scientific data

Check for updates

OPEN

DATA DESCRIPTOR

The high-quality chromosome-level genome assembly of *Dracocephalum rupestre* Hance

Qian Zhao[1,2], Zixu Fan[1], Hui Yu[1,3] ✉ & Zhanli Wang[1] ✉

Dracocephalum rupestre **Hance is China's traditional herbal medicine in the family Labiatae with numerous health benefits, including anti-inflammatory, antiviral and anti-tumor activities. However, the genus *Dracocephalum* has no reference genome currently, which restricts the research on the breeding, cultivation and exploration of medicinal properties in *D. rupestre*. Thus, we present the high-quality chromosome-level genome assembly of *D. rupestre* using a combination of Pacbio HiFi sequencing and Hi-C scaffolding technologies. The final genome was 435.45 Mb with a contig N50 of 49.83 Mb and a scaffold N50 of 59.06 Mb. The assembled sequences were anchored to 7 chromosomes with an integration efficiency of 96.96%. Furthermore, we predicted 25,865 protein-coding genes, 98.23% of which were functionally annotated. These results offer valuable resources for understanding the genetic basis of the unique phenotypes of *D. rupestre* and will facilitate further study of the functional genomics of this species.**

Background & Summary

Dracocephalum rupestre Hance, commonly referred to as Mao Jiancao and Yan Qinglan in China, is a traditional Chinese medicine that belongs to the plant family Labiatae[1]. The herb is mainly distributed in Shanxi, Inner Mongolia, Hebei, Qinghai and Liaoning provinces in northern China and grows in alpine meadows at an altitude from 650 to 2400 m (2650–3100 m in Qinghai province)[1–4] (Fig. 1A). As a perennial plant, the whole herb of *D. rupestre* has a special aroma and can be used as medicine commonly treating chronic diseases such as ischemic cardiomyopathy, hepatitis and hypertension[2,4]. Moreover, Maojian tea is a popular drink made from the tender stems and leaves of *D. rupestre* through traditional processing methods. Tea has healthy beneficial functions with an elegant aroma and pleasant taste, which can not only help digestion and clean the intestines but also prevent cancer and delay aging[2–5]. *D. rupestre* contains high levels of flavonoids, polyphenols and triterpenes[2–4]. The polyphenols and flavonoids in *D. rupestre* have a significant scavenging effect on free radicals, which reduces the damage of oxidative and aging[6]. Among the flavonoids in *D. rupestre*, apigenin, naringin and luteolin are the main components, which have anti-inflammatory, antioxidant, antiplatelet aggregation, antibacterial, neuroprotective and antihypertensive effects[7]. Pentacyclic triterpene compounds in *D. rupestre*, such as β-sitosterol, betulinic acid and oleanolic acid, have anti-inflammatory and anti-tumor activities[2]. Therefore, *D. rupestre* is a functional herb with high medicinal value that contributes to most health benefits.

Presently, there is no reference genome in the genus *Dracocephalum*, and the research on *D. rupestre* is mainly focused on the biologically active molecules and basic introduction and cultivation[2–4,6–8]. With the development of the medicinal value of Maojian tea, it's increasingly important for the study of *D. rupestre* in scientific cultivation and efficient extraction of active components. However, the lack of genomic information restricts the functional research of *D. rupestre* to a certain extent, which delays the further exploration and exploitation of medicinal properties of plant. Previously, some medicinal plants like Scutellaria baicalensis[9] and Salvia miltiorrhiza[10] found the biosynthesis pathways of related pharmaceutical compounds via combining with genome, transcriptome and metabolome to find key genes with the activity of catalyzing enzymes, which provides new clues for molecular improvement of medicinal plants and efficient synthesis of key compounds. Therefore, the publication of *D. rupestre* genomes lays the foundation for the identification and characterization of the

[1]Inner Mongolia Key Laboratory of Disease-Related Biomarkers, The Second Affiliated Hospital, Baotou Medical College, Baotou, 014030, China. [2]Translational Medicine Center, Baotou Medical College, Baotou, 014040, China. [3]School of Basic Medicine, Baotou Medical College, Baotou, 014040, China. ✉e-mail: huiyu2008@hotmail.com; wang.zhanli@hotmail.com

