High quality assembly of the reference genome for scarlet sage, *Salvia splendens*, an economically important ornamental plant

**Manuscript Number:** GIGA-D-18-00028R1

**Full Title:** High quality assembly of the reference genome for scarlet sage, *Salvia splendens*, an economically important ornamental plant

**Article Type:** Data Note

**Funding Information:**
- Beijing Key Laboratory of Greening Plants Breeding (Z201605) Dr. Hai-Bo Xin
- Fundamental Research Funds for the Central Universities (YX2013-41) Mr. Jian-Feng Mao

**Abstract:**
Background: *Salvia splendens* Ker-Gawler, scarlet or tropical sage, is a tender herbaceous perennial widely introduced and seen in public gardens all over the world. With few molecular resources, breeding is still restricted to traditional phenotypic selection, and the genetic mechanisms underlying phenotypic variation still remain unknown. Hence, a high quality reference genome will be very valuable for marker assisted breeding, genome editing or molecular genetics.

Findings: We generated 66 gigabases (Gb) and 37 Gb of raw DNA sequences, respectively, from whole-genome sequencing of a largely homozygous scarlet sage inbred line using PacBio Single-Molecule Real-Time (SMRT) and Illumina HiSeq sequencing platforms. PacBio de novo assembly yielded a final genome with a scaffold N50 size of 3.12 megabases (Mb), and a total length of 808 Mb. The repetitive sequences identified accounted for 57.52% of the genome sequence and 54,008 protein-coding genes were predicted collectively with ab initio and homology-based gene prediction from the masked genome. The divergence time between *S. splendens* and *S. miltiorrhiza* was estimated with 28.21 million years ago (Mya). Moreover, 3,797 species-specific genes and 1,187 expanded gene families were identified for the scarlet sage genome.

Conclusions: We provide the first genome sequence and gene annotation for the scarlet sage. The availability of these resources will be of great importance for further breeding strategies, genome editing and also for comparative genomics among related species.

**Corresponding Author:** Ri-Chen Cong, Ph.D
Beijing Institute of Landscape Architecture
Beijing, CHINA

**Order of Authors:**
- Ai-Xiang Dong
- Hai-Bo Xin
- Zi-Jing Li
- Hui Liu
- Yan-Qiang Sun
- Shuai Nie
We provide our comments directly underneath the points raised by you and within the three reviewers’ reports as follows:

AE: Please pay particular attention to reviewer #2's comment number 3: "Since the genomes of Salvia miltiorrhiza (Zhang et al. and Xu et al.) and Mentha longifolia have been published, a more detailed analysis about differences between Salvia splendens and the other two plants should be conducted, so as to highlight the importance of Salvia splendens."

R: We provided synteny analyses among detected metabolic gene cluster between the Salvia genomes. One section of comparative genomics was added (also see Figure S7 for synteny blocks). However, even mentha genome has been published, its gene annotation data are not publicly available. We wrote two emails to the corresponding authors for two times, we did not get any response. So mentha genome was not included in our comparative genomic studies.

AE: Your manuscript is under consideration as a Data Note, and although we do not require in-depth exploration of biological questions for this article type, I fully agree with the referee that it is crucially important that you provide some detailed context regarding the other published Salvia and Mentha genomes - what are similarities and differences, and what are unique features of Salvia splendens.

R: please see answer provided above.

AE: Please also clarify a number of technical issues mentioned by reviewer 3, e.g. regarding your scaffolding approach, as well as the use of Pilon and BUSCO.

R: Please see answers to Reviewer 3.

AE: As an editorial point, I notice that you indicate 4 "equally contributing" first authors. Please note that we allow a maximum of 3 co-first authors (and only if their contributions are really absolutely equal). Please revise the author role indications accordingly.

R: Revised. Now we have 3 co-first authors.

Reviewer #1: The authors of "High quality assembly of the reference genome for scarlet sage, Salvia splendens, an economically important ornamental plant" describe their efforts in generating a reference sequence for the plant Salvia splendens that is spread out in multiple gardens. Overall the authors relied mainly on Pacbio to obtain a high quality reference genome sequence using state of the art methods. Furthermore, they annotated the genome using RNA-Seq reads and state of the art methods such as maker, Augustus etc. Thus, I don't have any comments or concerns.

