Hybrid de novo genome assembly of the Chinese herbal fleabane

*Erigeron breviscapus*

Jing Yang¹, Guanghui Zhang²,⁹, Jing Zhang³,⁹, Hui Liu⁴,⁵, Wei Chen¹,⁶, Xiao Wang⁴,⁵, Yahe Li⁷, Yang Dong¹,⁶,⁸,¹⁰, Shengchao Yang²,¹⁰

¹Biological Big Data College, Yunnan Agricultural University, Kunming 650201, China.
²National-Local Joint Engineering Research Center on Germplasm Utilization and Innovation of Chinese Medicinal Materials in Southwest China, Yunnan Agricultural University, Kunming 650201, China.
³NOWBIO Technology Co. Ltd, Kunming 650202, China.
⁴State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China.
⁵University of Chinese Academy of Sciences, Beijing 100049, China.
⁶Yunnan Research Institute for Local Plateau Agriculture and Industry, Kunming 650201, China.
⁷Longjing Pharmaceutical Co. Ltd, Kunming 650228, China.
⁸College of Life Science, Kunming University of Science and Technology, Kunming, China.
⁹Co-first authors.
¹⁰Corresponding authors.

**Abstract**

**Background:** The plants in the *Erigeron* genus of the Compositae (Asteraceae) family are commonly called fleabanes, possibly due to the belief that certain
chemicals in these plants repel fleas. In the traditional Chinese medicine, *Erigeron brevisscapus*, which is native to China, was widely used in the treatment of cerebrovascular disease. A handful of bioactive compounds, including scutellarin, 3,5-dicaffeoylquinic acid, and 3,4-dicaffeoylquinic acid, have been isolated from the plant. With the purpose of finding novel medicinal compounds and understanding their biosynthetic pathways, we propose to sequence the genome of *E. brevisscapus*.

**Findings:** We assembled the highly heterozygous *E. brevisscapus* genome using a combination of PacBio single-molecular real-time sequencing method and next-generation sequencing method on the Illumina HiSeq platform. The final draft genome is approximately 1.2 Gb, with the contig and scaffold N50 sizes of 18.8 kb and 31.5 kb, respectively. Further analyses predicted 70,214 protein-coding genes in the *E. brevisscapus* genome, and 9,825 shared gene families among Compositae species.

**Conclusions:** The *E. brevisscapus* genome provides a valuable resource for the investigation of novel bioactive compounds in this Chinese herb.

**Keywords:** *Erigeron brevisscapus*, Illumina sequencing, PacBio sequencing.

**Background**

*Erigeron brevisscapus* (also known as *dengzhanhua* in Chinese) is a perennial flower in the *Erigeron* genus of the Compositae (Asteraceae) family. Its flower head is
comprised of yellow disk florets and multiple surrounding blue to purple ray florets (Fig. 1). This species is endemic to Southwestern China, which grows in mid-altitude mountains, subalpine open slopes, grasslands and forest margins from 1000 m to 3500 m [1,2]. In the traditional Chinese medicine, *E. breviscapus* is believed to improve blood circulation and ameliorate platelet coagulation [3,4]. Since the 1980s, the herbal extracts and bioactive compounds from *E. breviscapus* have been widely used for the treatment of cerebral embolism and its complications, cerebral thrombosis, coronary heart disease, angina pectoris, acute renal failure, and nephritic syndrome [5]. At present, more than 1,000 tons of dry *E. breviscapus* are collected and used in the pharmaceutical industry each year, greatly exhausting the wild resources of this species [6,7]. In this study, we report the draft genome assembly of *E. breviscapus*.

Because of the high heterozygosity of the *E. breviscapus* genome, we adopted both Illumina sequencing and PacBio single-molecular real-time sequencing in the assembly procedure.

