



OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of the vegetable leafminer *Liriomyza sativae* (Diptera: Agromyzidae)

Xulong Chen^{1,5}, Xiaodong Cai^{2,5}, Jiuzhou Liu², Shuangmei Ding³, Zheng Fan⁴, Zhuo Chen¹ & Ding Yang^{1,2}✉

So far, the chromosome-level genome assembly of only one species has been known for the family Agromyzidae. The vegetable leafminer, *Liriomyza sativae* Blanchard, 1938, is a highly polyphagous invasive pest that poses a significant threat to a wide range of vegetable crops. However, genomic resources for this species remain unavailable, impeding our understanding of its invasive traits and extensive adaptability. In this study, we present the high-quality chromosome-level genome assembly of *L. sativae*, using MGI short-read, PacBio HiFi long-read and Hi-C sequencing technology. The final assembly spans 110.98 Mb, of which 92.08% (102.21 Mb) was anchored to five pseudo-chromosomes, yielding a contig N50 of 1.16 Mb and 94.9% BUSCO completeness. The genome contains 30.39 Mb of repetitive sequences (accounting for 27.38% of the total) and encodes 10,312 predicted protein-coding genes, of which 9,978 (96.76%) are functionally annotated. The high-quality genome assembly provides a critical foundation for elucidating the adaptive evolution and invasion mechanisms of *Liriomyza* species.

Background & Summary

The genus *Liriomyza* Mik, 1894, a highly diversified taxon within the family Agromyzidae, comprises 464 described species distributed across all major zoogeographic regions¹. It holds the significant economic importance in agricultural ecosystems due to the unique larval feeding behavior—tunneling through mesophyll tissues to create serpentine mines that severely impair photosynthesis. Over 20 species are recognized as economically detrimental pests, affecting both agricultural crops (e.g., Solanaceae, Cucurbitaceae) and ornamental plants (e.g., Asteraceae, Rosaceae)². Among these, three invasive species, *L. sativae* Blanchard, 1938, *L. huidobrensis* (Blanchard, 1926), and *L. trifolii* (Burgess, 1880), have successively invaded China, resulting in substantial economic losses to a variety of vegetable and floriculture crops^{3,4}.

The vegetable leafminer fly *Liriomyza sativae* (Fig. 1) is characterized by its small size, high reproductive capacity, highly polyphagous, and strong ecological adaptability, making it an important international quarantine pest^{3,5,6}. As a thermophilic species with a short generation cycle, it thrives ranging from tropical to temperate zones, exploiting over 60 host plant species spanning 18 different families⁷. Originating from the Americas, *L. sativae* has rapidly spread globally over the past few decades, facilitated by international trade and its remarkable adaptability to diverse agroecosystems^{8–10}. Currently, it is the most widely distributed leafminer pest in China, having rapidly colonized 95% of vegetable-growing regions since its initial detection in Hainan Province in 1993. Larval mining behavior directly damages the foliar tissues of economically important crops, including tomatoes, cucumbers and beans^{11–14}, reducing the photosynthetic area by 40–70% in severe infestations (Fig. 1) and predisposing plants to secondary pathogens. Meanwhile, female adults puncture the leaves for feeding or

¹State Key Laboratory of Green Pesticides, Guizhou University, Guiyang, Guizhou, 550025, China. ²State Key Laboratory of Agricultural and Forestry Biosecurity, MARA Key Lab of Surveillance and Management for Plant Quarantine Pests, College of Plant Protection, China Agricultural University, Beijing, 100193, China. ³The Institute of Scientific and Technical Research on Archives, NAAC, Beijing, 100053, China. ⁴Key Laboratory of Eco-Environments in Three Gorges Reservoir Region (Ministry of Education), School of Life Sciences, Southwest University, Chongqing, 400715, China. ⁵These authors contributed equally: Xulong Chen, Xiaodong Cai. ✉e-mail: dyangcau@126.com

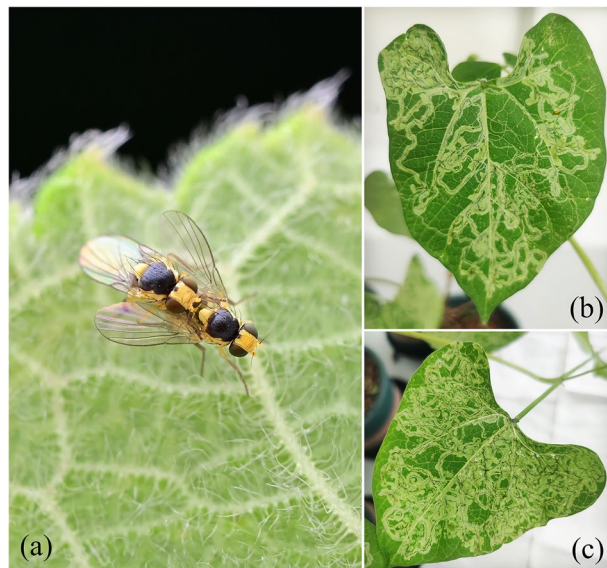


Fig. 1 Live adults of *Liriomyza sativae* and their infestation symptoms. **(a)** Copulating pair of *L. sativae*. **(b,c)** Infestation symptoms of *L. sativae* on common bean.

