scientific data

OPEN

DATA DESCRIPTOR

Chromosome-Level Genome Assembly and Annotation of *Piptanthus nepalensis* (Hook.) Sweet

Jiayin Zhang[1,4], Zhefei Zeng[2,3,4], Ngawang Bonjor[2,3], Ngawang Norbu[2,3], Norzin Tso[2,3], Xin Tan[2,3], Yuguo Wang[1]✉, Junwei Wang[2,3]✉ & La Qiong[2,3]✉

Piptanthus nepalensis **(Hook.) Sweet is a leguminous shrub native to the Himalayan region and is valued for its medicinal uses and ornamental yellow flowers. Here, we report a chromosome-scale genome assembly of *P. nepalensis* generated using PacBio HiFi long reads, Illumina short reads, and Hi-C chromatin interaction data. The assembled genome spans approximately 1.04 Gb, with a contig N50 of 39.5 Mb and a scaffold N50 of 111.5 Mb. Benchmarking universal single-copy orthologs (BUSCO) analysis indicated high completeness for both the genome (99.3%) and predicted protein set (99.2%). Approximately 99.0% of the assembled sequences were anchored onto nine pseudo-chromosomes. We annotated 26,035 protein-coding genes, and repetitive elements were estimated to comprise ~75.2% of the genome. This high-quality genome assembly provides a foundational genomic resource for *P. nepalensis* and supports future studies on its biology and utilization.**

Background & Summary

Piptanthus nepalensis (Hooker) Sweet is a semi-deciduous shrub in the family Fabaceae (Leguminosae), distributed in high-altitude regions of the Himalayas[1]. Commonly known as "Nepalese laburnum," it has long been used in Himalayan folk medicine and is cultivated worldwide as an ornamental plant for its bright yellow pea-like flowers[1]. In traditional herbal practices of Nepal, various parts of *P. nepalensis* have been used to treat ailments, and extracts of this plant have demonstrated notable antibacterial activity against both Gram-positive and Gram-negative bacteria[2,3]. The medicinal properties of *P. nepalensis* are attributed to its rich arsenal of secondary metabolites, potentially including quinolizidine alkaloids (such as cytisine found in its seeds), flavonoids, and terpenoids[4–6]. These bioactive constituents underline the plant's therapeutic potential and have sparked interest in its pharmacological evaluation. Beyond its medicinal use, *P. nepalensis* also holds ecological significance as part of alpine shrub communities and is valued for soil stabilization and habitat provision in its native range[2].

Most research on *P. nepalensis* to date has focused on basic phytochemical screening and antibacterial efficacy of its extracts[2,4]. However, detailed investigations into the biosynthetic pathways of its active compounds, as well as genetic studies for crop improvement, remain limited by the absence of a reference genome[7]. High-quality genomic data can illuminate the genes and pathways involved in the production of valuable secondary metabolites and facilitate the discovery of genetic markers for important traits[8,9]. The lack of a reference genome for *P. nepalensis* has thus hindered deeper insights into the molecular basis of its medicinal properties and constrained efforts in breeding or genetic enhancement of this species. Therefore, a chromosome-level genome assembly for *P. nepalensis* is crucial for dissecting its genetic background and understanding the regulation of its bioactive compounds. This knowledge will not only advance fundamental research on this Himalayan medicinal shrub

[1]Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, Institute of Biodiversity Science, School of Life Sciences, Fudan University, Shanghai, 200438, China. [2]Key Laboratory of Biodiversity and Environment on the Qinghai-Tibetan Plateau, Ministry of Education, School of Ecology and Environment, Xizang University, Lhasa, China. [3]Yani Observation and Research Station for Wetland Ecosystem of the Tibet (Xizang) Autonomous Region, Xizang University, Linzhi, China. [4]These authors contributed equally: Jiayin Zhang, Zhefei Zeng. ✉e-mail: wangyg@fudan.edu.cn; jwyx12240315@126.com; lhagchong@163.com

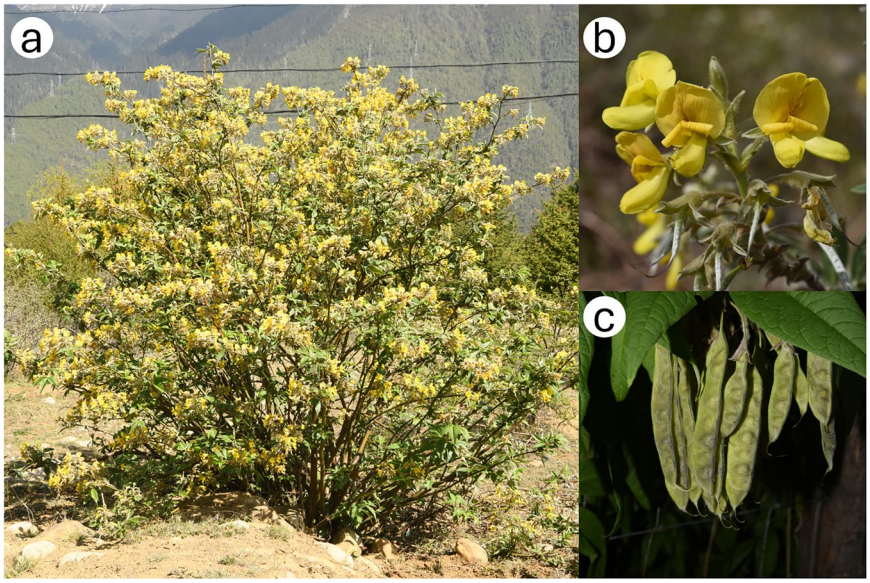

Fig. 1 Morphological characteristics of *Piptanthus nepalensis*. (**a**) A flowering shrub of *P. nepalensis* growing in its natural habitat, (**b**) Close-up view of the yellow flowers during blooming stage, and (**c**) Mature pods at the fruiting stage.

but also support the selection of superior genotypes and the molecular breeding of *P. nepalensis* for enhanced therapeutic and ornamental qualities.