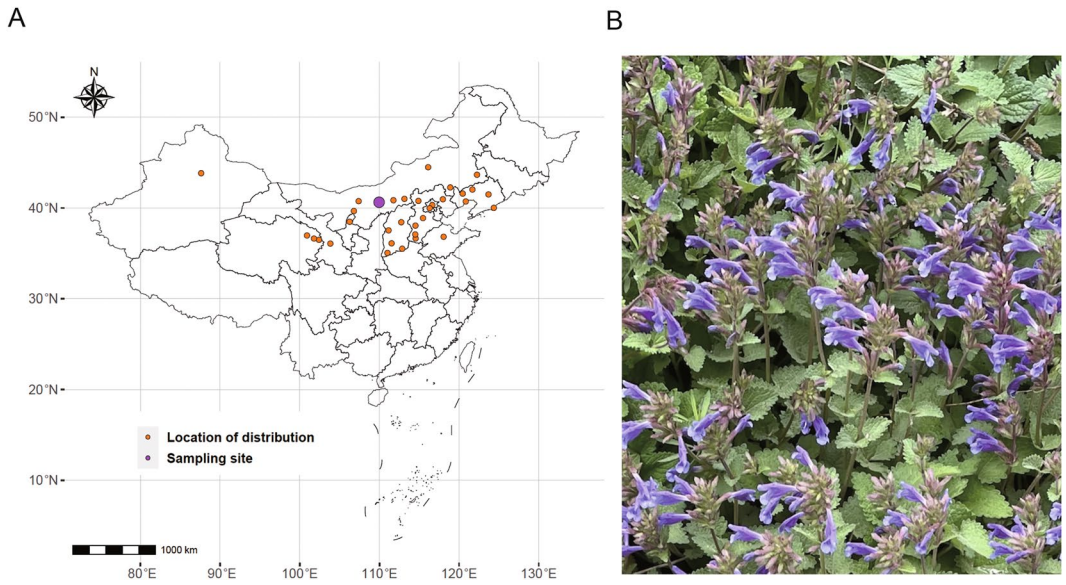


Fig. 1 Distribution and photograph of *D. rupestre*. (**A**) The location map of *D. rupestre* distribution and sampling site. (**B**) The photograph of *D. rupestre* in the sampling site.

important active compounds contained in plant and their biosynthetic pathways, which can also provide new direction for the functional research of other similar traditional Chinese herbs.

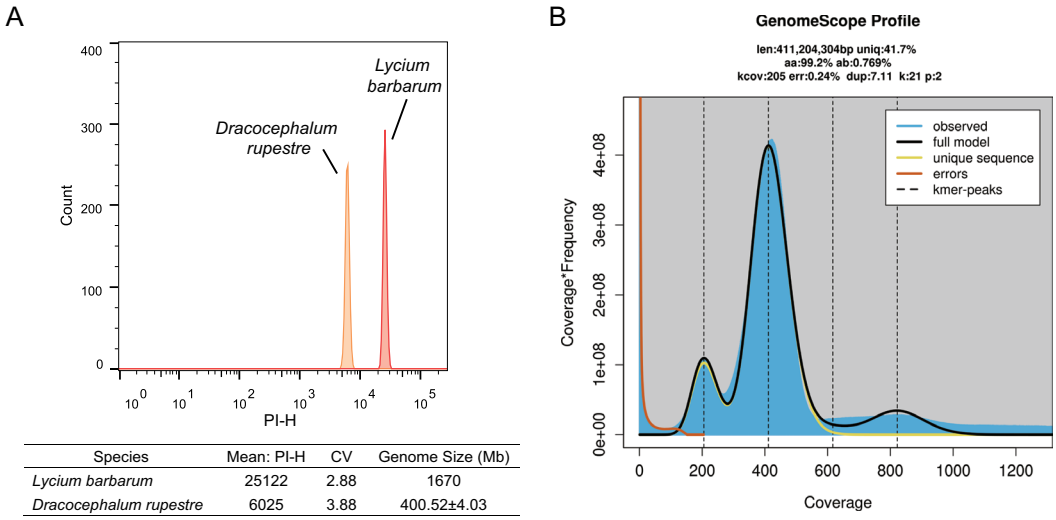
In this study, we assembled the first chromosome-level genome of *D. rupestre* with high quality using PacBio, HiFi (high-fidelity), and Hi-C (high-throughput chromatin conformation capture) technologies. During the process, 25.29 Gb of HiFi reads and 75.70 Gb of Hi-C data were used for assembly and chromosome anchoring. The final results showed that the size of genome assembly is approximately 435.45 Mb with a contig N50 of 49.83 Mb and a scaffold N50 of 59.06 Mb, and 96.96% of the assembled bases were associated with the 7 chromosomes. Here, we report a high-quality genome assembly of *D. rupestre*, one of the representative genomes for the genus *Dracocephalum*. These high-quality genomic data are expected to improve the identification efficiency of the important active compounds and provide new clues for the scientific introduction and cultivation of *D. rupestre*. Moreover, in-depth research on the synthetic pathways of the important active compounds in *D. rupestre* will provide useful value to the metabolic synthesis network of other plants in the genus *Dracocephalum*.

Methods

Sample collection. The *D. rupestre* was planted in the medicinal plants' conservation nursery of BaoTou Medical College in Baotou, China. Thereafter, the young leaf samples were collected and treated with liquid nitrogen for DNA extraction using the DNAsecure Plant Kit (Tiangen, Beijing, China) for genome survey and sequencing (Fig. 1B). Five tissues (roots, stems, young leaves, mature leaves and flowers) were harvested from a mature *D. rupestre* plant for total RNA extraction using EASYspin Plus Polysaccharide Polyphenol/Complex Plant RNA Rapid Extraction Kit (Aidlab, Beijing, China) for sequencing. Three biological replicates were performed, each with three technical replicates.

Genome survey. Flow cytometry was used to experimentally estimate the genome size of *D. rupestre*[11–13]. The young leaves were cleaned with sterile water and chopped up with a sharp razor blade in Galbraith's isolation buffer (45 mM $MgCl_2$, 30 mM Sodium citrate, 20 mM MOPS, 0.1% (v/v) TritonX-100). The homogenate was incubated on ice for 30 min and then filtered through 48-μm nylon mesh. Plant cell nuclei were stained by adding 150 μl of buffer containing propidium iodide (1 mg/mL) and RNAse A (0.2 mg/mL) in the dark for 10 min on ice. *Lycium barbarum* was used as an internal standard to analyze the genome size of *D. rupestre*. The experiment was performed on a BD FACSCanto II flow cytometer (Becton-Dickinson, San Jose, CA, USA) (Fig. 2A).