R: Thank you.

Reviewer #2: This manuscript described the construction of genome sequence and annotation for Salvia splendens Ker-Gawler. A hybrid approach using PacBio Single-Molecule Real-Time (SMRT) and Illumina HiSeq sequencing platforms was employed. Finally, a genome of 808Mb and 54,008 protein-coding genes were reported. The genome should be pretty completed because 1) the genome size is already bigger than the k-mer estimated genome size; 2) supported by BUSCO results and 3) satisfactory N50 and contig / scaffold number. However, this is not the first species of...
the same genus and more functional information should be included to improve the novelty and usefulness of this piece of work. Otherwise, this will be only another genome sequence deposited in the database.

R: Thank you. Regarding more functional information provision from genomic data, please see our comments immediately below.

Reviewer #2: Comments and suggestions:
2.1. As mentioned in the introduction, many species of this genus are extensively used for culinary purposes, essential oil production and Chinese herbal remedies. Therefore, it is expected that the active ingredients of the plant responsible for its biological and therapeutic functions should be quite well known. If the metabolic pathways responsible for the production of these ingredients could be dissected, the information reported could be more useful for researchers working on this plant species.

R: One section (lines 284-332) involving description and analysis of metabolic pathways, gene clusters and comparative genomics was added. Two pathways of flavonoid and menthol biosynthesis were constructed by homolog mapping with the help of the Plant Metabolic Network (PMN v12.5, https://www.plantcyc.org/). Results were summarized in Figure S5 and S6, Supplementary_File_1.

2.2. Regarding the transcriptome analysis, results had been generated using tissues obtained from roots, shoots, leaves, calyxes and corollas. For gene discovery, mixing all the datasets to generate the transcript set is reasonable. However, to highlight the therapeutic value of particular part(s) of the plant, differential expression analysis and gene clustering would be expected.

R: Yes, this true. Our immediate intention was to identify the overall metabolic gene clusters for the two Salvia genomes, and related gene co-expression profiles were further examined among the co-localized genes. These gene clusters were summarized in Table S13, and genomic composition of gene clusters and gene expression were detailed in Supplementary File 2 and 3 (lines 284-321). A follow-up study could now target more specifically the genes of interest that promise to be correlated with variation in the therapeutic value of certain compounds and in the different plant parts and confidently identify those with the highest value.

2.3. Since the genomes of Salvia miltiorrhiza (Zhang et al. and Xu et al.) and Mentha longifolia have been published, a more detailed analysis about differences between Salvia splendens and the other two plants should be conducted, so as to highlight the importance of Salvia splendens. Moreover, the functional significance of such differences should be extensively explored and discussed. Finally, certain experiments should be done if necessary.

R: We provided synteny analyses among the detected metabolic gene cluster between the Salvia genomes. One section (lines 284-332) of comparative genomics was added to our manuscript (also see Figure S7). However, even though the mentha genome has been published, curiously, its gene annotation data is not publicly available! We wrote two emails to the corresponding authors, but we did not get any response. Thus, at this time, unfortunately, the mentha genome could not be included in our comparative genomic studies.

Reviewer #3: Dong et al. provide a near complete reference genome for the ornamental crop Salvia splendens using a PacBio sequencing approach. The assembly is high quality and will be useful for the plant comparative genomics community. The approaches are technically sound and adequate details on the assembly and annotation of this genome are provided. I have a few minor concerns I feel should be addressed before this manuscript is published.

R: Thanks.

Reviewer #3: Comments and suggestions:
3.1 The assembly metrics of the Salvia genome are exceptionally good and the near completeness of this assembly will make it useful for the comparative genomics community. The scaffolding is potentially problematic given the short read lengths of the Illumina data and the lack of an additional set of PacBio data that was not utilized in the initial assembly. The authors used 4-5 different scaffolding algorithms on the same datasets, potentially introducing errors. Most of these scaffolding and gap filling programs were designed to utilize mate pair data to bridge repeats and not the short insert libraries produced by the authors. The Illumina data could falsely bridge gaps
creating chimeric, misassembled scaffolds. R: Indeed, we used two sets of PacBio reads from two individual plants, and just one set of Illumina reads. Genome assembly was processed in two main steps in this study as follows: We firstly generated the primary assemblies with different algorithms based on one set of PacBio reads. Then, the other set of PacBio reads was utilized in a further scaffolding step starting from the best assembly from the primary step. We provided a detailed description for genome assembly in this revision now to avoid ambiguity in the method description. We were trying to explore extra information from the Illumina short reads in the second scaffolding step, while taking care of the potential false bridge. In fact, Illumina did provide us only few values.