In this study, we report a high-quality, chromosome-level genome assembly of *P. nepalensis* (2n = 18). We generated approximately 119.1 Gb of PacBio HiFi long-read data and 28.9 Gb of transcriptome sequencing data from Illumina, and we integrated 184.1 Gb of Hi-C reads to achieve chromosome scaffolding. The resulting assembly is well annotated and provides comprehensive insight into the genome organization of *P. nepalensis*, including its repeat landscape and gene repertoire. This genomic resource lays the groundwork for future functional genomics studies, such as identifying genes involved in the biosynthesis of medicinal compounds, and will facilitate genetic improvement and conservation of *P. nepalensis*.

Methods

Sample collection and genome sequencing. Fresh young tissues of *P. nepalensis* were collected from a healthy cultivated shrub at Mainling County, Nyingchi City, Xizang Autonomous Region, China (29°20′N, 94°25′E) (Fig. 1). The sample was authenticated morphologically and preserved as a voucher specimen. Leaves were washed with distilled water, flash-frozen in liquid nitrogen, and stored at −80 °C until DNA extraction. High-molecular-weight genomic DNA was extracted using a CTAB-based protocol and purified for library construction[10]. A SMRTbell library was prepared using the PacBio SMRTbell Express Template Prep Kit 2.0 following the manufacturer's instructions. The library was sequenced on the PacBio Revio platform (Pacific Biosciences) to generate HiFi long reads. This yielded approximately 119.1 Gb of HiFi data with an average read length of ~16.1 kb and a read N50 of ~16.1 kb (Table 1). In addition, a short-read whole-genome shotgun library was constructed and sequenced on the Illumina HiSeq 2500 platform to provide data for initial genome survey analysis. This Illumina sequencing produced about 79.8 Gb of 150 bp paired-end reads (Table 1).

To assist in genome annotation, transcriptome sequencing was performed on fresh *P. nepalensis* leaves, root, stems and flowers. Total RNA was extracted from the same source plant, and Illumina cDNA libraries were prepared using the Illumina TruSeq RNA Sample Prep Kit. The libraries were sequenced on the Illumina NovaSeq 6000 platform. After filtering out low-quality reads and adaptors using fastp[11], a total of 28.90 Gb of clean RNA-seq data were obtained. Additionally, a Hi-C library was prepared from young leaf tissue to capture chromatin conformation information using Biomarker Hi-C Library Prep Kit. Following a standard Hi-C protocol[12], cross-linked DNA was digested with *MboI* and ligated to produce chimeric junctions, then size-selected and sequenced on the Illumina NovaSeq 6000. This Hi-C sequencing generated 184.1 Gb of raw paired-end reads, representing approximately 180× coverage of the *P. nepalensis* genome (Table 1).

Genomic characteristics estimation. We employed a kmer approach to estimate the genome size and heterozygosity of *P. nepalensis* prior to assembly. A k-mer frequency distribution analysis was performed on the sequencing data to estimate genome size and heterozygosity *in silico*. Jellyfish[13] was used to count 21-mers in the Illumina short-read dataset, and the resulting k-mer frequency histogram was analyzed with GenomeScope[14]. In addition, we generated a k-mer (k = 19) spectrum from the HiFi reads to complement the short-read analysis. The genome size and heterozygosity rate were inferred by fitting models to the k-mer distribution using GenomeScope2[15]. The k-mer analysis indicated a predominantly diploid genome with distinct peaks corresponding to homozygous and heterozygous k-mers (Fig. 2). The estimates from k-mer profiling were ~0.78 Gb and

Species	tissue	Library strategy	Sequencing platform	Sum raw length	Sum clean length
Piptanthus nepalensis (Hook.) Sweet	root	Transcriptome	Illumina NovaSeq 6000	6,811,745,000	6,748,764,000
Piptanthus nepalensis (Hook.) Sweet	flowers	Transcriptome	Illumina NovaSeq 6000	6,155,576,000	6,106,663,000
Piptanthus nepalensis (Hook.) Sweet	stems	Transcriptome	Illumina NovaSeq 6000	6,470,668,000	6,417,087,000
Piptanthus nepalensis (Hook.) Sweet	leaves	Transcriptome	Illumina NovaSeq 6000	9,460,529,000	9,439,730,000
Piptanthus nepalensis (Hook.) Sweet	leaves	WGS	Illumina HiSeq 2500	79,781,333,000	79,111,258,000
Piptanthus nepalensis (Hook.) Sweet	small leaves	Hi-C	Illumina NovaSeq 6000	184,075,000,000	182,706,000,000
Piptanthus nepalensis (Hook.) Sweet	small leaves	HiFi	PacBio Revio	119,060,727,646	—

Table 1. Samples and sequencing statistics.