K-mer frequency analysis was also used to evaluate the genome size[13]. The genomic DNA of *D. rupestre* was extracted from young leaves using the CTAB method[14]. Qubit fluorometer, NanoDrop and agarose gel electrophoresis were used for concentration detection and quality control. The qualified samples were used for the construction of the sequencing library with the inserted fragment of about 300–400 bp, then the DNBseq sequencing platform was used to perform the double-terminal sequencing and a total of 221.92 Gb of raw data was obtained (Table 1). After the adaptor removing and low-quality reads filtering by Fastp (v0.23.2)[15], 205.85 Gb of clean data was retained to generate k-mer (k = 21). According to the k-mer frequency calculated by Jellyfish (v2.3.0)[16] and GenomeScope (v2.0)[17], the genome size was estimated to be 411 Mb, which was approximate with the result of 400.52 ± 4.03 Mb by flow cytometry. The heterozygous ratio and repeat ratio were estimated to be 0.769% and 58.326%, respectively (Fig. 2B).



Species	Mean: PI-H	CV	Genome Size (Mb)
Lycium barbarum	25122	2.88	1670
Dracocephalum rupestre	6025	3.88	400.52±4.03

Fig. 2 The results of genome survey in *D. rupestre*. (**A**) The detection of flow cytometry in *D. rupestre* and *L. barbarum*. (**B**) Depths and frequency distributions of K-mers in *D. rupestre*.

Platform (type)	Total Reads	Raw Bases (Gb)	Read Length/Mean Length(bp)	N50 Length(bp)	Q30 (%)
DNBseq (DNA)	800042338	240.01	150	/	93.31
DNBseq (RNA)	1071187116	160.68	150	/	96.82
PacBio Revio (DNA)	1228658	25.29	20584.89 (CCS reads)	20515	/
PacBio Sequel (RNA)	556899	49.24	88422.78 (Polymerase reads)	150561	/
Hi-C	505853218	75.88	150	/	93.87

Table 1. Summary of sequencing data.

Statistical Level	Contig Length (bp)	Contig Number
N90	23111541	10
N80	25430900	8
N70	41473441	6
N60	48753916	5
N50	49830734	4
Total length	461177119	
Number(>=2 kb)		217
Max length	67724620	

Table 2. Statistics of assembly.

PacBio library construction, sequencing and *de novo* assembly using HiFi reads. For long-read DNA sequencing, the SMRTbell library was constructed using genomic DNA according to the standard PacBio protocol with an insert size of approximately 20 kb (Pacific Biosciences, USA), and the SMRT cell was sequenced on the PacBio Revio platform in the circular consensus sequencing (CCS) model (Pacific Biosciences, USA). In total, 25.29 Gb of HiFi (high-fidelity) reads with N50 of 20.52 Mb were generated for the *de novo* assembly. Hifiasm (v0.19.6)[18] with default parameters was used to generate primary contigs. After redundancy removal and depollution by Purge-Haplotigs (v1.0.4)[19] and BLAST (v2.11.0+)[20], the haploid genome yielded with a size of 461.18 Mb composed by 217 contigs, and the N50 length was 49.83 Mb (Tables 1, 2).

For long-read RNA sequencing, total RNA was isolated from 5 tissues (roots, stems, young leaves, mature leaves and flowers) of a plant. After the detection of quality and integrity by Agilent 2100 Bioanalyzer, the high-quality RNA with RIN >9 was used for reverse transcription by UMI base PCR cDNA Synthesis Kit (BGI, China) to form double-stranded cDNA. A library of single-molecule long-read isoform sequencing (ISOseq) was constructed by the large-scale amplification of cDNA. The PacBio Sequel platform was used to perform the sequencing of a SMRT cell and 49.24 Gb of total data as well as 48.50 Gb of subreads data were obtained. After identification and classification of reads of insert (ROI) by Smrt analysis tool (v12.0) (https://www.pacb.com/products-and-services/analytical-software/smrt-analysis/), there were 347,309 full-length non-chimeric (FLNC) reads using for clustering and polishing using ICE (iterative clustering and error correction) algorithm and

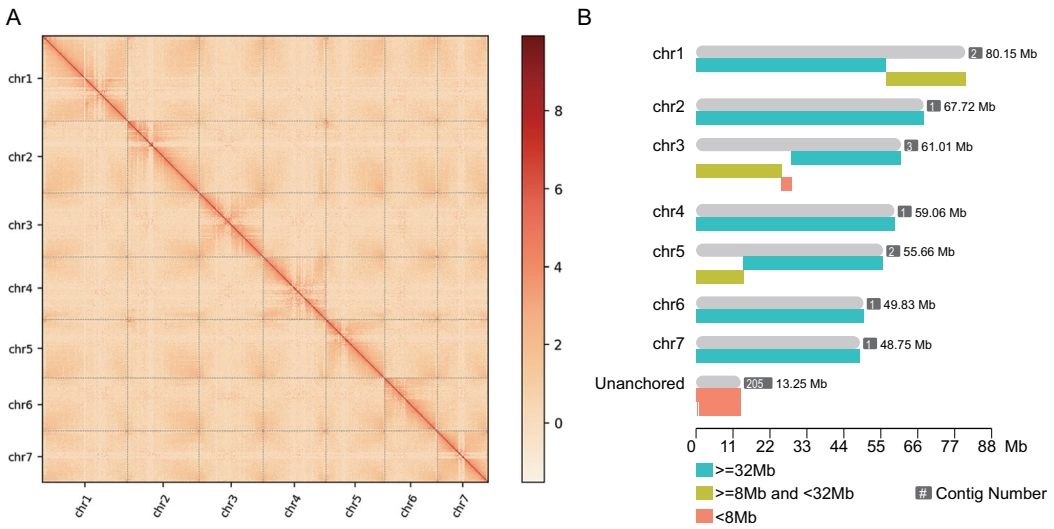


Fig. 3 The Hi-C results of *D. rupestre*. Chromosome-level Hi-C interaction heatmap (**A**) and contig distribution map (**B**).