### Data description

**Whole-genome shotgun sequencing of *E. breviscapus* on Illumina platform**

*E. breviscapus* seedlings were provided by the Longjing Pharmaceutical Co. Ltd and maintained in a greenhouse at the Yunnan Agricultural University. Genomic DNA was extracted from the leaf tissues of a single *E. breviscapus* plant using the GenElute™ Plant Genomic DNA Miniprep Kit (Sigma-Aldrich; St. Louis, USA). Paired-end
libraries with insert sizes ranging from 150 bp to 800 bp were constructed using NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB, USA), and mate pair libraries with insert sizes from 2 kb to 20 kb were constructed using Illumina Nextera Mate Pair Library Preparation Kit (Illumina, USA). All constructed libraries were sequenced on a HiSeq 2500 platform (Illumina, USA) using either a PE-100 or PE-90 module (Additional file 1: Table S1). In total, about ~413.4 Gb raw data were generated on the Illumina platform. The raw data were initially filtered by removing reads with more than 10 % N or more than 40 bp low quality bases. Next, redundant reads resulting in duplicate base calls were filtered at a threshold of Euclidean distance ≤ 3 and mismatch rate of ≤ 0.1. Only one copy of any duplicated paired-end reads was retained. Finally, both read 1 and read 2 were removed if they contained an adapter ≥ 10 bp with a mismatch rate ≤ 0.1. This process yielded ~275.1 Gb of clean data for the de novo assembly of the E. breviscapus genome (Additional file 1: Table S1).

Single-molecule real-time sequencing of long reads on PacBio platform

Single-molecule real-time (SMRT) sequencing of long reads on a PacBio RS II platform (Pacific Biosciences, USA) was used to assist the subsequent de novo genome assembly process [8]. In brief, 40 μg of sheared DNA was used to construct 26 SMRT Cell libraries with an insert size of 17 kb. These libraries were sequenced in 105 SMRT DNA sequencing cells using the P6 polymerase/C4 chemistry combination,
and a data collection time of 240 min per cell. The sequencing produced about 62.4 Gb clean data, consisting of 6,802,553 reads with an average read length of 9,175 bp (Additional file 1: Table S1).

Estimation of the *E. breviscapus* genome size

The genome size of *E. breviscapus* was estimated by flow cytometry, using *Oryza sativa* Nipponbare as internal standard and propidium iodide as the stain. The result showed that the genome size of *E. breviscapus* was approximately 1.52 Gb (Additional file 1: Figure S1).

Hybrid de novo genome assembly of *E. breviscapus*

A hybrid genome assembly pipeline was used to overcome challenges posed by the heterozygous *E. breviscapus* genome (Fig. 2). HiSeq reads were first assembled using MaSuRCA [9] with default parameters, and also using Platanus [10] with parameters “-m 500 -k 43 -s 5 -d 0.3 -u 0.15 -c 3”, resulting in two contig assemblies. The Platanus-generated contigs, together with PacBio reads, were used to generate a third contig assembly by DBG2OLC with default parameters [11]. The three different contig assemblies were merged together by Minimus2 using default parameters [12]. To eliminate possible errors of the merged contig assembly, Bowtie2 [13] was used to align HiSeq reads back to this assembly. The result was further polished by PICARD and GATK using default parameters [14,15]. Polished contigs were then used to build
scaffolds using OPERA [16] with a k-mer of 39. This process yielded a final draft E. breviscapus genome of 1.2 Gb, with a contig N50 size of 18.8 kb and a scaffold N50 size of 31.5 kb (Additional file 1: Table S2).

Evaluation of the completeness of the E. breviscapus genome assembly

We evaluated the completeness of the final assembly using CEGMA [17] with a set of 248 ultra-conserved core eukaryotic genes and BUSCO [18] with the Embryophyta gene set. CEGMA assessment showed that our assembly captured 240 (96.9 %) of the 248 ultra-conserved core eukaryotic genes, of which 217 (87.5 %) were complete (Table 1). BUSCO analysis showed that 80.6 % and 63.3 % of the 1440 expected embryophytic genes were identified as complete and fragmented, respectively (Table 2).

Transcriptome sequencing

Total RNA was extracted from the leaf, root, stem, and flower tissues of a cultivated E. breviscapus individual using Qiagen RNeasy Plant Mini Kits. Additional RNA samples of the leaf tissues were acquired from six more cultivated individuals and five wild individuals (Additional file 1: Table S3). All cultivated samples were acquired from the greenhouse and all wild samples were collected from Dali, Yunnan Province. Total RNA-seq libraries were prepared using TruSeq RNA Library Preparation Kit v2 (Illumina, CA, USA) according to the manufacturer’s instructions.
and subsequently sequenced on the HiSeq 2500 platform. In total, about 1.1 billion RNA-seq reads were obtained, representing ~117.6 Gb raw data. We aligned all the RNA-seq reads back to the *E. breviscapus* genome assembly using TopHat [19] with default parameters (Additional file 1: Table S3). The percentage of aligned reads ranged from 60.6 % for the root to 80.9 % for the leaf. The FPKM value was calculated for each protein-coding gene by Cufflinks using default parameters. FPKM >0.05 was used as the cutoff value to identify expressed genes.