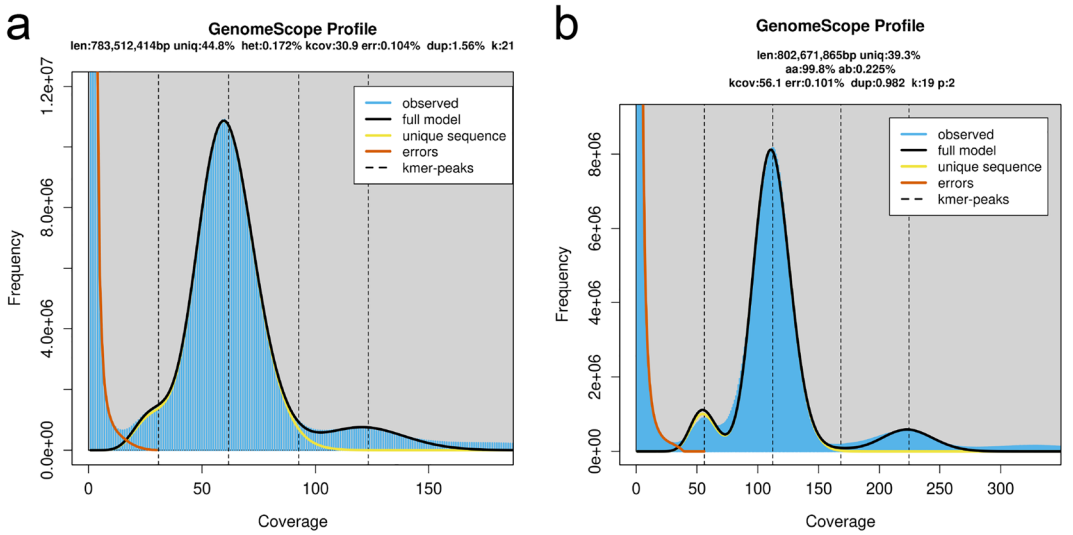

Fig. 2 Genome survey of *P. nepalensis* based on kmer distribution. (**a**) Genome survey based on next generation sequencing data (kmer = 21). (**b**) Genome survey based on third generation sequencing data (kmer = 19).

~0.80 Gb for the genome size based on Illumina and HiFi data, respectively, with inferred heterozygosity rates of approximately 0.172% and 0.225% for each data source (Fig. 2).

Genome assembly and scaffolding. We assembled the *P. nepalensis* genome using the PacBio HiFi reads as the primary data input. The HiFi reads were initially assembled *de novo* with Hifiasm v0.19.9[16] using default parameters, producing a set of contigs. To refine the assembly, we aligned the raw HiFi reads back to these contigs using Minimap2[17] in asm20 mode and calculated the read depth across each contig. Simultaneously, we screened the contigs for organelle-derived sequences by BLASTn[18] search against a database of organelle genomes (including mitochondrial and chloroplast sequences) with a relaxed stringency (e-value 1e-5, identity ≥80%). Contigs showing abnormally high read depth (>200 × the mean coverage) and significant homology to organelle genomes (>85% of length) were identified as likely organelle assemblies and removed from the dataset.

The remaining contigs were then processed to eliminate redundant haplotigs arising from heterozygosity. We applied Purge_Dups (v1.2.6)[19] with default settings, which uses coverage depth and sequence similarity criteria to detect and remove secondary contigs corresponding to alternative haplotypes. This step reduced the total assembly size and improved continuity by collapsing the assembly into a representative haploid set of contigs. After purging, the draft assembly spanned 1.04 Gb with a contig N50 of 39.5 Mb (Table 2).

To construct chromosome-scale scaffolds, we integrated the Hi-C data with the purged contigs. a total of 1,227,165,156 read pairs were processed, of which 1,132,182,573 were retained as valid unique interaction pairs after duplicate removal, representing 92.26% of uniquely mapped pairs. The duplicate rate was very low (0.105%), indicating high library complexity and excellent data quality. These high-quality valid interaction pairs were subsequently used for chromosome-level genome scaffolding. Hi-C paired-end reads were aligned to the contigs using BWA-MEM (v0.7.17)[20] with default parameters, and valid Hi-C contacts (uniquely mapped, non-duplicated read pairs) were retained. We utilized HapHiC[21] to order and orient the contigs into chromosome-length scaffolds. HapHiC clustered the contigs into 9 linkage groups, consistent with the haploid chromosome number (n = 9) of *P. nepalensis*. The contigs within each cluster were arranged and joined based on Hi-C contact frequency, producing 9 draft pseudo-chromosomes.

The scaffolded assembly was further refined by closing gaps and polishing base accuracy. We applied TGS-GapCloser2 (v1.1.1)[22] to fill sequence gaps in the scaffolds using PacBio HiFi reads. Subsequently, two rounds of polishing were carried out with NextPolish2[23] using the Illumina short reads, which correct any remaining base errors. The final genome assembly comprises 9 pseudo-chromosomes and a small number of

Stat Type	Contig Length(bp)	Contig Number	Genome Length(bp)	Genome Number
N50	39,462,528	9	111,527,884	4
N60	32,856,413	12	105,459,186	5
N70	28,221,395	15	103,101,562	6
N80	21,058,044	20	100,488,355	7
N90	14,590,025	26	96,088,255	9
Longest	111,527,884	1	148,012,131	1
Total	1,040,127,259	227	1,040,133,159	168

Table 2. Statistics of *P. nepalensis* genome assembly.