3.3 Line 162. The aligner used to map the Illumina reads to the Salvia genome for Pilon based polishing should be provided. Parameters for Pilon and the number of corrected indels/SNPs should also be listed. R: Yes, we did it. Pl. see lines 164-170.

3.4 Line 216 and Line 225: It is unclear why two different BUSCO datasets were used to verify the completeness of the genome assembly/annotation. R: We assured that only one BUSCO dataset (1,440 single copy orthologs of the Viridiplantae database) was used in this study. We wrongly input the description for BUSCO dataset. Now we corrected it throughout the text.

3.5 It would be interesting to include more downstream comparative genomics analyses for this species, but I suspect this is beyond the scope of this manuscript. R: We did further provide functional analyses according to the second reviewer. However, no real comparative genomic analyses were provided as published genomes of Salvia miltiorrhiza (Zhang et al. and Xu et al.) and Mentha longifolia are really low quality or no protein annotation has yet been released which prevented further comparative study.

Additional Information:

<table>
<thead>
<tr>
<th>Question</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are you submitting this manuscript to a special series or article collection?</td>
<td>No</td>
</tr>
</tbody>
</table>

**Experimental design and statistics**

Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends.

Have you included all the information requested in your manuscript? | Yes |

**Resources**

A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. | Yes |
Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?

<table>
<thead>
<tr>
<th>Availability of data and materials</th>
<th>Yes</th>
</tr>
</thead>
</table>

All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in [publicly available repositories](#) (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.

Have you met the above requirement as detailed in our Minimum Standards Reporting Checklist?
High quality assembly of the reference genome for scarlet sage, *Salvia splendens*, an economically important ornamental plant

Ai-Xiang Dong1‡, Hai-Bo Xin1,2‡, Zi-Jing Li1‡, Hui Liu2, Yan-Qiang Sun2, Shuai Nie2, Zheng-Nan Zhao1, Rong-Feng Cui1, Ren-Gang Zhang3, Quan-Zheng Yun3, Xin-Ning Wang3, Fatemeh Maghuly4, Ilga Porth5, Ri-Chen Cong1*, Jian-Feng Mao2*

1 Beijing Key Laboratory of Greening Plants Breeding, Beijing Institute of Landscape Architecture, Beijing, 10020, China.
2 Beijing Advanced Innovation Center for Tree Breeding by Molecular Design, National Engineering Laboratory for Tree Breeding, Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants, Ministry of Education, College of Biological Sciences and Technology, Beijing Forestry University, Beijing, 100083, China.
3 Beijing Ori-Gene Science and Technology Co. Ltd, Beijing, 10226, China.
4 Plant Biotechnology Unit (PBU), Dept. Biotechnology, BOKU-VIBT, University of Natural Resources and Life Sciences, Muthgasse 18, 1190 Vienna, Austria. ORCID: 0000-0001-5433-0070.
5 Département des sciences du bois et de la forêt, Pavillon Charles-Eugène-Marchand, 1030, Avenue de la Médecine, Université Laval, Québec (Québec) G1V 0A6, Canada. ORCID: 0000-0002-9344-6348
‡These authors contributed equally to this paper.

*Correspondence to: hardhopeee@163.com (RCC), ORCID: 0000-0002-4619-6120;
jianfeng.mao@bjfu.edu.cn (JFM), ORCID: 0000-0001-9735-8516

Abstract

Background: *Salvia splendens* Ker-Gawler, scarlet or tropical sage, is a tender herbaceous perennial widely introduced and seen in public gardens all over the world. With few molecular resources, breeding is still restricted to traditional phenotypic selection, and the genetic mechanisms underlying phenotypic variation still remain unknown. Hence, a high quality reference genome will be very valuable for marker assisted breeding, genome editing or molecular genetics.

