and accuracy of genome assembly. The completeness is qualified using the QV from the Merqury, while the accuracy is evaluated by the R-AQI and S-AQI values from the CRAG. (**b**) BUSCO evaluations from the assembled genome and annotated protein-coding genes.

as a reference genome[50,51]. CRAQ (v1.0.9)[52] further validated assembly integrity, with R-AQI (regional assembly quality indicators) >97 and S-AQI (structural assembly quality indicators) of 77~97 across chromosomes (Fig. 4a). We also evaluated the genome assembly using CRAQ with both short-read and long-read (PacBio HiFi) data. The short-read-based assessment yielded a genome-wide R-AQI of 99 and an S-AQI of 100, whereas the HiFi-based evaluation resulted in a genome-wide R-AQI of 99 and an S-AQI of 90.

The completeness of the genome assembly was further assessed using BUSCO (v5.2.2) with the embryophyta_odb10 orthologous database[53]. BUSCO evaluation of T2T genome assembly revealed a completeness of 99.1%, with 96.3% complete and single copy genes and 2.8% complete and duplicated genes (Fig. 4b). The BUSCO analysis also indicated a completeness of 98.2% for the annotated protein-coding genes (Fig. 4b). These results collectively demonstrate the high accuracy, continuity, and biological utility of the our T2T genome assembly of *L. purpureus* cv. Linghu.

Data availability

The raw sequencing dataset has been deposited to NCBI Sequence Read Archive (SRP583761). The genomic sequence has been deposite to NCBI GenBank (GCA_051403735.1). The annotation results are available at https://doi.org/10.6084/m9.figshare.28953455.v1.

Code availability

All software and pipelines were executed in accordance with the guidelines and protocols outlined by the developers of the respective bioinformatics tools. Software versions and non-default parameters are detailed in the Methods section.

Received: 12 June 2025; Accepted: 26 September 2025;
Published online: 13 November 2025

References

1. Kongjaimun, A. *et al*. Molecular Analysis of Genetic Diversity and Structure of the Lablab (Lablab purpureus (L.) Sweet) Gene Pool Reveals Two Independent Routes of Domestication. *Plants* **12** (2023).
2. Maass, B. L. & Chapman, M. A. in *Underutilised Crop Genomes* (ed Chapman, M. A.) 229–253 (Springer International Publishing, 2022).
3. Maass, B. L. *et al*. Lablab purpureus—A Crop Lost for Africa? *Tropical Plant Biology* **3**, 123–135, https://doi.org/10.1007/s12042-010-9046-1 (2010).
4. Njaci, I. *et al*. Chromosome-level genome assembly and population genomic resource to accelerate orphan crop lablab breeding. *Nature Communications* **14**, 1915, https://doi.org/10.1038/s41467-023-37489-7 (2023).
5. Habib, H. M., Theuri, S. W., Kheadr, E. E. & Mohamed, F. E. Functional, bioactive, biochemical, and physicochemical properties of the Dolichos lablab bean. *Food & Function* **8**, 872–880, https://doi.org/10.1039/C6FO01162D (2017).
6. Minde, J. J., Venkataramana, P. B. & Matemu, A. O. Dolichos Lablab-an underutilized crop with future potentials for food and nutrition security: a review. *Critical Reviews in Food Science and Nutrition* **61**, 2249–2261, https://doi.org/10.1080/10408398.2020.1775173 (2021).
7. Nord, A., Miller, N. R., Mariki, W., Drinkwater, L. & Snapp, S. Investigating the diverse potential of a multi-purpose legume, Lablab purpureus (L.) Sweet, for smallholder production in East Africa. *PLOS ONE* **15**, e0227739, https://doi.org/10.1371/journal.pone.0227739 (2020).
8. Missanga, J. S., Venkataramana, P. B. & Ndakidemi, P. A. Recent developments in genomics: A focus on drought stress tolerance and use of genomic resources to develop stress-resilient varieties. *Legume Science* **3**, e99, https://doi.org/10.1002/leg3.99 (2021).