**Repeat annotation of the *E. breviscapus* genome assembly**

The *E. breviscapus* genome was searched for tandem repeats using the Tandem Repeat Finder [20]. RepeatMasker and RepeatProteinMasker [21] were used against Repbase library [22] to identify known transposable element repeats. *De novo* evolved transposable element annotation was performed using RepeatModeler [21] and LTR FINDER [23]. The combined results show that the total length of repeated sequences is about 664.2 Mb, accounting for ~54.58 % of the *E. breviscapus* genome assembly (Additional file 1: Table S4 and S5).

**Gene prediction**

We used multiple methods to annotate protein-coding genes in the *E. breviscapus* genome, including homology-based predictions, *de novo* predictions, and transcriptome-based predictions. For homology-based predictions, protein sequences...
of *Arabidopsis thaliana, Fragaria vesca, Malus domestica, Oryza sativa, Prunus persica* and *Vitis vinifera* were obtained from Phytozome v9.1 (http://www.phytozome.net/), *Pyrus communis* from Genome Database for Rosaceae (https://www.rosaceae.org), and *Prunus mume* from NCBI (ftp://ftp.ncbi.nih.gov/genomes/Prunus_mume). First, query sequences were subjected to TBLASTN analysis with a cutoff E-value of 1 e−5. BLAST hits corresponding to reference proteins were concatenated by Solar [24] (The Beijing Genomics Institute BGI) development) after low-quality records were removed. The genomic sequence of each reference protein was extended upstream and downstream by 2,000 bp to represent a protein-coding region. GeneWise [25] was used to predict gene structure contained in each protein region. For *de novo* predictions, AUGUSTUS [26], GENSCAN [27] and glimmerHMM [28] analyses were performed on the repeat-masked genome, with parameters trained from *A. thaliana*. For transcriptome-based predictions, RNA-seq data from the leaves of six cultivated individuals were used for gene annotation, processed by Tophat and Cufflinks [19]. The homology, *de novo* and transcriptomic-based predicted gene sets were merged to form a comprehensive and non-redundant reference gene set using EVidenceModeler [29]. Our analysis indicates that the *E. breviscapus* genome contains 70,214 protein-coding genes with an average CDS length of 839 bp (Additional file 1: Table S6).
Non-coding RNA annotation

tRNAscan-SE [30] with default parameters for eukaryotes was used for tRNA annotation. Homology-based rRNA annotation was performed by mapping plant rRNAs to the *E. breviscapus* genome using BLASTN with parameters of “E-value = 1e⁻⁵”. miRNA and snRNA genes were predicted by INFERNAL [31] using the Rfam database (release 11.0) [32]. The final results include 504 miRNAs, 751 tRNAs, 159 rRNAs, and 385 snRNAs (Additional file 1: Table S7).

Gene family clustering analysis

To identify and estimate the number of potential orthologous gene families between *E. breviscapus*, *Helianthus annuus*, *Cynara cardunculus*, *Solanum tuberosum*, *Solanum lycopersicum*, *V. vinifera*, and *O. sativa*, we applied the OrthoMCL pipeline [33] using standard settings (BLASTP E-value < 1e⁻⁵) to compute the all-against-all similarities. Gene sequences from *S. tuberosum*, *S. lycopersicum*, *V. vinifera*, and *O. sativa* were downloaded from Phytozome v11.0. Gene sequences from *H. annuus* and *C. cardunculus* were downloaded from Sunflower Genome Database (http://www.sunflowergenome.org) and Globe artichoke GBrowse (http://gviewer.ge.ucdavis.edu/cgi-bin/gbrowse/Artichoke_v1_1), respectively.

Among the total 19,565 *E. breviscapus* gene families, 5,501 (28.1%) appear to be lineage specific. There are 9,825 (50.2%) gene families shared among Compositae species including *E. breviscapus*, *H. annuus*, and *C. cardunculus*. In addition, *E.
*breviscapus* shared 7,957 (40.7%) gene families with *S. tuberosum* (Fig. 3).