oviposition, which accelerates water loss and nutrient depletion. Conventional insecticide-based control strategies face increasing challenges due to the species' propensity to develop resistance, underscoring the urgent need for genomic resources to decipher its adaptive mechanisms.

Prior to this study, the chromosome-level genome assembly of only one species (*L. trifolii*) had been published for the family Agromyzidae¹⁵. Although these leafminers occupy similar ecological niches, *L. sativae* invaded China substantially earlier than *L. trifolii* (1993 versus 2005). Demonstrating remarkable invasion velocity, *L. sativae* had colonized 21 of China's 34 provincial-level administrative regions by 1995, impacting approximately 1.488 million hectares of agricultural land. The high-quality chromosome-level genome of *L. sativae* presented in this study provides critical insights into the genomic basis of its invasion success. The final genome assembly spans 110.98 Mb, with 102.21 Mb (92.08% of the total) anchored to five pseudo-chromosomes. The assembly exhibits contig and scaffold N50 values of 1.16 Mb and 18.57 Mb, respectively, and achieves a Benchmarking Universal Single-Copy Orthologs (BUSCO) completeness score of 94.9%. This chromosome-level genome assembly establishes a foundational genomic platform for elucidating the molecular basis of *Liriomyza*'s polyphagous adaptations, developing RNA interference (RNAi)-based management strategies, and deciphering the molecular mechanisms driving their invasive success across agroecosystems.

Methods

Sample collection and rearing. The lab-reared colonies of *Liriomyza sativae* used in this study were originally collected from the East Campus of Guizhou University, Guiyang City, Guizhou Province, China (26°26'48"N, 106°40'0"E) in 2024. Subsequently, these leafminers were maintained on common bean plants (*Phaseolus vulgaris*) for six generations under controlled conditions of $26 \pm 1^\circ\text{C}$, a photoperiod of 16 h light and 8 h darkness, and a relative humidity of $60 \pm 5\%$. Inbred female and male adults were starved for 6 h, rapidly frozen in liquid nitrogen, and stored at -80°C until further processing.

Genome sequencing. Whole genomic DNA was extracted from 600 adult male individuals using the Sodium Dodecyl Sulfate (SDS) method, with specific sample allocations as follows: (1) 265 individuals for MGI short-read sequencing, (2) 200 individuals for PacBio long-read sequencing, and (3) 135 individuals for Hi-C library construction, followed by purification with the Monarch Genomic DNA Purification Kit T3010 (New England Biolabs, USA). The integrity and quality of the extracted genomic DNA were assessed using 0.7% agarose gel electrophoresis, a NanoDrop One Spectrophotometer (NanoDrop Technologies, Wilmington, DE) and a Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA).

The MGI paired-end sequencing library with an insert size of 350 bp was prepared using the MGI Plus DNA Library Prep Kit and subsequently sequenced on the DNBSEQ T7 platform (MGI). A total of 23.74 Gb of MGI short reads were generated. The raw reads were filtered using the software fastp v0.21.0¹⁶ (parameter: `-detect_adapter_for_pe`), with low-quality reads removed according to the following criteria: 1) removal of reads containing $> 5\%$ ambiguous bases (N); 2) exclusion of reads with $\geq 50\%$ low-quality bases (Phred score ≤ 5); 3) elimination of adapter-contaminated reads; and 4) deduplication of PCR duplicates caused by amplification bias. After filtering, 23.61 Gb of clean reads (Table 1) were retained for subsequent genome survey analysis.

An SMRT bell sequencing library containing approximately 15 kb fragments was constructed using the Pacific Biosciences SMRTbell Express Template Prep Kit 2.0 and sequenced on the PacBio Revio platform. The polymerase reads obtained from sequencing were processed to yield subreads containing only insert fragments as follows: 1) removal of reads shorter than 50 bp; 2) exclusion of reads with a quality value < 0.8 ; 3) elimination

Library	Platform	Usage	Clean data (Gb)	Coverage (×)
Short reads	MGI DNBSEQ T7	Genome survey	23.61	212.74
Long reads	PacBio Revio	Assembly	20.81	187.51
Hi-C	MGI DNBSEQ T7	Hi-C assembly	51.65	465.40
RNA-sr	MGI DNBSEQ T7	Annotation	13.08	—
RNA-ONT	Oxford Nanopore PromethION	Annotation	11.22	—

Table 1. Library sequencing data for genome assembly and annotation of *L. sativae*.