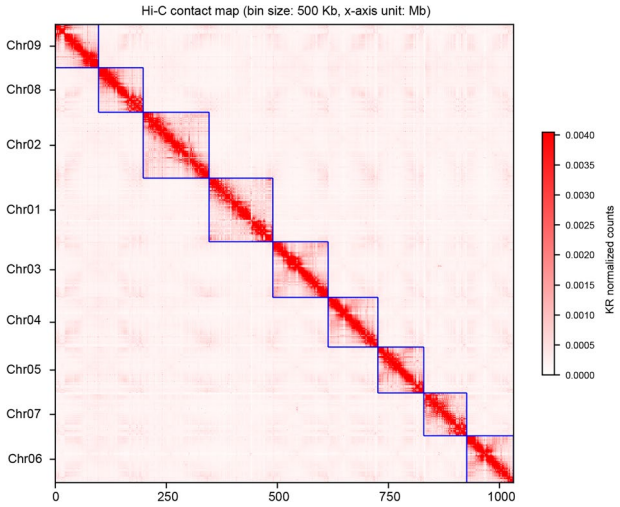

Fig. 3 Heatmap of strength of DNA-DNA interactions discovered by the Hi-C dataset. Interaction frequencies are represented by color intensity, with warmer colors indicating stronger contact signals. Blue frames highlight the nine assembled pseudo-chromosomes of *P. nepalensis*, demonstrating clear intra-chromosomal interaction enrichment and supporting the accuracy and continuity of the chromosome-level assembly.

unplaced small contigs, for a total length of 1.04 Gb (Table 2). The scaffold N50 of the final assembly is 111.5 Mb, indicating that the assembly is highly contiguous (Table 2). The assembled genome size (1.04 Gb) is larger than the k-mer based estimate (~0.78–0.80 Gb), which is likely attributable to the high repeat content of the genome and the known tendency of k-mer approaches to underestimate genome size in repeat-rich plant genomes. Approximately 99.0% of the assembled bases are anchored in the 9 pseudo-chromosomes, reflecting the near-completeness of chromosomal assembly (Fig. 3).

To evaluate the completeness and accuracy of the genome assembly, we employed several strategies. First, we mapped back the raw sequencing data to the assembly. Over 99.91% of the Illumina short reads and 99.99% of the HiFi reads could be aligned to the genome, indicating that the assembly captures nearly all sequenced genomic sequences (Table 3). The high mapping rates also suggest low contamination and that very few genomic segments are missing from the assembly. We then assessed base-level accuracy and phasing using Merqury[24], which leverages k-mer concordance between the assembly and raw reads. The Merqury analysis yielded high quality values (QV) for each pseudo-chromosome (Table 4), with most chromosomes exceeding QV60, and an estimated consensus accuracy >99.99%. The overall assembly QV was approximately 66, corresponding to an estimated error rate on the order of 10^{-7} or lower. Additionally, the LTR Assembly Index (LAI)[25] was computed to evaluate assembly continuity through repetitive long terminal repeat (LTR) retrotransposon regions. The *P. nepalensis* genome achieved an LAI of 12.9 for the assembled portion, which qualifies as a reference-quality assembly (LAI > 10). These quality metrics demonstrate that the assembly is both highly contiguous and accurate.

We further performed a BUSCO (Benchmarking Universal Single-Copy Orthologs) analysis using the embryophyta_odb10 dataset to quantify the completeness of gene space representation in the assembly[26]. The BUSCO results showed that 99.3% of the expected plant orthologs were present in the *P. nepalensis* assembly as complete sequences (95.4% as single-copy and 3.9% as duplicated), with only 0.2% fragmented and 0.5% missing. Likewise, assessment of the annotated protein set (see below) found 99.1% of BUSCOs complete. These high scores are comparable to those of other high-quality plant genomes and confirm the assembly's completeness and the accuracy of gene model predictions (Table 5). Collectively, the mapping statistics, k-mer-based evaluations, and BUSCO analysis all indicate that we have obtained a robust and nearly complete genome assembly for *P. nepalensis*.

	Mapping coverage	meandepth
NGS	99.91%	75.46
TGS	99.99%	112.85

Table 3. Summary of reads mapping rate and meandepth of *P. nepalensis*.

Chromosome	Kmers in assembly	Kmers in assembly and HiFi reads	QV	Error rate
chr01	724	148,011,051	65.8931	2.57449E-07
chr02	1,098	143,230,898	63.9419	4.03472E-07
chr03	791	124,422,575	64.7547	3.34599E-07
chr04	462	111,527,866	66.6149	2.18025E-07
chr05	472	105,458,578	66.2789	2.35563E-07
chr06	480	103,101,426	66.1078	2.45033E-07
chr07	348	100,487,275	67.3928	1.8227E-07
chr08	378	97,216,722	66.8900	2.04644E-07
chr09	321	96,087,293	67.5491	1.75827E-07

Table 4. Evaluation of the assembly quality of the *P. nepalensis* genome by Merqury analysis.

BUSCOs	assembly (%prop)	annotation (%prop)
Complete (C)	1,603 (99.3)	1,600 (99.1)
Complete and single-copy (S)	1,540 (95.4)	1,538 (95.3)
Complete and duplicated (D)	63 (3.9)	62 (3.8)
Fragmented (F)	3 (0.2)	6 (0.4)
Missing (M)	8 (0.5)	8 (0.5)
Total	1,614 (100.0)	1,614 (100.0)

Table 5. BUSCO assessment results for P. nepalensis genome assembly and annotation.

Annotation of repetitive elements. We annotated repetitive elements in the *P. nepalensis* genome using a combination of *de novo* and homology-based approaches. First, we built a custom repeat library *de novo* from the genome assembly. We used LTR_FINDER (v1.07)[27], RepeatScout (v1.0.5)[28], and RepeatModeler (v2.0)[29] to identify and characterize repetitive sequences, especially transposable elements (TEs), within the assembly. The consensus sequences of repeats predicted by these tools were merged into a non-redundant custom library. Next, we combined this custom library with the Repbase-derived library from the RepeatMasker database and the RiTE database of repetitive elements to create a comprehensive reference for repeat masking[30,31]. We then employed RepeatMasker (v4.1.2) to scan the genome, using the combined library to detect and mask repetitive sequences. Additionally, interspersed repeats at the protein level were identified using RepeatMasker's RepeatProteinMask utility, which searches for TE-related protein domains in the genome. Tandem repeats were independently annotated using Tandem Repeats Finder (TRF) (v4.09)[32] to identify satellites and other tandem arrays.