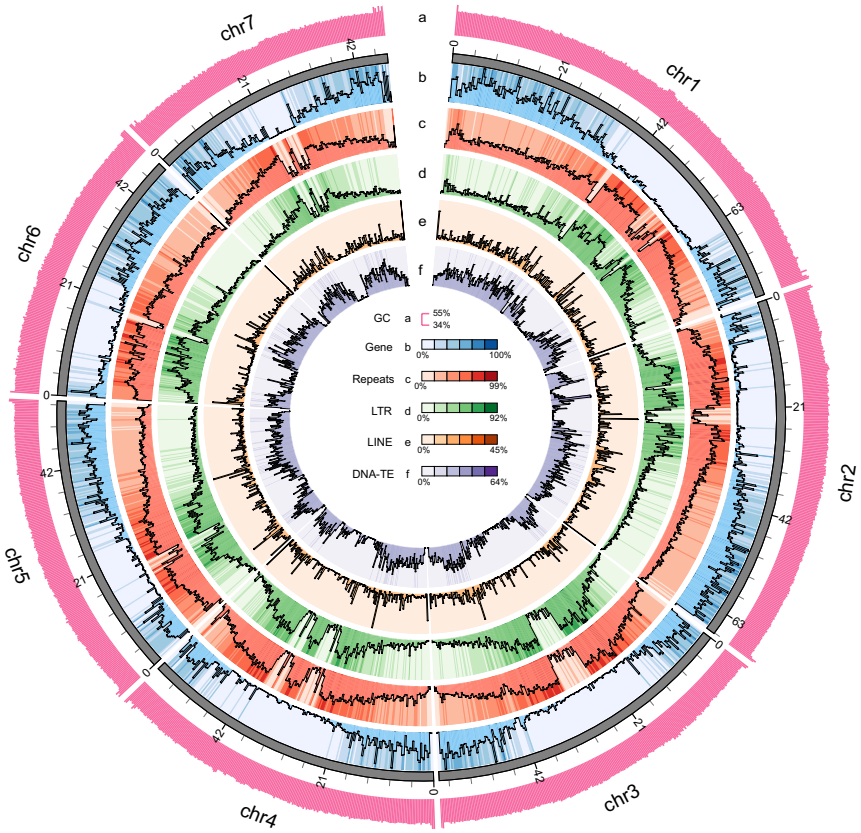


Fig. 4 Genomic landscape of the 7 assembled chromosomes of *D. rupestre*. Note: Window size = Genome size/3000 (kb).

Quiver algorithm, and 211,626 isoforms with the N50 of 1402 bp were finally obtained (Table 1; Supplementary Table 1).

Hi-C library construction, sequencing and chromosome anchoring. The crosslinked DNA was fixed in the cell nucleus by formaldehyde, and the restriction enzyme Mbo I was used to digest the DNA. The ends were filled and marked with biotin, then ligated by DNA ligase to form a loop. The crosslinked fragments were obtained via protease digestion and purification. After being interrupted again using ultrasound, the crosslinked

Type	Repeat Size (bp)	% of genome
TRF	63279706	14.53
Repeatmasker	45424723	10.43
Proteinmask	45828036	10.52
De novo	130241911	29.91
Total	218157857	50.1

Table 3. The statistical results of repeat sequences.

	Gene set	Protein codinggene number	Average gene length (bp)	Average CDS length (bp)	Average exon per gene	Average exon length (bp)	Average intron length (bp)
De novo	GlimmmerHMM	82376	4482.13	1538.47	3.18	483.95	1350.93
	AUGUSTUS	53252	2113.29	1027.1	4.05	253.42	355.78
Homolog	Salvia miltiorrhiza	59974	5353.47	1408.45	4.67	301.75	1075.64
	Salvia splendens	58009	4057.03	1262.03	5.04	250.45	692
	Arabidopsis thaliana	44318	4931.9	1322.99	5.99	220.77	722.84
	Salvia hispanica	52057	4841.57	1373.34	5.16	266.41	834.71
	Scutellaria baicalensis	31164	3729.31	1188.96	4.44	268.04	739.37
Liftoff	Scutellaria baicalensis	13602	2097.49	1140.13	4.28	266.15	291.54
	Salvia miltiorrhiza	20239	2420.28	1245.48	4.91	253.43	300.12
Trans ORF	RNAseq	17568	4555.15	1495.21	6.68	355.01	384.45
	ISOseq	6706	10709.21	1181.71	6.44	270.89	1646.68
BUSCO		1659	4721.77	1691.44	9.9	170.81	340.4
MAKER		25865	4032.92	1176.15	5.16	296.15	601.24
HiCESAP		25899	4555.51	1325.14	5.59	333.23	586.3

Table 4. The statistics of predicted protein-coding genes.

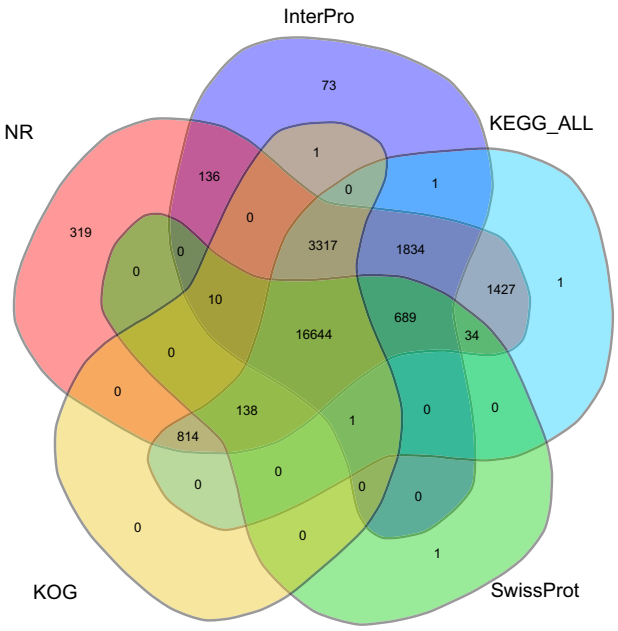


Fig. 5 Venn diagram showed the annotation results of five databases' intersection.