Item	Value
Genome size (Mb)	110.98
Number of psedo-chromosomes	5
Contig N50 (Mb)	1.16
Number of contigs	529
Scaffold N50 (Mb)	18.57
Number of scaffolds	338
GC content (%)	31.43

Table 2. Summary statistics of the genome assembly of *L. sativae*.

of reads containing self-ligated adapters; and 4) trimming of adapter sequences. The subreads were then processed with SMRTLink v13.0 (<https://www.pacb.com/smrt-link/>) to generate HiFi reads. In total, 20.81 Gb of HiFi reads with an N50 of 20,035 bp were obtained (Table 1).

Hi-C (High-throughput Chromosome Conformation Capture) technology was employed to achieve chromosomal-scale scaffolding of the genome assembly while resolving the three-dimensional chromatin interaction landscape through genome-wide pairwise contact analysis. A Hi-C library was constructed following the standard protocol and sequenced on the MGI DNBSEQ-T7 platform using the PE150 module to generate paired-end reads. A total of 345,243,968 Hi-C raw reads (51.79 Gb) were generated. Raw data were processed with fastp v0.23.2¹⁶ to remove adapter sequences and low-quality reads. Ultimately, 51.65 Gb of Hi-C clean data (Table 1) was retained for the chromosome assembly.

Transcriptome sequencing. Total RNA was extracted from 70 adult female individuals using the Trizol reagent. The quality and concentration of the RNA were determined through 1% agarose gel electrophoresis, a NanoDrop One Spectrophotometer (NanoDrop Technologies, Wilmington, DE) and a Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA). Second-generation and third-generation sequencing libraries were constructed and sequenced on the DNBSEQ T7 platform (MGI Tech) and the Oxford Nanopore PromethION platform (Oxford Nanopore Technologies), respectively. The second-generation sequencing reads and Oxford Nanopore Technologies (ONT) reads were processed using fastp v0.23.2¹⁶ (parameter: -L -w 16) and NanoFilt v2.8.0¹⁷ (parameter: -l 50 -q 7), respectively, to filter out adapter sequences and low-quality reads for subsequent genome annotation analyses.

Genome size estimation. To evaluate the genome size of *L. sativae*, we performed a k-mer analysis following these steps: First, high-quality short reads were preprocessed and subsequently analyzed using Jellyfish v2.2.10¹⁸ (parameter: count -C -m 19 -s 1 G -g generators -G 4/stats) with a k-mer size of 19. The resulting k-mer frequency distribution was then analyzed with GenomeScope v2.0¹⁹ (parameter: -i histo_all -k kmer_size -o./ genomescope -p ploidy). Based on this analysis, the genome size of *L. sativae* was estimated to be approximately 100 Mb. This predicted genome size is comparable to that of the congeneric species *L. trifolii* (genome size: 122.64 Mb). Furthermore, genome complexity was assessed by analyzing heterozygosity rate (1.29%), repeat content (27.50%), and GC content (31.43%). These genomic features provide a critical foundation for future in-depth studies on *L. sativae*, including functional annotation and comparative genomic analyses.

Genome assembly. The PacBio HiFi long-read data were used to generate a contig-level assembly of the *L. sativae* genome. Using the default parameters of WTDBG2 v0.0²⁰, a preliminary genome assembly was produced, resulting in 529 contigs with a total length of 110.98 Mb and a contig N50 of 1.16 Mb (Table 2). Subsequently, Hi-C data were employed to assemble the contigs at the chromosome level. Initially, raw Hi-C data were trimmed and filtered using fastp v0.23.2¹⁶ (parameter: -detect_adapter_for_pe) to obtain high-quality reads. Clean reads were then aligned to the draft genome using BWA v0.7.17-r1188²¹ with default parameters. The valid interaction data were extracted for chromatin interaction analysis. Scaffolding tools AllHiC v0.9.8²² (parameter: -e GATC), 3D-DNA v201008²³ (parameter: -q 30) and Juicer v1.6²⁴ (parameter: -g matrial -s MboI -t 30 -S early) were applied to cluster, order, and orient contigs. Manual adjustments were performed in Juicebox v1.11.08²⁵ (parameter: Coverage) to finalize contig assignments. Finally, chromosome-level scaffolds were constructed by integrating the contig-level genome sequences, with gaps between contigs within chromosomes filled with 100 consecutive Ns. In this process, a total of five psedo-chromosomes were successfully anchored (Fig. 2), with lengths ranging from 16.19 Mb to 25.09 Mb (Fig. 3, Table 3). Remarkably, a high Hi-C mapping rate of 92.08% was attained, along