9. Workneh, S. T., Feyissa, T., Asfaw, Z. & Disasa, T. Genetic Diversity and Population Structure of Lablab (Lablab purpureus L. Sweet) Accessions from Ethiopia Using SSR Markers. *Plant Molecular Biology Reporter* **42**, 688–699, https://doi.org/10.1007/s11105-024-01447-4 (2024).
10. Maass, B. L., Robotham, O. & Chapman, M. A. Evidence for two domestication events of hyacinth bean (Lablab purpureus (L.) Sweet): a comparative analysis of population genetic data. *Genetic Resources and Crop Evolution* **64**, 1221–1230, https://doi.org/10.1007/s10722-016-0431-y (2017).
11. Zhao, Y. *et al*. Nuclear phylotranscriptomics and phylogenomics support numerous polyploidization events and hypotheses for the evolution of rhizobial nitrogen-fixing symbiosis in Fabaceae. *Molecular Plant* **14**, 748–773, https://doi.org/10.1016/j.molp.2021.02.006 (2021).
12. Pootakham, W. *et al*. A de novo chromosome-scale assembly of the Lablab purpureus genome. *Frontiers in Plant Science* **ume 15**, 2024, https://doi.org/10.3389/fpls.2024.1347744 (2024).
13. James, M. E., Brodribb, T., Wright, I. J., Rieseberg, L. H. & Ortiz-Barrientos, D. Replicated Evolution in Plants. *Annual Review of Plant Biology* **74**, 697–725, https://doi.org/10.1146/annurev-arplant-071221-090809 (2023).
14. Kress, W. J. *et al*. Green plant genomes: What we know in an era of rapidly expanding opportunities. *Proceedings of the National Academy of Sciences* **119**, e2115640118, https://doi.org/10.1073/pnas.2115640118 (2022).
15. Jha, U. C., Bohra, A. & Nayyar, H. Advances in "omics" approaches to tackle drought stress in grain legumes. *Plant Breeding* **139**, 1–27, https://doi.org/10.1111/pbr.12761 (2020).
16. Robotham, O. & Chapman, M. Population genetic analysis of hyacinth bean (Lablab purpureus (L.) Sweet, Leguminosae) indicates an East African origin and variation in drought tolerance. *Genetic Resources and Crop Evolution* **64**, 139–148, https://doi.org/10.1007/s10722-015-0339-y (2017).
17. Wang, B. *et al*. Identification of drought-inducible regulatory factors in Lablab purpureus by a comparative genomic approach. *Crop and Pasture Science* **69**, 632–641 (2018).
18. Kumar Rai, K., Rai, N. & Pandey Rai, S. Downregulation of γECS gene affects antioxidant activity and free radical scavenging system during pod development and maturation in Lablab perpureus L. *Biocatalysis and Agricultural Biotechnology* **11**, 192–200, https://doi.org/10.1016/j.bcab.2017.07.005 (2017).
19. Hendre, P. S. *et al*. African Orphan Crops Consortium (AOCC): status of developing genomic resources for African orphan crops. *Planta* **250**, 989–1003, https://doi.org/10.1007/s00425-019-03156-9 (2019).
20. Jamnadass, R. *et al*. Enhancing African orphan crops with genomics. *Nature Genetics* **52**, 356–360, https://doi.org/10.1038/s41588-020-0601-x (2020).
21. Chang, Y. *et al*. The draft genomes of five agriculturally important African orphan crops. *GigaScience* **8**, giy152, https://doi.org/10.1093/gigascience/giy152 (2019).
22. Kokot, M., Długosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759–2761, https://doi.org/10.1093/bioinformatics/btx304 (2017).
23. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* **11**, 1432, https://doi.org/10.1038/s41467-020-14998-3 (2020).
24. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
25. Hu, J. *et al*. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biology* **25**, 107, https://doi.org/10.1186/s13059-024-03252-4 (2024).
26. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255, https://doi.org/10.1093/bioinformatics/btz891 (2020).
27. Zhang, H. *et al*. Fast alignment and preprocessing of chromatin profiles with Chromap. *Nature Communications* **12**, 6566, https://doi.org/10.1038/s41467-021-26865-w (2021).
28. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, btac808, https://doi.org/10.1093/bioinformatics/btac808 (2023).
29. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3**, 95–98, https://doi.org/10.1016/j.cels.2016.07.002 (2016).
30. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems* **3**, 99–101, https://doi.org/10.1016/j.cels.2015.07.012 (2016).
31. Lin, Y. *et al*. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Horticulture Research* **10**, uhad127, https://doi.org/10.1093/hr/uhad127 (2023).
32. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
33. Ou, S. *et al*. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology* **20**, 275, https://doi.org/10.1186/s13059-019-1905-y (2019).
34. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Current Protocols in Bioinformatics* **25**, 4.10.11–14.10.14, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
35. Keilwagen, J., Hartung, F. & Grau, J. in *Gene Prediction: Methods and Protocols* (ed Kollmar, M.) 161–177 (Springer New York, 2019).
36. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37**, 907–915, https://doi.org/10.1038/s41587-019-0201-4 (2019).
37. Niknafs, Y. S., Pandian, B., Iyer, H. K., Chinnaiyan, A. M. & Iyer, M. K. TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nature Methods* **14**, 68–70, https://doi.org/10.1038/nmeth.4078 (2017).
38. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics* **3**, lqaa108, https://doi.org/10.1093/nargab/lqaa108 (2021).
39. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**, W435–W439, https://doi.org/10.1093/nar/gkl200 (2006).
40. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes. *Genome Research* **34**, 757–768, https://doi.org/10.1101/gr.278373.123 (2024).
41. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**, R7, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
42. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution* **38**, 5825–5829, https://doi.org/10.1093/molbev/msab293 (2021).
43. Huerta-Cepas, J. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* **47**, D309–D314, https://doi.org/10.1093/nar/gky1085 (2019).
44. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology* **20**, 277, https://doi.org/10.1186/s13059-019-1911-0 (2019).
45. Cabanettes, F. & Klopp, C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**, e4958, https://doi.org/10.7717/peerj.4958 (2018).
46. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP583761 (2025).

47. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_051403735.1 (2025).
48. Liu, H. The complete telomere-to-telomere genome assembly of Lablab purpureus (L.) Sweet. *figshare. Dataset.* https://doi.org/10.6084/m9.figshare.28953455.v1 (2025).
49. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**, 245, https://doi.org/10.1186/s13059-020-02134-9 (2020).
50. Mokhtar, M. M., Abd-Elhalim, H. M. & El Allali, A. A large-scale assessment of the quality of plant genome assemblies using the LTR assembly index. *AoB PLANTS* **15**, plad015, https://doi.org/10.1093/aobpla/plad015 (2023).
51. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Research* **46**, e126–e126, https://doi.org/10.1093/nar/gky730 (2018).
52. Li, K., Xu, P., Wang, J., Yi, X. & Jiao, Y. Identification of errors in draft genome assemblies at single-nucleotide resolution for quality assessment and improvement. *Nature Communications* **14**, 6556, https://doi.org/10.1038/s41467-023-42336-w (2023).
53. Waterhouse, R. M. *et al.* BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution* **35**, 543–548, https://doi.org/10.1093/molbev/msx319 (2018).

Acknowledgements

This work was supported by the Precision Identification Project for Agricultural Germplasm Resources by Zhejiang Provincial Department of Agriculture and Rural Affairs (2023JZJD002).

Author contributions

Heyun Chen and Qiaojun Lou conceived and designed the study. Heqin Liu and Xiuhui Liu prepared samples, and wrote the draft manuscript. Haomin Lyu analyze the data and submitted the genome data. Guihua Zou revised the manuscript. Xueqiang Zheng performed experiments. Tao Zheng supervise this project. All authors critically reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.L., Q.L. or H.C.