Findings: We generated 66 gigabases (Gb) and 37 Gb of raw DNA sequences, respectively, from whole-genome sequencing of a largely homozygous scarlet sage inbred line using PacBio Single-Molecule Real-Time (SMRT) and Illumina HiSeq sequencing platforms. PacBio *de novo* assembly yielded a final genome with a scaffold N50 size of 3.12 megabases (Mb), and a total length of 808 Mb. The repetitive sequences identified accounted for 57.52% of the genome sequence and 54,008 protein-coding genes were predicted collectively with *ab initio* and homology-based gene prediction from the masked genome. The divergence time between *S. splendens* and *S. miltiorrhiza* was estimated with 28.21 million years ago (Mya). Moreover, 3,797 species-specific genes and 1,187 expanded gene families were identified for the scarlet sage genome.
Conclusions: We provide the first genome sequence and gene annotation for the scarlet sage. The availability of these resources will be of great importance for further breeding strategies, genome editing and also for comparative genomics among related species.

Keywords: annotation, evolution, reference genome, *Salvia splendens*, scarlet sage

Data description

Background information

*S. splendens* (NCBI taxon ID:180675), scarlet or tropical sage, is a herbaceous perennial species, which is native to Brazil. While it is a perennial in warmer climate zones, it grows as an annual in cooler areas. *S. splendens* is a very popular bedding plant, and is widely cultivated in public gardens all over the world [3, 5], characterized by its dense flowers, and wide variation of colours (scarlet, purple, pink, blue, lavender, salmon, yellow green, white and bicolor), as well as long lasting flowering (3-9 weeks or even longer). Additionally, *S. splendens* can provide...
outstanding visual effects when grown in beds, borders and containers with
long-lasting lifespans ranging from late spring to the occurrence first frost.
Furthermore, the flower is easy to maintain and fairly free of pests and diseases due to
Lamiaceae’s characteristic insect repellent fragrance content [6]. The plant blends
nicely with other annuals or perennial plants for the best visual effects in an ensemble
setting; in addition this plant requires little deadheading as well it attracts various
butterfly species. *S. splendens* is a prolific and durable bloomer, thrives in full sun,
and survives in a large range of soil moisture regimes.

Traditional breeding activities using phenotypic selection as well as performing
targeted variety hybridizations between elite cultivars have resulted in a large number
of new cultivars with different performances regarding flowering characters (related
to colour, flowering time, flowering period amongst others), individual growth
performance, height, and/or tolerance to moisture or temperature extremes. However,
little is known about the molecular mechanisms underlying such economically
important characteristics for ornamental varieties. To date, only few genetic markers
[7] are available for marker assistant breeding or genetic modification.

In the current study, we present the first high quality genome assembly for *S.
splendens* with a hybrid assembly strategy using PacBio Single-Molecule Real-Time
and Illumina HiSeq short-read sequencing platforms. The genome assembly, its
structural and functional annotation, provide a valuable reference for the genomic
dissection of the phenotypic variation in *Salvia*, and new breeding strategies. This
reference genome could also be used in comparative genomics with the recently released *Salvia* genome (*S. miltiorrhiza*) [8, 9] and the mint genome (*Mentha longifolia*) [10] to study the biosynthesis of important fragrant and medicinal compounds.

**Plant material**

We chose the elite variety *S. splendens*, "Aoyunshenghuo (Olympic flame)" (Fig. 1a-b) for whole genome sequencing, which was originally developed by multiple rounds of selection/selfing of one hybrid to obtain this inbred line. This cultivar is characterized by resistance to drought, high temperature, and improved performance related to a longer flowering period; it is well adapted to climate conditions across North China, and therefore grows well in Beijing. Because of the high homozygosity obtained due to advanced generation selfing, this cultivar shows no phenotypic segregation, a characteristic of important commercial value. Seeds of this cultivar were provided by the Beijing Institute of Landscape Architecture germplasm bank.