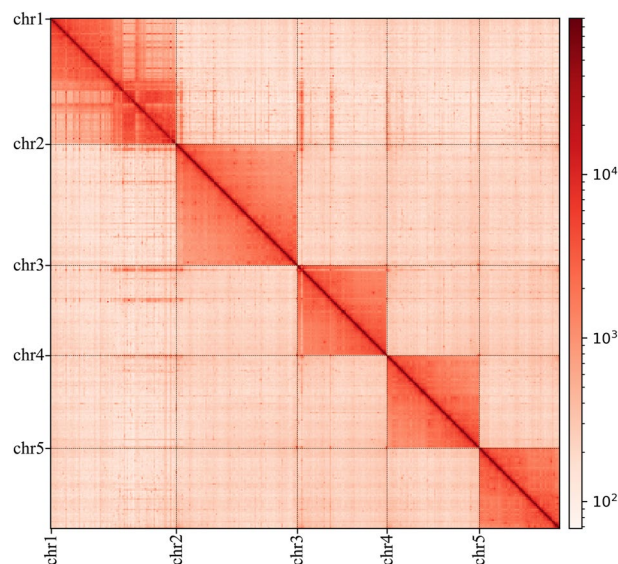


Fig. 2 Hi-C heatmap of *L. sativae* genome assembly.

with a scaffold N50 value of 18.57 Mb (Table 2). Hi-C interaction patterns were visualized using HiCExplorer v3.7.5²⁶ (parameter: -dpi 300 -log1p -colorMap Reds -clearMaskedBins -rotationX 45 -vMin 11), with signal intensity displayed in Fig. 2.

Genome annotation. Repetitive sequences, comprising transposable elements (TEs) and tandem repeat sequences, were systematically characterized. Tandem repeat sequences were predicted using TRF v4.09²⁷ (parameter: 2 7 7 80 10 50 2000 -d) and MISA v2.1²⁸ (parameter: default). Two methods, *de novo* assembly and homology-based prediction, were employed to identify TEs. First, *de novo* annotation was carried out using a range of conventional softwares, including RepeatModeler v2.0.6²⁹ (parameter: -database mydb -threads 16), LTR_FINDER (Official release of LTR_FINDER_parallel, parameter: -threads 16 -harvest_out -size 1000000 -time 300)³⁰, LTRharvest v1.6.5³¹ (parameter: -minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1 -similar 85 -vic 10 -seed 20 -seqids yes), and LTR_retriever v3.0.1³² (parameter: -threads 16 -noanno), to generate a *de novo* repetitive sequence library. Then, the RepBase library was merged with the *de novo* library, and RepeatMasker v4.1.7³³ (parameter: -noLowSimple -pvalue 0.0001) was used to align and predict repetitive sequences, yielding the “De novo + RepBase” results. Finally, RepeatProteinMask v4.1.7³³ (parameter: -noLowSimple -pvalue 0.0001) was used to predict repetitive sequences of the TE_protein type, receiving the “TE proteins” results. The analysis revealed a total of 30.39 Mb of repetitive sequences, accounting for 27.38% of the genome, which comprises 7.21% long terminal repeats (LTRs), 0.20% short interspersed nuclear elements (SINES), 3.77% long interspersed nuclear elements (LINEs), and 15.54% DNA elements (Table 4).

Non-coding RNAs (ncRNAs) with essential biological functions, including transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs) involved in protein synthesis, were systematically annotated. To identify tRNA sequences in the genome, we used tRNAscan-SE v2.0.12³⁴ with default parameters, based on the structural characteristics of tRNAs. For rRNA prediction, we employed RNAmmer v1.2³⁵ (parameter: -S euk -m tsu,lsu,ssu). Additionally, we used INFERNAL v1.1.4³⁶ (parameter: -cut_ga -rfam -nohmmonly -fmt 2), referencing the Rfam database, to detect other ncRNA sequences in the genome, such as small nuclear RNAs (snRNAs) and microRNAs (miRNAs).