DNA with the size of 300–500 bp marked with biotin was enriched for the Hi-C library construction. The DNBseq sequencing platform was used to perform the double-terminal sequencing and a total of 75.70 Gb of clean data was obtained from 75.88 Gb of raw data (Table 1).

The Hi-C data was further used to anchor the contigs to chromosomes. HiCUP (v0.7.2)[21] with default parameters was used to map the data to the genome, and then the valid pairs (65.17%) were obtained by the Juicer pipeline (v1.6)[22] to anchor the contigs into chromosomes with 3D-DNA pipeline (v180,922, parameters: -r 0) (https://github.com/theaidenlab/3d-dna). Finally, Juicebox (v1.11.08)[23] was used for manual checking

Classification	Databses	Gene		mRNA	
		Number	Percent (%)	Number	Percent (%)
Total		25899	100	35602	100
Annotated		25441	98.23	34795	97.73
	NR	25362	97.93	34667	97.37
	SwissProt	17517	67.64	24132	67.78
	TrEMBL	25251	97.5	34525	96.97
	KOG	20925	80.79	28667	80.52
	TF	1842	7.11	2449	6.88
	InterPro	22706	87.67	31196	87.62
	GO	16757	64.7	23185	65.12
	KEGG_ALL	24900	96.14	34049	95.64
	KEGG_KO	10373	40.05	14795	41.56
	Pfam	20811	80.35	28343	79.61
Unannotated		458	1.77	807	2.27

Table 5. The annotated statistics of protein-coding genes.

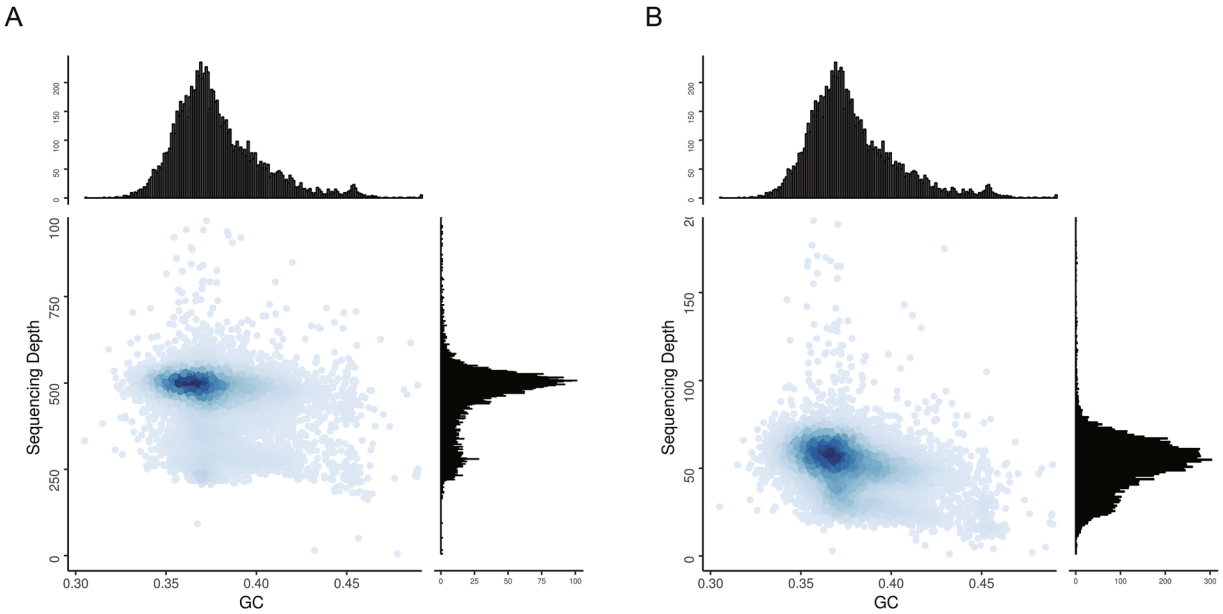


Fig. 6 GC Content and Sequencing Depth of short reads (**A**) and Pacbio HiFi reads (**B**). The x-axis represents the GC content; the y-axis represents the average depth.

and refinement of the Hi-C maps. After the above processes, a high-quality genome with a size of 435.45 Mb was obtained. The contig N50 and scaffold N50 were 49.83 Mb and 59.06 Mb, respectively, and approximately 422.19 Mb (96.96%) of the contig sequences were anchored to 7 chromosomes (Fig. 3).