**PacBio SMRT sequencing**

Plants for DNA extraction have been placed in the dark for 48 h before harvesting the leaf material. DNA was purified with Mobio PowerClean® Pro DNA Clean-Up Kit and quality was assessed by standard agarose gel electrophoresis and Thermo Fisher Scientific Qubit Fluorometry. Genomic DNA was sheared to a size range of 15–40 kb using either AMPure beads (Beckman Coulte) or g-TUBE (Covaris), and enzymatically repaired and converted into SMRTbell template libraries as recommended by Pacific Biosciences. Following this procedure, hairpin adapters were ligated following exonuclease-based digestion (of the remaining damaged DNA fragments and those fragments without adapters at both ends). Subsequently, the resulting SMRTbell templates were size-selected by Blue Pippin electrophoresis (Sage Sciences) and templates ranging from 15 to 50 kb were sequenced on a PacBio RS II instrument using P6-C4 sequencing chemistry (25 Single-Molecule Real-Time (SMRT) cells for individual huo1) and on a PacBio Sequel instrument using S/P2-C2 sequencing chemistry (8 SMRT cells for the other individual, huo1_1). A total of 8,858,116 PacBio post-filtered reads were generated. This produced a total of 65,962,079,028 bp (roughly 82x of the assembled genome) of single-molecule sequencing data, with an average read length of 7,446 bp (Fig. S1 and Table S1).

Illumina short-read sequencing

DNA was extracted from leaf tissue of the same soil-grown seedlings (huo1 and
huo1_1) using the Qiagen DNeasy Plant Mini Kit. Two 500 bp paired-end (PE)
libraries (huo1 and huo1_1) were prepared using the NEBNext Ultra DNA Library
Prep Kit for Illumina sequencing with an Illumina HiSeq X Ten machine. Short reads
were processed with Trimmomatic v0.33 (Trimmomatic, RRID:SCR_011848) [11,
12] and Cutadapt v1.13 (cutadapt, RRID:SCR_011841) [13, 14] to remove adapter
sequences and leading and trailing bases with a quality score below 20 and reads with
an average per-base-quality of 20 over a 4 bp sliding window. Reads < 70 nucleotides
in length after trimming were removed from further analysis. A total of 265.53 million
reads were generated. This produced a total of 36.83 Gb (roughly 40x of the
assembled genome) of raw sequencing data, with an average cleaned read length of
137 bp (Table S1).

Estimation of genome size, heterozygosity, and repeat content

All generated PacBio reads were filtered and corrected with Canu v1.5 (Canu,
RRID:SCR_015880) [15], thereafter, Jellyfish (Jellyfish, RRID:SCR_005491) [16]
was used to count occurrence of k-mers based on the processed data. Finally, gce
1.0.0 [17] was employed to estimate the overall characteristics of the genome such as
genome size, repeat contents and heterozygous rate. In this study, a total of
22,117,819,357 k-mers were generated and the peak k-mer depth was 31 (Fig. S2).
The genome size was estimated to be approximately 711 Mb (Table S2) and the final
cleaned data corresponded to the coverage of about 33-fold. Repeat and error rates
were estimated to be 47.99% and 0.27%, respectively, and heterozygosity rate was 0.06%.