A three-tiered strategy integrating transcriptomic, ab initio, and homology-based approaches was implemented to annotate the protein-coding genes in the *L. sativae* genome. For transcriptome-based gene prediction, the second-generation sequencing reads were aligned using Hisat2 v2.2.1³⁷ (parameter: -dta -p 16) and assembled with StringTie v2.2.1³⁸ (parameter: -p 16 -R -L). The third-generation sequencing reads were aligned to the genome assembly by Minimap2 v2.26-r1175³⁹ (parameter: -t 16 -ax splice -uf -secondary=no), and the aligned sequences were subsequently assembled into transcripts with StringTie v2.2.1³⁸ (parameter: -p 16 -R -L). The predicted transcripts were merged using Tama v1.0⁴⁰ (parameter: -f filelist.txt -p merge), and then candidate coding regions within the transcript sequences were identified using the default parameters of TransDecoder v5.7.1 (<https://github.com/TransDecoder/TransDecoder>). For ab initio gene prediction, Augustus v3.5.0⁴¹ (parameter: -uniqueGenes = true -noInFrameStop = true -gff3 = on -strand = both) was trained with RNA-Seq-guided hints and ortholog evidence, and GenScan v1.0⁴² (parameter: default) predicted additional gene models. For homology-based gene prediction, MiniProt v0.13⁴³ (parameter: -gff -lut50) aligned *L. trifolii* proteomes to the genome, retaining high-confidence orthologs.

Overall, we predicted a total of 10,312 protein-coding genes. Genome assembly completeness was assessed using BUSCO v5.7.0⁴⁴ (database: insecta_odb10; n = 1,367 orthologs), yielding a completeness score of 90.5% according to the Benchmarking Universal Single-Copy Orthologues (BUSCO) standard (Table 5). The average length of gene, CDS, exon and intron sequences was 5,561.81 bp, 864.76 bp, 453.09 bp, and 1,552.75 bp, respectively (Table 6).