In total, approximately 75.2% of the *P. nepalensis* genome assembly was identified as repetitive DNA, based on the combined homology-based and *de novo* predictions. This exceptionally high repeat proportion is comparable to that of many legume genomes, reflecting extensive proliferation of transposable elements. Among these, retroelements were the most abundant, occupying about 44.1% of the genome (~459 Mb), whereas DNA transposons accounted for approximately 6.7% (~70 Mb) (Fig. 4, Table 6). Long terminal repeat (LTR) retrotransposons, particularly of the *Gypsy* and *Copia* superfamilies, dominated the repetitive fraction, constituting the major component of the genome. These LTR elements showed a non-uniform chromosomal distribution, being enriched in pericentromeric regions, a pattern consistent with observations in other plant genomes. In addition, roughly 20.1% of the assembly comprised unclassified repeats and other minor repeat types such as simple sequence repeats and low-complexity regions. This comprehensive repeat annotation provides valuable insights into genome expansion and organization, and will serve as an essential foundation for subsequent gene annotation.

Gene prediction and functional annotation. We performed gene prediction on the repeat-masked *P. nepalensis* genome using an integrative strategy that combined evidence from ab initio prediction, homologous proteins, and transcriptome data. First, to incorporate empirical transcription evidence, we mapped the RNA-seq reads to the genome. Four sets of Illumina RNA-seq reads derived from *P. nepalensis* tissues were aligned to the assembly using HISAT2 (v2.2.1) with default parameters[33]. The aligned reads were processed with Samtools to sort and index, producing BAM alignment files[34]. These alignments were then used to assemble transcript sequences *de novo* and genome-guided using StringTie2 (v2.1.5), which reconstructed expressed gene models based on the read coverage[35]. The assembled transcripts provided high-confidence exon-intron structures that served as RNA evidence for training gene models.

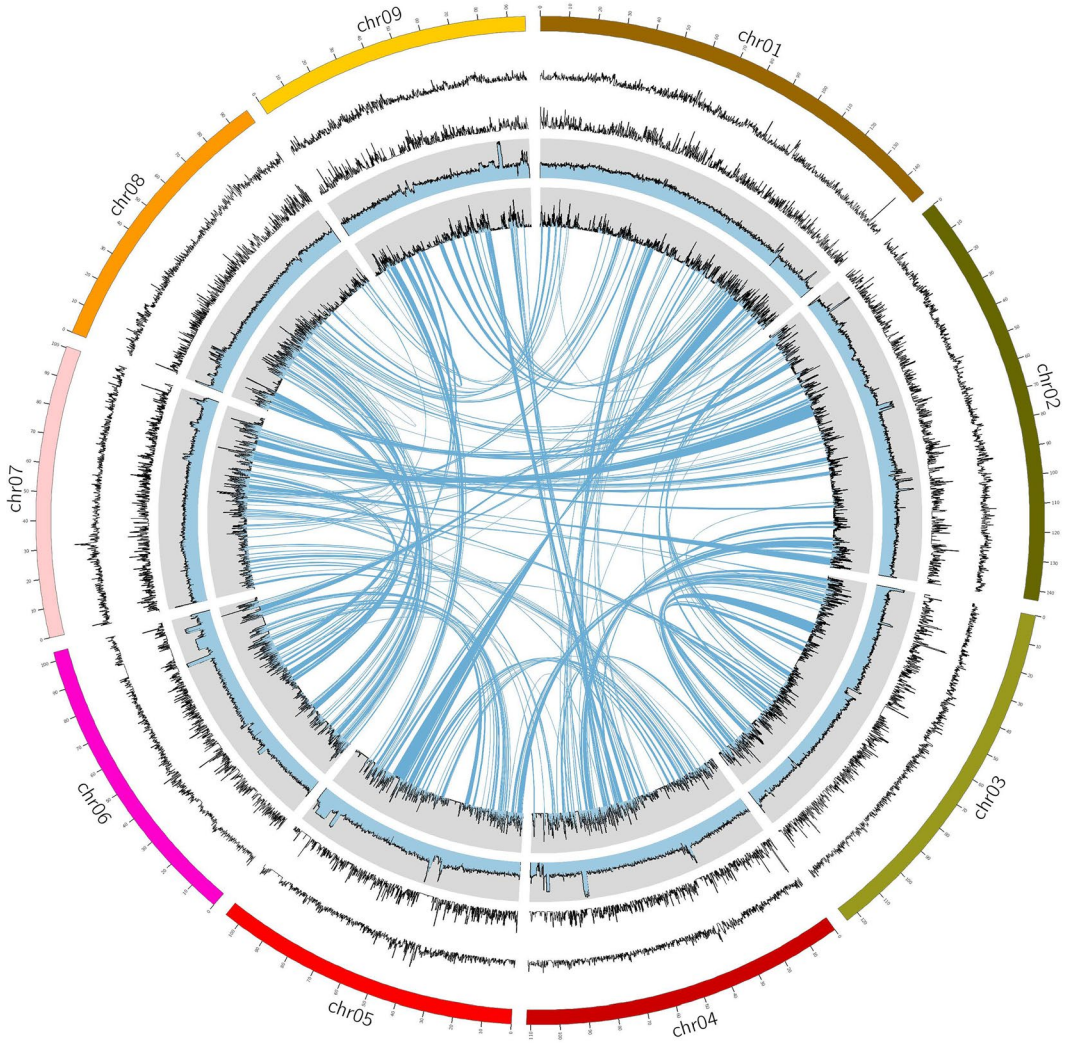

Fig. 4 Illustration of genomic features of 9 pseudo-chromosomes of the *P. nepalensis* genome. From the outermost to the innermost layers, the circular tracks represent *Gypsy* retrotransposon density, *Copia* retrotransposon density, overall repeat element density, GC content, gene density, and collinear blocks, respectively. Together, these features provide an integrated view of chromosome-scale structural organization, repeat distribution patterns, and gene arrangement within the assembled genome.