**Phylogenetic Tree Construction and Divergence Time Estimation**

All 413 single-copy orthologous genes identified in the gene family clustering analysis from the *S. lycopersicum, V. vinifera, O. sativa, E. breviscapus, H. annuus, C. cardunculus*, and *S. tuberosum* were used to construct a phylogenetic tree.

Orthologous genes from the seven species were aligned using MUSCLE with default settings [32] for each gene. Four-fold degenerate sites were extracted from each gene and concatenated into a “super gene” for each species. PhyML [35] was used to reconstruct phylogenetic trees between species. We implemented a Monte Carlo Markov chain (MCMC) algorithm for the estimation of divergence times using the program MCMCtree from the PAML package [36]. The result showed that *E. breviscapus* shared a closer phylogenetic relationship with *H. annuus* than *C. cardunculus* in the Compositae family (Additional file 1: Figure S2). The estimated divergence time was 29.4 million years ago between *E. breviscapus* and *H. annuus* (Additional file 1: Figure S3).

**Expansion and Contraction of Gene Families**

CAFE [37] is a tool for analyzing the evolution of gene family size based on the stochastic birth and death model. With the calculated phylogeny and the divergence time, this software was applied to identify gene families that had undergone
expansion and/or contraction in *S. lycopersicum*, *V. vinifera*, *O. sativa*, *E. breviscapus*, *H. annuus*, *C. cardunculus*, and *S. tuberosum* with the parameters “p-value = 0.05, number of threads = 10, number of random = 1000, and search for lambda”. We identified 10,845 expanded gene families in the *E. breviscapus* genome, which is more than that in two other species *C. cardunculus* (1,059) and *H. annuus* (3,480) in *Compositae* (Additional file 1: Figure S4).

In summary, we reported the genome sequencing, assembly, annotation, and evolution analysis of the *E. breviscapus*. This genome assembly will provide a valuable resource for studying the biosynthetic pathways of the medicinal components in *E. breviscapus*. This information will also help find novel bioactive compounds, and improve the molecular breeding of this medicinal herb.

**Availability of supporting data**

Sequencing reads of each sequencing library and RNA-seq data have been deposited at NCBI with the project ID PRJNA352312. Supporting data are also available in the GigaScience database, GigaDB [38]. All supplementary figures and tables are provided in Additional file 1.

**Additional file**

**Additional file 1: Supplemental tables and figures. Table S1.** Raw sequencing
statistics from the Illumina platform and PacBio platform. **Table S2.** Summary of genome assembly. **Table S3.** Summary of transcriptomes. **Table S4.** Statistics of repeats in the *E. breviscapus* genome. **Table S5.** Repeat annotation of the *E. breviscapus* genome assembly. **Table S6.** Gene annotation statistics for the *E. breviscapus* genome. **Table S7.** Summary of non-protein-coding gene annotation in the *E. breviscapus* genome assembly. **Figure S1.** The estimated genome size of *E. breviscapus* with flow cytometry. **Figure S2.** Phylogenetic reconstruction of the *E. breviscapus* and six other plant species. **Figure S3.** Divergence time estimation of the *E. breviscapus* and six other plant species. **Figure S4.** Gene family expansions and contractions in the *E. breviscapus*.

**Abbreviations**

CDS: Coding DNA sequence; NCBI: National Center for Biotechnology Information; CEGMA: Core Eukaryotic Genes Mapping Approach; BUSCO: Benchmarking Universal Single-Copy Orthologs.

**Funding**

This work was support by the pilot project for establishing new socialized service system by agricultural science and education combination in Yunnan Province (Medical Plant Unit) (2014NG003) and National Natural Science Foundation of China (81260614).
Competing interests

The authors declare that they have no competing interests.

Authors’ contributions

WC, YD, GZ and SY designed the study. HL assembled the genome. JY, JZ analyzed the data. JY, WC and YD wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Longjing Pharmaceutical Co. Ltd for providing samples of *E. breviscapus* plant.