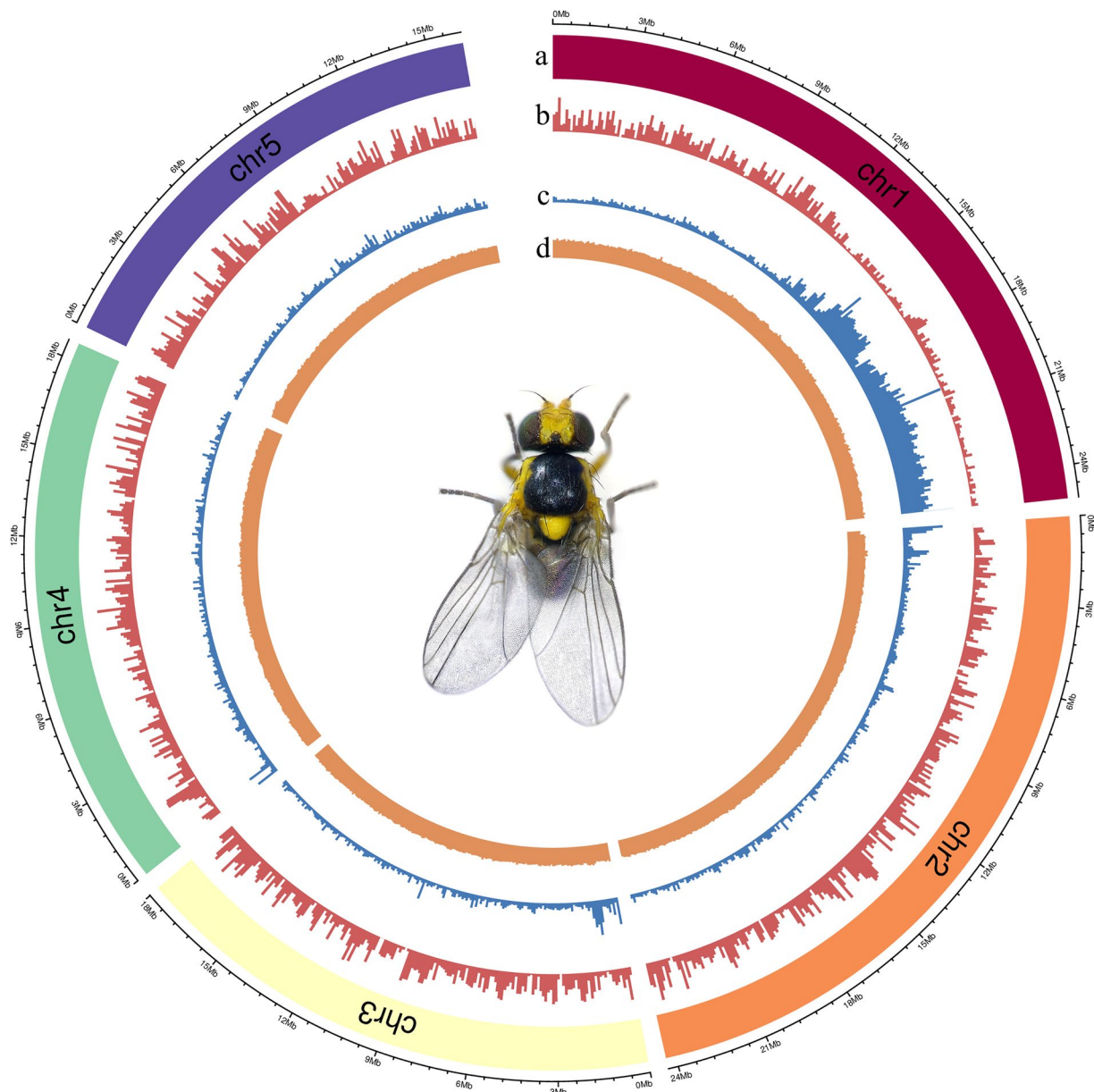


Fig. 3 Circos plot of genomic features in *L. sativae* genome in a 100-kb window size. (a) chromosome. (b) gene density. (c) tandem repeat density. (d) GC content.

Chromosome	Length (bp)	Contig number
chr1	25,088,002	64
chr2	24,335,156	36
chr3	18,023,508	32
chr4	18,573,768	28
chr5	16,192,392	36
chrUnn	8,785,088	333

Table 3. Chromosome length and contig number.

Functional annotations were systematically assigned to all predicted genes by integrating evidence from seven major biological databases: Kyoto Encyclopedia of Genes and Genomes (KEGG)⁴⁵, NCBI non-redundant protein (Nr)⁴⁶, UniProt⁴⁷, Gene Ontology (GO)⁴⁸, Eukaryotic Orthologous Groups (KOG)⁴⁹, Pfam⁵⁰, and InterPro⁵¹. This multi-database approach resulted in the annotation of 9,978 genes (representing 96.76% of the predicted *L. sativae* protein-coding gene repertoire), with detailed annotations cataloged in Table 6.

Type	TE protiens		De novo + rebase		Combined TEs	
	Length (Bp)	% in genome	Length (Bp)	% in genome	Length (Bp)	% in genome
DNA	393,093	0.35	17,183,372	15.48	17,247,090	15.54
LINE	258,535	0.23	4,113,426	3.71	4,183,177	3.77
SINE	0	0.00	220,149	0.20	220,149	0.20
LTR	2,951,195	2.66	6,906,050	6.22	8,000,174	7.21
LTR-Gypsy	1,635,279	1.47	2,808,243	2.53	3,382,869	3.05
LTR-Copia	384,842	0.35	654,424	0.59	730,409	0.66
Other	0	0.00	194	0.00	194	<0.01
Unknown	0	0.00	1,992,771	1.80	1,992,771	1.80
Total	3,600,389	3.24	29,176,948	26.29	30,387,806	27.38

Table 4. Statistics of repeated sequences in the assembled genome of *L. sativae*.

	Assembly		Annotation	
	Proteins	Percentage (%)	Proteins	Percentage (%)
Complete BUSCOs	1297	94.9	1237	90.5
Complete Single-Copy BUSCOs	1282	93.8	1217	89.0
Complete Duplicated BUSCOs	15	1.1	20	1.5
Fragmented BUSCOs	7	0.5	33	2.4
Missing BUSCOs	63	4.6	97	7.1
Total BUSCO groups searched	1367	100.00	1367	100.00

Table 5. Statistics of the BUSCO evaluation results for genome assembly and annotation.

Gene structure annotation	
Total number of gene	10,312
Average of mRNA length (bp)	5,561.81
Average CDS length of per gene (bp)	1,552.75
Average exon number of per gene	4.88
Average of exon length (bp)	453.09
Average of intron length (bp)	864.76
Total number of exon	50,287
Total number of intron	39,975
Total intron length (bp)	34,568,709
Gene function annotation	
Annotation database	Number (Percent)
Annotation	9,978 (96.76%)
KEGG	8,003 (77.61%)
Pathway	3,492 (33.86%)
NR	9,645 (93.53%)
Uniprot	9,594 (93.04%)
GO	8,358 (81.05%)
KOG	1,002 (9.72%)
Pfam	7,925 (76.85%)
Interpro	9,687 (93.94%)

Table 6. Statistics of coding gene structure and functional annotation of the *L. sativae* genome.