For *ab initio* and homology-based gene prediction, we used BRAKER2 (v2.1), a pipeline that integrates AUGUSTUS and GeneMark-ET, to predict protein-coding genes[36–38]. BRAKER2 was run in a mode that incorporates the RNA-seq alignments (BAM files) as extrinsic hints to guide gene model inference. In this process, GeneMark-ET was first employed to produce an initial gene prediction set and to train the algorithm using untranslated regions identified by the RNA evidence. The hints (intron positions, exon boundaries) derived from the RNA-seq were then used by AUGUSTUS (v3.4.0) to refine gene models, with iterative training to improve accuracy.

In parallel, we incorporated protein homology evidence. A comprehensive protein sequence database was constructed, including the UniProt/SwissProt[39] plant proteins and the predicted proteomes of three legume species with high-quality genomes—*Ammopiptanthus mongolicus* (PRJCA024714), *Ammopiptanthus nanus* (PRJNA413722), and *Glycine max* (GCA_000004515.5), *Robinia pseudoacacia* (PRJCA011483), *Sophora flavescens* (PRJNA973122), *Sophora koreensis* (SAMN31373487). These species were chosen for their phylogenetic proximity and well-annotated gene sets. We aligned this protein set to the *P. nepalensis* genome using TBLASTN[18] and provided the high-confidence matches to BRAKER2 as homology hints. AUGUSTUS was also supplied with the aligned protein sequences to improve gene model predictions, allowing homologous gene structures to inform exon-intron boundaries in *P. nepalensis*.

The gene prediction step yielded an initial set of gene models, which we filtered and refined. We removed any putative genes that overlapped with known transposable elements or repeat sequences, as well as very short gene models (<150 bp coding sequence) lacking support from either RNA or protein evidence. After filtering, the final annotated gene set of *P. nepalensis* comprises 26,035 high-confidence protein-coding genes (Table 7). These genes encode 44,451 transcripts (including alternatively spliced isoforms), with an average transcript

Repeat elements			Number of elements	Length (bp)	Percentage
Retroelements			703,014	458,789,125	44.1%
	SINEs:		21,339	12,837,266	1.23%
	Penelope:		1,398	342,706	0.03%
	LINEs:		117,614	54,983,364	5.29%
		CRE/SLACS	17	438	0%
		L2/CR1/Rex	0	0	0%
		R1/LOA/Jockey	0	0	0%
		R2/R4/NeSL	0	0	0%
		RTE/Bov-B	7,201	1,828,592	0.18%
		L1/CIN4	110,292	53,146,518	5.11%
	LTR elements:		564,061	390,968,495	37.58%
		BEL/Pao	0	0	0%
		Ty1/Copia	214,491	144,015,480	13.85%
		Gypsy/DIRS1	256,195	200,374,260	19.26%
		Retroviral	0	0	0%
DNA transposons			198,184	69,805,815	6.71%
		hobo-Activator	51,881	10,004,632	0.96%
		Tc1-IS630-Pogo	3,898	1,231,756	0.12%
		En-Spm	0	0	0%
		MULE-MuDR	32,935	8,403,226	0.81%
		PiggyBac	0	0	0%
		Tourist/Harbinger	12,626	2,594,344	0.25%
		Other (Mirage, P-element, Transib)	0	0	0%
Rolling-circles			8,943	3,410,602	0.33%
Unclassified:			772,626	209,237,485	20.12%
Total interspersed repeats:				738,175,940	70.95%
Small RNA:			35,160	15,125,217	1.45%
Satellites:			28,147	3,918,534	0.38%
Simple repeats:			367,073	17,899,658	1.72%
Low complexity:			83,727	4,654,993	0.45%

Table 6. Statistics of repeat elements in the *P. nepalensis* genome.

length of 1,243 bp and an average coding sequence length of 414.5 bp. The gene density in gene-rich regions reaches about 35 genes per Mb, whereas it is markedly lower in repeat-rich regions, reflecting the heterogeneous genome composition (Fig. 4).

We then performed functional annotation of the predicted protein-coding genes to determine their potential biological roles. The protein sequences of *P. nepalensis* were compared against several public databases. Homology-based BLASTP searches (E-value $\leq 1 \times 10^{-5}$) were conducted against the SwissProt and NCBI non-redundant (NR) protein databases, resulting in functional assignments for 17,583 (67.54%) and 22,760 (87.42%) genes, respectively[18,39,40]. In addition, EggNOG and COG/KOG databases provided orthology-based functional categories for 21,074 genes (80.94%) (Table 7)[41,42].

To identify conserved domains and associated functions, HMMER (v3.3) searches were performed against the Pfam database[43,44], and Gene Ontology (GO) terms were inferred based on both sequence similarity and domain annotations, covering 12,703 genes (48.79%)[45]. Furthermore, 7,218 genes (27.72%) were mapped to pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Table 7)[46].