Repeat annotation. Both homolog- and *de novo*-based methods were used to annotate repeat elements. Homolog repeat elements, including DNA transposons, long interspersed nuclear elements (LINE), short interspersed nuclear elements (SINE), long terminal repeats (LTRs) and satellites, were identified by RepeatMasker (v open-4.0.9) (http://www.repeatmasker.org) and RepeatProteinMask (v open-4.0.9) (http://www.repeatmasker.org) against the RepBase database (http://www.girinst.org/repbase) with default parameters. RepeatModeler (v open-1.0.11)[24] and LTR_FINDER_parallel (v1.0.7)[25] were used to build the *de novo* library with default parameters, and RepeatMasker (v open-4.0.9) was also applied to finish the annotation using the *de novo* library. Tandem Repeats Finder (v4.09)[26] was used to identify tandem repeats. Finally, the proportion of repeat elements in the genome was 50.10%, and approximately half of the repeat elements (25.33%) belonged to the LTRs subfamily (Fig. 4; Table 3; Supplementary Table 2).

Gene prediction and functional annotation. For the structural annotation of protein-coding genes, homology-, *de novo*- and transcriptome-based methods were applied and the integrated result was regarded as a final high-quality gene set. Firstly, five closely related species [*Arabidopsis thaliana*[27], *Salvia hispanica*[28],

Type	Assembly		Annotation	
	Number	Percentage(%)	Number	Percentage (%)
Complete BUSCOs (C)	1587	98.33	1585	98.2
Complete and single-copy BUSCOs (S)	1514	93.8	1513	93.7
Complete and duplicated BUSCOs (D)	73	4.52	72	4.5
Fragmented BUSCOs (F)	7	0.43	6	0.4
Missing BUSCOs (M)	20	1.24	23	1.4
Total BUSCO groups searched	1614	100	1614	100

Table 6. The BUSCO results of assembly and annotation.

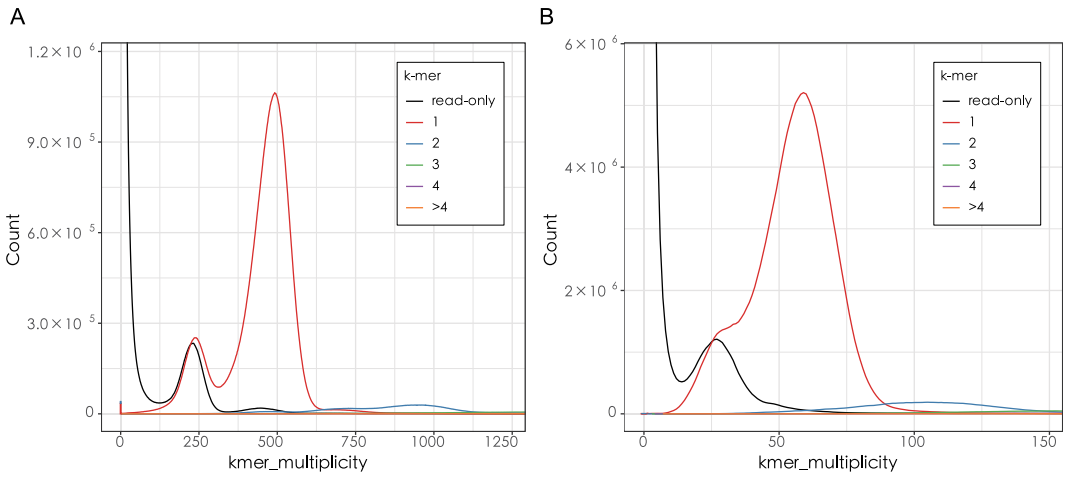


Fig. 7 Histogram of k-mer multiplicity collected from short reads (**A**) and Pacbio HiFi reads (**B**) The x-axis represents the depth of k-mer; the y-axis represents the count of k-mer.

Salvia miltiorrhiza[10], *Salvia splendens* (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_023119035.1/) and *Scutellaria baicalensis*[29], annotation files were downloaded from the National Center for Biotechnology Information (NCBI)] were used for homology-based prediction by Liftoff (v1.6.3)[30] and Exonerate (v2.2.0, parameters:–model protein2genome–showtargetgff1)[31]. Then the prediction tools GlimmerHMM (v3.0.4)[32] and AUGUSTUS (v3.3.2)[33] were applied to construct models for *de novo* prediction. Moreover, both of ISOseq and RNAseq data were used for transcriptome-based prediction. In RNAseq, 15 RNA samples from five tissues of *D. rupestre* were sequenced by DNBseq sequencing platform, then Hisat2 (v2.1.0)[34], Stringtie (v1.3.5)[35] and TransDecoder (v5.5.0) (https://github.com/TransDecoder/TransDecoder) with default parameters were employed to align and assemble the reads as well as identify the coding sequence. All the predictions from three methods were integrated by MAKER2 (v2.31.10)[36], and then the software HiCESAP was used to obtain the reliable gene set. Finally, a non-redundant and more complete gene set was generated containing 25,899 genes and 35,602 transcripts (Fig. 4; Table 4). In addition, tRNAscan-SE (v1.3.1)[37] and BLAST (v2.11.0+)[20] were used to annotate the transfer RNA (tRNA) and ribosomal RNA (rRNA), Rfam (v 14.8, parameters: cmscan–rfam–no hmmonly)[38] was used for the identification of small nuclear RNA (snRNA) and microRNA (miRNA). A total of 748 tRNA, 6899 rRNA, 555 snRNA and 103 miRNAs were identified (Supplementary Table 3).