Author details

1 Biological Big Data College, Yunnan Agricultural University, Kunming 650201, China. 2 National-Local Joint Engineering Research Center on Germplasm Utilization and Innovation of Chinese Medicinal Materials in Southwest China, Yunnan Agricultural University, Kunming 650201, China. 3 NOWBIO Technology Co. Ltd, Kunming 650202, China. 4 State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China. 5 University of Chinese Academy of Sciences, Beijing 100049, China. 6 Yunnan Research Institute for Local Plateau Agriculture and Industry, Kunming 650201, China. 7 Longjing Pharmaceutical Co. Ltd, Kunming 650228, China. 8 College of Life Science, Kunming University of Science and Technology, Kunming, China. 9 Co-first authors. 10 Corresponding authors.
References


17. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core

18. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO:
assessing genome assembly and annotation completeness with single-copy

Differential gene and transcript expression analysis of RNA-seq experiments with

20. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic

21. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements

Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome

23. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length

genome assembly of domesticated apple (Malus x domestica). GigaScience,
2016;5:35.


Table 1 Statistics of the completeness of the hybrid *de novo* assembly genome of *E. breviscapus* by CEGMA.

<table>
<thead>
<tr>
<th>Group</th>
<th>Protein Num.</th>
<th>Completeness (%)</th>
<th>Total Num.</th>
<th>Average Ortholog (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>217</td>
<td>87.50</td>
<td>633</td>
<td>2.92</td>
</tr>
<tr>
<td>Group1</td>
<td>58</td>
<td>87.88</td>
<td>158</td>
<td>2.72</td>
</tr>
<tr>
<td>Group2</td>
<td>49</td>
<td>87.50</td>
<td>126</td>
<td>2.57</td>
</tr>
<tr>
<td>Group3</td>
<td>53</td>
<td>86.89</td>
<td>171</td>
<td>3.23</td>
</tr>
<tr>
<td>Group4</td>
<td>57</td>
<td>87.69</td>
<td>178</td>
<td>3.12</td>
</tr>
<tr>
<td>Partial</td>
<td>240</td>
<td>96.77</td>
<td>856</td>
<td>3.57</td>
</tr>
<tr>
<td>Group1</td>
<td>63</td>
<td>95.45</td>
<td>206</td>
<td>3.27</td>
</tr>
<tr>
<td>Group2</td>
<td>55</td>
<td>98.21</td>
<td>185</td>
<td>3.36</td>
</tr>
<tr>
<td>Group3</td>
<td>59</td>
<td>96.72</td>
<td>232</td>
<td>3.93</td>
</tr>
<tr>
<td>Group4</td>
<td>63</td>
<td>96.92</td>
<td>233</td>
<td>3.70</td>
</tr>
</tbody>
</table>

*Protein Num.: Number of 248 ultra-conserved core eukaryotic genes (CEGs) present in the *E. breviscapus* genome.*

*Completeness (%): Percentage of 248 ultra-conserved CEGs present in the *E. breviscapus* genome.*

*Total Num.: Total number of CEGs including putative orthologs present in the *E. breviscapus* genome.*

*Average Num.: Average number of orthologs per CEG.*
Table 2 Statistics of the completeness of the hybrid de novo assembly genome of *E. breviscapus* by BUSCO.

<table>
<thead>
<tr>
<th>BUSCO benchmark</th>
<th>Number</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total BUSCO groups searched</td>
<td>1440</td>
<td>-</td>
</tr>
<tr>
<td>Complete BUSCOs</td>
<td>1161</td>
<td>80.63</td>
</tr>
<tr>
<td>Complete and single-copy BUSCOs</td>
<td>635</td>
<td>44.10</td>
</tr>
<tr>
<td>Complete and duplicated BUSCOs</td>
<td>526</td>
<td>36.53</td>
</tr>
<tr>
<td>Fragmented BUSCOs</td>
<td>90</td>
<td>6.25</td>
</tr>
<tr>
<td>Missing BUSCOs</td>
<td>189</td>
<td>13.13</td>
</tr>
</tbody>
</table>

Figure Legend

**Fig. 1** Example of the *E. breviscapus* (image from Shengchao Yang).

**Fig. 2** Assembly pipeline for the *E. breviscapus* genome.

**Fig. 3** Venn diagram showing unique and shared gene families among four sequenced dicotyledonous species.
HiSeq Reads

- MaSuRCA Assembly
- Platanus Assembly
- Contigs
- Minimus2 Merge
- Contigs
- OperA Assembly
- Final Assembly

PacBio Reads

- DBG2OLC Assembly
- Contigs

Bowtie2

- PICARD GATK Corrected
- Polished Contigs

Figure 2
Click here to access/download 
**Supplementary Material** 
Supplementary Figure 3.pdf
Click here to access/download
**Supplementary Material**
Supplementary Material.docx