In addition to protein-coding genes, we annotated non-coding RNA genes in the *P. nepalensis* genome. For non-coding RNA annotation, rRNA genes were predicted using Barrnap v0.9 (https://github.com/tseemann/barrnap), identifying 27,891 rRNA features across the *P. nepalensis* genome[47]. tRNA genes were detected with tRNAscan-SE v2.0, yielding 2,162 candidate tRNA loci[48]. Additional non-coding RNAs, including miRNAs (237), snRNAs (1,905), and other ncRNAs (2,851), were identified by searching against the Rfam.cm database using Cmscan, with related families grouped according to the Rfam.clanin file[43,49]. In total, 7,155 non-coding RNA genes were annotated in the genome.

Data Records

The raw Illumina, PacBio, Hi-C and RNAseq sequencing data have been deposited in the NCBI SRA database under accession number SRP679460[50]. The finalized chromosome-level genome assembly has been deposited in the NCBI GenBank JBVQTV000000000[51]. The corresponding genome annotation files have been deposited in the Figshare (https://doi.org/10.6084/m9.figshare.30416158.v1)[52].

Database	Number	Percentage
EggNOG	21,074	80.94%
GO	12,703	48.79%
COG/KOG	21,074	80.94%
KEGG Pathway	7,218	27.72%
SwissProt	17,583	67.54%
NR	22,760	87.42%
Total	26,035	100%

Table 7. Summary of gene function annotations of *P. nepalensis* genome.

Technical Validation

To ensure the reliability of the *P. nepalensis* genome data, stringent quality control and validation procedures were applied during sequencing, assembly, and annotation.

Sequencing data quality. High-quality sequencing reads were obtained from multiple platforms. After adapter trimming and quality filtering with *fastp* ($Q \geq 20$), 99.2% of the Illumina reads (79.1 Gb, ~80×) and 99.3% of the Hi-C reads (182.7 Gb, ~180×) were retained. PacBio HiFi data showed an N50 read length of 16.6 kb and a mean read QV of ~30, indicating very low base-calling error rates. RNA-seq libraries achieved >98% bases with Q20 quality or higher, providing a robust dataset for transcript support and gene model validation.

Assembly quality and completeness. The final assembly spans 1.04 Gb across nine pseudo-chromosomes, consistent with the haploid chromosome number of *P. nepalensis*. It achieved a scaffold N50 of 111.5 Mb, with ~99% of bases anchored to the pseudo-chromosomes. Read mapping confirmed the completeness and accuracy of the assembly, with alignment rates of 99.91% (Illumina) and 99.99% (HiFi). Merqury analysis estimated a consensus QV ≈ 66 (error rate $<10^{-7}$), and the BUSCO completeness reached 99.3%, comparable to reference-grade plant genomes. The *Hi-C* contact map further verified correct chromosomal organization with clear intra-chromosomal interaction signals.

Collectively, these metrics demonstrate that the *P. nepalensis* genome assembly is highly contiguous, complete, and accurate, providing a reliable foundation for downstream comparative and functional analyses.

Data availability

The raw Illumina, PacBio, Hi-C, and RNA-seq sequencing data are available in the NCBI Sequence Read Archive (SRA) under accession number SRP679460. The final chromosome-level genome assembly is available in NCBI GenBank under accession number JBVQTV000000000. The genome annotation files are available in Figshare (https://doi.org/10.6084/m9.figshare.30416158.v1).

Code availability

All software and pipelines were implemented in full compliance with the manuals and protocols specified by the respective published bioinformatics tools. No custom programming or coding was employed.

Received: 25 October 2025; Accepted: 25 March 2026;
Published online: 31 March 2026

References

1. Wu, Z. Y., Raven, P. H. & Hong, D. Y. (eds.) Flora of China, Vol. 10 (Fabaceae). Beijing: Science Press; St. Louis: Missouri Botanical Garden Press (2010).
2. Bhattarai, S. & Basukala, O. Antibacterial activity of selected ethnomedicinal plants of Sagarmatha region of Nepal. *Int. J. Ther. Appl.* **31**, 27–31, https://doi.org/10.20530/IJTA_31_27-31 (2016).
3. Tamang, R. *et al*. Ethnomedicinal uses of plants and their antibacterial activities in Solukhumbu district, eastern Nepal. *BMC Complement. Med. Ther.* **20**, 201, https://doi.org/10.1186/s12906-020-02986-9 (2020).
4. Paris, R. R., Faugeras, G. & Dobremez, J.-F. Isoflavones des rameaux de *Piptanthus nepalensis*. *Planta Med.* **29**(1), 32–36, https://doi.org/10.1055/s-0028-1097625 (1976).
5. Sener, B. *et al*. Quinolizidine alkaloids in the Leguminosae: distribution and biological activities. *Phytochem. Rev.* **2**, 123–136, https://doi.org/10.1023/A:1026064906835 (2003).
6. Wink, M. Evolution of secondary metabolites from an ecological and molecular phylogenetic perspective. *Phytochemistry* **64**, 3–19, https://doi.org/10.1016/S0031-9422(03)00300-5 (2003).
7. Sun, H. *et al*. The complete chloroplast genome of *Piptanthus nepalensis*, a medicinal plant. *Mitochondrial DNA B Resour.* **7**, 796–797, https://doi.org/10.1080/23802359.2021.1994897 (2022).
8. Chen, S. *et al*. Herbal genomics: examining the biology of traditional medicines. *Science* **347**(6219), S27–S29 (2015).
9. Liu, X. *et al*. The genome of the medicinal plant *Macleaya cordata* provides new insights into benzylisoquinoline alkaloid metabolism. *Mol. Plant* **10**, 975–989, https://doi.org/10.1016/j.molp.2017.05.007 (2017).
10. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* **19**, 11–15 (1987).
11. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890, https://doi.org/10.1093/bioinformatics/bty560 (2018).
12. Belton, J. M. *et al*. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276, https://doi.org/10.1016/j.ymeth.2012.05.001 (2012).

13. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).
14. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204, https://doi.org/10.1093/bioinformatics/btx153 (2017).
15. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432, https://doi.org/10.1038/s41467-020-14998-3 (2020).
16. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Hifiasm: a haplotype-resolved assembler for accurate HiFi reads. *Nat. Methods* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
17. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
18. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421, https://doi.org/10.1186/1471-2105-10-421 (2009).
19. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898, https://doi.org/10.1093/bioinformatics/btaa025 (2020).
20. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324 (2009).
21. Zeng, X. *et al*. Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nat. Plants* **10**, 1184–1200, https://doi.org/10.1038/s41477-024-01755-3 (2024).
22. Xu, M. *et al*. TGS-GapCloser: a fast and accurate gap closer for large genomes with third-generation sequencing reads. *GigaScience* **9**, giaa067, https://doi.org/10.1093/gigascience/giaa067 (2020).
23. Hu, J. *et al*. NextPolish2: A repeat-aware polishing tool for genomes assembled using HiFi long reads. *Genomics Proteomics Bioinformatics* **22**(1), qzad009, https://doi.org/10.1093/gpbjnl/qzad009 (2024).
24. Rhie, A. *et al*. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245, https://doi.org/10.1186/s13059-020-02134-9 (2020).
25. Ou, S. *et al*. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126, https://doi.org/10.1093/nar/gky730 (2018).
26. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212, https://doi.org/10.1093/bioinformatics/btv351 (2015).
27. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268, https://doi.org/10.1093/nar/gkm286 (2007).
28. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358, https://doi.org/10.1093/bioinformatics/bti1018 (2005).
29. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
30. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **25**, 4.10.1–4.10.14, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
31. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467, https://doi.org/10.1159/000084979 (2005).
32. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580, https://doi.org/10.1093/nar/27.2.573 (1999).
33. Kim, D., Langmead, B. & Salzberg, S. L. HISAT2: graph-based alignment of next-generation sequencing reads to a population of genomes. *Nat. Methods* **12**, 357–360, https://doi.org/10.1038/nmeth.3317 (2015).
34. Danecek, P. *et al*. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008, https://doi.org/10.1093/gigascience/giab008 (2021).
35. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278, https://doi.org/10.1186/s13059-019-1910-1 (2019).
36. 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 Bioinformatics* **3**, lqaa108, https://doi.org/10.1093/nargab/lqaa108 (2021).
37. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439, https://doi.org/10.1093/nar/gkl200 (2006).
38. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O. & Borodovsky, M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**, 6494–6506, https://doi.org/10.1093/nar/gki937 (2005).
39. O'Leary, N. A. *et al*. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745, https://doi.org/10.1093/nar/gkv1189 (2016).
40. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531, https://doi.org/10.1093/nar/gkac1052 (2023).
41. Huerta-Cepas, J. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource. *Nucleic Acids Res.* **47**, D309–D314, https://doi.org/10.1093/nar/gky1085 (2019).
42. Tatusov, R. L. *et al*. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **11**, 41, https://doi.org/10.1186/1471-2105-11-41 (2010).
43. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195, https://doi.org/10.1371/journal.pcbi.1002195 (2011).
44. Mistry, J. *et al*. Pfam: the protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419, https://doi.org/10.1093/nar/gkaa913 (2021).
45. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29, https://doi.org/10.1038/75556 (2000).
46. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30, https://doi.org/10.1093/nar/28.1.27 (2000).
47. Seemann, T. *Barrnap 0.9: rapid ribosomal RNA prediction*. https://github.com/tseemann/barrnap (2014).
48. Chan, P. P. & Lowe, T. M. tRNAscan-SE: searching for tRNA genes in genomic sequences. *Nucleic Acids Res.* **47**, 1–14, https://doi.org/10.1093/nar/gky1048 (2019).
49. Ontiveros-Palacios, N. *et al*. Rfam 15: RNA families database in 2025. *Nucleic Acids Res.* **53**(D1), D258–D267, https://doi.org/10.1093/nar/gkae1023 (2025).
50. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP679460 (2026).
51. Zhang, C. *et al*. *Piptanthus nepalensis* Genome sequencing and assembly. *Genbank* https://identifiers.org/ncbi/insdc:JBVQTV000000000 (2026).
52. Zhang, J. *et al*. The genome annotation files of *Piptanthus nepalensis*. *Figshare*. https://doi.org/10.6084/m9.figshare.30416158.v1 (2026).

Acknowledgements
The computations in this research were performed using the CFFF platform of Fudan University. This work was supported by the Science and Technology Projects of Xizang Autonomous Region, China (Project Nos. XZ202402ZD0005, XZ202402ZY0023, XZ202402JX0003, XZ202401ZR0028, and XZ202303ZY0002G), and by the Open Project of the Key Laboratory of Biodiversity and Environment on the Qinghai-Tibet Plateau, Ministry of Education (Project No. KLBE2025003).

Author contributions
Y.W., J.W., and L.Q. conceived and designed the research framework. J.Z., and Z.Z. collected the samples, prepared experimental materials, and conducted laboratory work. N.B., N.N., N.T., and X.T. contributed to data acquisition, curation, and preliminary analyses. Z.Z. and J.Z. performed the formal data analysis and drafted the initial manuscript, including figures and tables. Y.W., J.W., and L.Q. critically revised the manuscript for intellectual content and provided guidance throughout the study. All authors contributed to improving the manuscript and approved the final version for publication.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Y.W., J.W. or L.Q.

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.