Sequence similarity and domain/motif conservation were two basic methods for the functional annotation of protein-coding genes. The gene set of *D. rupestre* was firstly mapped against KEGG (v105.0, 2023-01-01)[39], Swiss-Prot (2023-03-01)[40], TrEMBL (2023-03-01)[40], InterPro (v93.0, 2023-03-02)[41], KOG (https://ftp.ncbi.nih.gov/pub/COG/KOG/, 2003-03-01), GO (2023-04-01)[42], Pfam (v37.2)[43], NR (NCBI Non-redundant protein, 2023-04-01)[44] and PlantTFDB (v5.0)[45] databases using diamond (v2.0.14, parameters:–evalue 1e-05)[46]. Meanwhile, the annotation information of KEGG was correlated with KEGG ORTHOLOGY and PATHWAY using KOBAS (v3.0, parameters: -t blastout:tab -s ko)[47]. Then the software InterProScan (v5.61-93.0, parameters:–seqtype p–formats TSV–goterms–pathways -dp)[48] acquired the information of conserved sequences from InterPro[41], and HMMER3 (v3.3.1)[49] was used to identify the motif and domain of *D. rupestre*. In total, 98.23% and 97.73% of the predicted genes and mRNAs were functionally annotated, respectively, and 16,644 common genes were identified by 5 databases (InterPro; GO; KEGG_KO; Swissprot; NR) (Fig. 5; Table 5).

Data Records

The data used for this paper had been deposited in National Center for Biotechnology Information (NCBI) with the BioProject accession number of PRJNA1108787, including 16 records of raw data[50] (DNA sequencing data from genome survey library: SRR29084948, DNA sequencing data from Hi-C library: SRR28956423-SRR28956429 and SRR28956435, DNA sequencing data from PacBio HiFi library: SRR28956436, RNAseq data from ISOseq library: SRR28956430, RNAseq data from RNAseq library: SRR28956422 and

SRR28956431- SRR28956434) in Sequence Read Archive (SRA). This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JBDOCY000000000[51]. The final genome assembly and annotation files are also available in Figshare database[52].

Technical Validation

Gene prediction and functional annotation. To assess genome correctness, short reads and PacBio HiFi reads were aligned to genome assembly using bwa (v0.7.12-r1039)[53] and minimap2 (v 2.24-r1122)[54]. The mapping rates of short and long reads were 99.41% and 99.96%, respectively. Besides, the coverages of short and long sequencing reads were 99.90% and 99.95%, respectively (Supplementary Table 4). Moreover, the association analysis of GC content and sequencing depth showed the uniformity of sequencing (Fig. 6). Then the genome completeness was firstly evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO, v 5.3.1, parameters: -m genome–limit 10)[55] with embryophyta_odb10 database (https://busco-data.ezlab.org/v5/data/lineages/), and 98.33% of expected conserved genes including 93.80% single-copy and 4.52% duplicated ones were identified as complete. In addition, 0.43% and 1.24% were identified as fragmented and missing, respectively (Table 6). The LTR assembly index (LAI) was also used to assess the completeness of genome via LTR_Finder (v1.07, parameters: -D 15000 -d 1000 -L 7000 -l 100 -p 20 -C -M 0.9) and LTR_retriever (v2.6), 17.21 of LAI meant the assembly result could be used as a reference genome[56,57]. The consensus quality value (QV) was calculated using Merqury (v1.3) for k-mer-based assembly validation[58]. The QV of short and long reads were 53.31 and 63.32, respectively, implied the high quality of genome assembly (Fig. 7; Supplementary Table 5). Moreover, the genome continuity was assessed by the number of contigs (217 contigs), N50 length of contigs and scaffold (49.83 Mb and 59.06 Mb, respectively), and the chromosome anchoring rate (96.96%). In conclusion, the result of all methods of evaluation showed high genome assembly completeness, quality and correctness.

Evaluation of the gene annotation. BUSCO analysis was also performed to evaluate the completeness of the gene set using the embryophyta_odb10 database, and 98.2% of expected conserved genes including 93.7% single-copy and 4.5% duplicated ones were identified as complete. Besides, there were only 0.4% and 1.4% identified as fragmented and missing, respectively. The assessment result of BUSCO showed a high degree of integrity for genome annotation (Table 6).

Code availability

No specific code or script was used in this work. Commands used for data analyses were all executed according to the manuals and protocols of the corresponding bioinformatics tools. Unless otherwise indicated, data processing was performed using default parameters.