Data Records

The genome sequencing data generated in this study have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession number PRJNA1232231. The MGI short reads, PacBio HiFi long reads, and Hi-C sequencing data used for the genome assembly have been deposited with the accession numbers SRR32581919⁵², SRR32581918⁵³ and SRR32581917⁵⁴. The transcriptome and full-length transcriptome (Iso-Seq) data used for genome annotation have been deposited with the accession numbers SRR32581921⁵⁵ and SRR32581920⁵⁶. The genome assembly of *L. sativae* has been deposited at GenBank under the accession number GCA_051363215.1⁵⁷. It has also been deposited in ScienceDB Digital Repository at <https://doi.org/10.57760/sciencedb.21720>⁵⁸. The genome annotation files are available in the Figshare database (<https://doi.org/10.6084/m9.figshare.28532852.v1>)⁵⁹.

Technical Validation

To accurately assess the integrity and quality of genomic DNA, a comprehensive quality control protocol was implemented, which included 0.7% agarose gel electrophoresis, NanoDrop One spectrophotometry (NanoDrop Technologies, Wilmington, DE), and Qubit 3.0 fluorometry (Life Technologies, Carlsbad, CA, USA). The extracted DNA exhibited optimal quality parameters: concentration = 77.1 ng/μL, OD260/280 value = 1.86, and OD260/230 = 2.38.

The final genome assembly exhibited a total size of 110.98 Mb (compared to 122.64 Mb in *L. trifolii*) with a scaffold N50 of 18.57 Mb (*L. trifolii*: 23.84 Mb). Through advanced assembly techniques, 92.08% of contigs (*L. trifolii*: 96.25%) were successfully anchored to five pseudo-chromosomes, mirroring the conserved karyotype observed in *L. trifolii*. During genome assembly and annotation processes, 94.9% and 90.5% of BUSCO genes were identified, respectively (Table 5), indicating exceptional completeness of the *L. sativae* genome assembly. For quality validation, next-generation sequencing short reads were aligned to the final assembly using BWA v0.7.17-r1188, achieving an alignment rate of 99.06%. Hi-C interaction heatmaps revealed well-organized interaction patterns along chromosomal diagonals and inversion regions (Fig. 2), providing orthogonal validation for the accuracy of chromosomal assembly. Collectively, these multi-platform validation results confirm the high contiguity, completeness, and chromosomal-level precision of the *L. sativae* genome assembly.

Code availability

All bioinformatics analyses in this study were conducted using standardized operational protocols as specified in the official documentation of respective software packages. No custom scripts, proprietary algorithms, or non-standardized code implementations were employed throughout the research.

Received: 11 March 2025; Accepted: 1 August 2025;

Published online: 08 August 2025

References

1. von Tschirnhaus, M. & Groll, E. World Agromyzidae Online. <https://sdei.senckenberg.de/tschirnhaus-agromyzidae> (2024).
2. Kang, L., Chen, B., Wei, J. N. & Liu, T. X. Roles of thermal adaptation and chemical ecology in *Liriomyza* distribution and control. *Annual Review of Entomology* **54**, 127–145 (2009).
3. Spencer, K. A. Agromyzidae (Diptera) of Economic Importance. Series Entomologica, **9**, Dr. W. Junk B.V., The Hague, The Netherlands (1973).
4. Reitz, S. R., Gao, Y. L. & Lei, Z. R. Insecticide use and the ecology of invasive *Liriomyza* leafminer management. In Trdan S (ed.), *Insecticides-Development of Safer and More Effective Technologies*. Rijeka: InTech, pp. 235–255 (2013).
5. Frederick, L. P. & David, O. W. Laboratory Rearing and Life History of *Liriomyza sativae* (Diptera: Agromyzidae) on Lima Bean. *Environmental Entomology* **23**, 1416–1421 (1994).
6. Zhao, Y. X. & Kang, L. Cold tolerance of the leafminer *Liriomyza sativae* (Dipt., Agromyzidae). *Journal of Applied Entomology* **124**, 185–189 (2000).
7. Bragard, C. et al. Pest categorisation of *Liriomyza sativae*. *EFSA Journal* **18**, 6037 (2020).
8. Scheffer, S. J. & Lewis, M. L. Mitochondrial phylogeography of vegetable pest *Liriomyza sativae* (Diptera: Agromyzidae): Divergent clades and invasive populations. *Annals of the Entomological Society of America* **98**, 181–186 (2005).
9. Andersen, A. et al. Polyphagous *Liriomyza* species (Diptera: Agromyzidae) in vegetables in Vietnam. *Tropical Agriculture* **779**, 241–246 (2002).
10. Chen, X., Lang, F., Xu, Z., He, J. & Ma, Y. The occurrence of leafminers and their parasitoids on vegetables and weeds in Hangzhou area, Southeast China. *Biocontrol* **48**, 515–527 (2003).
11. Namvar, P., Safaralizadeh, M. H., Baniameri, V., Pourmirza, A. A. & Isfahani, J. K. Spatial distribution and fixed-precision sequential sampling of *Liriomyza sativae* Blanchard (Diptera: Agromyzidae) on cucumber greenhouse. *Middle-East Journal of Scientific Research* **10**, 157–163 (2011).
12. Ridland, P. M., Umina, P. A., Pirtle, E. I. & Hoffmann, A. A. Potential for biological control of the vegetable leafminer, *Liriomyza sativae* (Diptera: Agromyzidae), in Australia with parasitoid wasps. *Austral Entomology* **59**, 16–36 (2020).
13. Alaei Verki, S. T., Iranipour, S. & Karimzadeh, R. Vegetable leafminer, *Liriomyza sativae* (Diptera: Agromyzidae) damage mediated yield loss of cucumber. *North-Western Journal of Zoology* **16**, 134–140 (2020).
14. Shubh, P. S. et al. Evaluation of different insecticides against *Liriomyza sativae* (Diptera: Agromyzidae) on cucumber plants. *Journal of Agriculture and Food Research*. **15**, 100987 (2024).
15. Chang, Y. W., Wang, Y. C., Wang, Y. C. & Du, Y. Z. Chromosome-level genome assembly of the invasive leafminer fly, *Liriomyza trifolii* (Diptera: Agromyzidae). *Scientific Data* **11**, 1326 (2024).
16. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
17. De Coster, W., D’hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. Nanopack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
18. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
19. Vurture, G. W. et al. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
20. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nature Methods*. **17**, 155–158 (2020).
21. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).
22. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nature Plants* **5**, 833–845 (2019).
23. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
24. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell systems* **3**, 95–98 (2016).
25. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell systems* **3**, 99–101 (2016).
26. Wolff, J. et al. Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Research* **48**, W177–W184 (2020).
27. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580 (1999).
28. Beier, S., Thiel, T., Münch, T., Scholz, U. & Mascher, M. MISA-web: a web server for microsatellite prediction. *Bioinformatics* **33**, 2583–2585 (2017).
29. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America* **117**, 9451–9457 (2020).

30. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**, W265–W268 (2007).
31. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
32. Ou, S. & Jiang, N. LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiology* **176**, 1410–1422 (2018).
33. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. <http://www.repeatmasker.org> (2013–2015).
34. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25**, 955–964 (1997).
35. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research* **35**, 3100–3108 (2007).
36. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
37. 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 (2019).
38. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology* **33**, 290–295 (2015).
39. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).
40. Kuo, R. I. *et al.* Illuminating the dark side of the human transcriptome with long read transcript sequencing. *BMC genomics* **21** (2020).
41. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
42. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268**, 78–94 (1997).
43. Li, H. Protein-to-genome alignment with minimap. *Bioinformatics* **39**, btad014 (2023).
44. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution* **38**, 4647–4654 (2021).
45. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28**, 27–30 (2000).
46. Deng, Y. *et al.* Integrated nr database in protein annotation system and its localization. *Computer Engineering* **32**, 71–72 (2006).
47. Apweiler, R. *et al.* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research* **32**, D115–119 (2004).
48. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29 (2000).
49. Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Expanded microbial genome coverage and improved protein family annotation in the cog database. *Nucleic Acids Research* **43**, D261–D269 (2015).
50. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Research* **42**, D222–D230 (2014).
51. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research* **49**, D344–D354 (2021).
52. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR32581919> (2025).
53. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR32581918> (2025).
54. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR32581917> (2025).
55. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR32581921> (2025).
56. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR32581920> (2025).
57. NCBI GenBank https://identifiers.org/ncbi/insdc.gca:GCA_051363215.1 (2025).
58. Chen, X. L. & Yang, D. Chromosome-level genome assembly of *Liriomyza sativae*. *Science Data Bank* <https://doi.org/10.57760/sciencedb.21720> (2025).
59. Chen, X. L. The genome annotation of *Liriomyza sativae*. *figshare* <https://doi.org/10.6084/m9.figshare.28532852.v1> (2025).

Acknowledgements

We sincerely thank Ms. Pengyan You from Henan University of Science and Technology for her valuable assistance in field sample collection. We also extend our thanks to Mr. Fan Gao for providing the photo of *L. sativae*. This study was funded by the Natural Science Special (Special Post) Scientific Research Fund Project of Guizhou University (2023-06) and the Qiankehe Platform Talent (BQW[2024]012).

Author contributions

D.Y. and Z.C. designed the study. X.L.C. collected and reared *L. sativae*. X.L.C., X.D.C., J.Z.L. and Z.F. conducted the bioinformatics analyses. X.L.C. and X.D.C. wrote the original draft manuscript. J.Z.L., S.M.D. and D.Y. revised the manuscript. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.