Received: 14 June 2024; Accepted: 7 March 2025;
Published online: 21 March 2025

References

1. Wu, Z. Y. & Li, X. W. Flora Reipublicae Popularis Sinicae (Zhongguo Zhiwu Zhi), vol. 65. Beijing: Science Press; (1977).
2. Gao, J. *et al*. Metabolomic characterization of the chemical compositions of Dracocephalum rupestre Hance. *Food Res Int* **161**, 111871 (2022).
3. Ren, D. M., Qu, Z., Wang, X. N., Shi, J. & Lou, H. X. Simultaneous determination of nine major active compounds in Dracocephalum rupestre by HPLC. *J Pharm Biomed Anal* **48**(5), 1441–1445 (2008).
4. Wang, F., Dong, P. Z., Pei, X. P. & Lian, Y. L. Research progress of Dracocephalum rupestre Hance. *Chinese Journal of Ethnomedicine and Ethnopharmacy* **31**(22), 50–55 (2022).
5. Jing, W. J. & Ren, L. Shanxi province purpureaa identification and antioxidant activity of Mao Jiancao tea. *Shanxi Chemical Industry* **43**(12), 44–46+56 (2023).
6. Li, H. Q., Cao, Y. X. & Huang, F. J. Comparative study of DPPH · and ABTS+ · methods on the antioxidant synergistic effects of polyphenols, flavonoids and polysaccharides of Dracocephalum rupestre Hance. *Journal of Shanxi Agricultural University (Natural Science Edition)* **42**, 98–105 (2022).
7. Ding, C., YH, L. I. & Kang, H. J. Research review in chemical compositions and pharmacological effects of Dracocephalum rupestre Hance. *Journal of Pharmaceutical Research* **32**(11), 663–664 (2013).
8. Tian, X. P., Guo, Q. H., Sun, M. Y., LI, Q. & Lu, T. Comparison on bubil formation of Dracocephalum rupestre Hance originated from different provenances. *ACTA AGRESTIA SINICA* **29**(06), 1357–1362 (2021).
9. Zhao, Q. *et al*. The reference genome sequence of Scutellaria baicalensis provides insights into the evolution of Wogonin biosynthesis. *Mol Plant* **12**(7), 935–950 (2019).
10. Ma, Y. *et al*. Expansion within the CYP71D subfamily drives the heterocyclization of tanshinones synthesis in Salvia miltiorrhiza. *Nat Commun* **12**(1), 685 (2021).
11. Tian, X. M., Zhou, X. Y. & Gong, N. Applications of flow cytometry in plant research-Analysis of nuclear DNA content and ploidy level in plant cells. *Chinese Agricultural Science Bulletin* **27**(09), 21–27 (2011).
12. Wang, Y. *et al*. Operation Skills of Flow Cytometer for Detecting Nuclear DNA Contents in Higher Plant Cells. *Plant Science Journal* **33**(01), 126–131 (2015).
13. Ma, Q. Y. *et al*. Estimation of genome sizes of six Acer species by flow cytometry and K-mer analysis. *Journal of Nanjing Forestry University (Natural Sciences Edition)* **47**(01), 163–170 (2023).
14. Allen, G. C., Flores-Vergara, M. A., Krasynanski, S., Kumar, S. & Thompson, W. F. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nat Protoc* **1**(5), 2320–2325 (2006).
15. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**(17), i884–i890 (2018).
16. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**(6), 764–770 (2011).
17. 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**(1), 1432 (2020).
18. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**(2), 170–175 (2021).

19. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**(1), 460 (2018).
20. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**(3), 403–410 (1990).
21. Wingett, S. *et al*. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res* **4**, 1310 (2015).
22. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**(1), 95–98 (2016).
23. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**(1), 99–101 (2016).
24. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21**(Suppl 1), i351–358 (2005).
25. Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob DNA* **10**, 48 (2019).
26. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**(2), 573–580 (1999).
27. Sloan, D. B., Wu, Z. & Sharbrough, J. Correction of Persistent Errors in Arabidopsis Reference Mitochondrial Genomes. *Plant Cell* **30**(3), 525–527 (2018).
28. Jia, K. H. *et al*. Chromosome-scale assembly and evolution of the tetraploid Salvia splendens (Lamiaceae) genome. *Hortic Res* **8**(1), 177 (2021).
29. Pei, T. *et al*. Gap-free genome assembly and CYP450 gene family analysis reveal the biosynthesis of anthocyanins in Scutellaria baicalensis. *Hortic Res* **10**(12), uhad235 (2023).
30. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**(12), 1639–1643 (2021).
31. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
32. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**(16), 2878–2879 (2004).
33. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**(Web Server issue), W435–439 (2006).
34. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**(8), 907–915 (2019).
35. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**(3), 290–295 (2015).
36. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
37. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**(5), 955–964 (1997).
38. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33**(Database issue), D121–124 (2005).
39. Kanehisa, M. Goto S: KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**(1), 27–30 (2000).
40. Coudert, E. *et al*. Annotation of biologically relevant ligands in UniProtKB using ChEBI. *Bioinformatics* **39**(1) (2023).
41. Paysan-Lafosse, T. *et al*. InterPro in 2022. *Nucleic Acids Res* **51**(D1), D418–d427 (2023).
42. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**(1), 25–29 (2000).
43. Finn, R. D. *et al*. Pfam: the protein families database. *Nucleic Acids Res* **42**(Database issue), D222–230 (2014).
44. Sayers, E. W. *et al*. Database resources of the national center for biotechnology information. *Nucleic Acids Res* **50**(D1), D20–d26 (2022).
45. Tian, F., Yang, D. C., Meng, Y. Q., Jin, J. & Gao, G. PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Res* **48**(D1), D1104–d1113 (2020).
46. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* **18**(4), 366–368 (2021).
47. Bu, D. *et al*. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res* **49**(W1), W317–w325 (2021).
48. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**(9), 1236–1240 (2014).
49. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res* **41**(12), e121 (2013).
50. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP506236 (2024).
51. Dracocephalum rupestre cultivar YanQinglan, whole genome shotgun sequencing project. GenBank https://identifiers.org/ncbi/insdc.gca:GCA_040181865.1.
52. Zhao, Q., Fan, Z. X., Yu, H. & Wang, Z. L. Chromosome-level genome assembly of *Dracocephalum rupestre* Hance. *Figshare* https://doi.org/10.6084/m9.figshare.25778184 (2024).
53. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14), 1754–1760 (2009).
54. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**(18), 3094–3100 (2018).
55. 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**(19), 3210–3212 (2015).
56. Ou, S. & Jiang, N. LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol* **176**, 1410–1422 (2018).
57. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res* **46**, e126 (2018).
58. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245 (2020).

Acknowledgements

This work was supported by grants from National Natural Science Foundation of China (82260092).

Author contributions

Wang Z.L., Yu H. and Zhao Q. conceived and designed the study; Zhao Q. and Fan Z.X. collected and prepared the samples; Zhao Q. and Fan Z.X. performed bioinformatics analysis; Zhao Q. wrote the manuscript and then Wang Z.L. and Yu H. edited and revised the writing with significant contributions. All authors read and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04778-3.

Correspondence and requests for materials should be addressed to H.Y. or Z.W.