**De novo genome assembly**

**De novo** assembly was conducted as follows in a progressive manner. Firstly, primary assemblies were generated from PacBio long reads of the 31 Gb from the ‘huo1’ sequenced individual by four different Overlap-Layout-Consensus (OLC) based assemblers, Canu (produced assembly v0.1), MECAT 1.1 (assembly v0.2) [18], FALCON v0.7 (Falcon, RRID:SCR_016089) [19] (https://github.com/PacificBiosciences/FALCON/) after Canu correction (v0.3) and SMARTdenovo 1.0.0 (https://github.com/ruanjue/smartdenovo) after Canu correction (v0.4) (Table S3). Based on the size of the assembled genome, the total number of assembled contigs, N50, the L50, maximum length of the contigs, and also the completeness of the genome assembly as assessed by using BUSCO criteria v2.0.1 (BUSCO, RRID:SCR_015008)[20] (1,440 single copy orthologs of the Viridiplantae database) with the BLAST E-value cutoff of $10^{-5}$, assembly (v0.1) from Canu was chosen for further polishing and scaffolding. In this selected primary assembly, the assembled genome size was 808 Mb distributed across 2,306 contigs with N50 of 2.06 Mb, L50 of 109 and maximum contig length of 8.88 Mb. We also confirmed on average 92.1% gene completeness in this assembly (Table S3). In the following steps, the arrow algorithm v2.2.1...
(https://github.com/PacificBiosciences/GenomicConsensus) was used to further improve the assembly based on PacBio long reads (v1.0), after which SSPACE-LongRead 1.1 [21] and SSPACE-standard 3.0 (SSPACE, RRID:SCR_005056) [22] were used for subsequent scaffold assembly based on PacBio long reads of 35 Gb from the second sequenced individual ‘hou1_1’ and Illumina short reads, respectively. Finally, after scaffold processing and subsequent gap filling with SOAPdenovo and GapCloser (GapCloser, RRID:SCR_015026) [23] (v1.1), arrow v2.2.1 algorithm (based on PacBio long reads) and pilon (Pilon, RRID:SCR_014731) (based on Illumina short reads, and run two times, parameters for Pilon: --changes --diploid --dumpreads.), we got the final genome assembly (v1.2).

Mapping of Illumina reads was done by using Bowtie2 v2.3.0 (Bowtie, RRID:SCR_005476) [24]. We detected 400,170 SNPs, 96,854 insertions and 62,637 deletions, respectively, for the first pilon run, while, subsequently, a greatly decreased number of variants for the second pilon run (40,465 SNPs, 6,935 insertions and 9,976 deletions, respectively). In this final assembly, we gained an assembled genome size of 808 Mb characterized by 2,204 contigs and 1,525 scaffolds (with contig N50 of 2.27 Mb and scaffold N50 of 3.12 Mb), and by gene completeness of 92.2% (Table 1 and Table S3). This assembly represents the highest continuity and completeness among the recently released genome assemblies for the Salvia genus [8, 9] and the mint [10], as it was examined by length distribution plotting of contigs and scaffolds as shown in Fig. 2a, b.
DNA repeats annotation

RepeatModeler v1.0.10 (RepeatModeler, RRID:SCR_015027) (http://www.repeatmasker.org/RepeatModeler.html) was employed to de novo identify and classify repeat families in the genome assembly. Subsequently, the outputs from RepeatModeler and RepBase [25] library were combined and used as repeat library for subsequent RepeatMasker (RepeatMasker, RRID:SCR_012954) (v4.0.7, rmblast-2.2.28) (http://www.repeatmasker.org/) analyses, which was used to fully discover and identify repeats within the assembled genome. In summary, 57.52% of the genome was annotated as repeats among which we found 1.08% simple repeats and 40.35% known transposable elements (TE). Long terminal repeats (LTRs) constituted the greatest proportion, 26.49% of the genome, and DNA TE made up 11.91% of the genome. Gypsy (18.15% of the genome) and Copia (7.92%) TEs were the largest components of LTRs. The results of repeat annotations are summarized in Table S4.

RNA sequencing, transcriptome assembly and functional annotation

RNA was extracted from the two cultivated lines with different flower colours (red and purple) using tissue obtained from, roots, shoots, leaves, calyxes and corollas. Frozen tissue from all samples was ground manually using mortar and pestil, and RNA was isolated using the NEBNext Poly(A) mRNA Magnetic Isolation Module.
RNA quality was assessed using an Agilent 2100 BioAnalyzer. Sequencing libraries were prepared using the NEBNext Ultra RNA Library Prep Kit for Illumina. 150 bp PE sequencing was performed using an Illumina HiSeq X Ten.

1,344 million raw reads from RNA sequencing were processed by Trimmomatic and Cutadapt and aligned to the genome assembly with HiSat2 v2.1.0 (HiSat2, RRID:SCR_015530) [26]. Base quality was checked with FastQC (FastQC, RRID:SCR_014583) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) before and after data cleaning, and respective statistics of RNA sequencing data are shown in Table S1. Reference genome guided transcriptome assemblies were independently prepared with Cufflinks v2.1.1 (Cufflinks, RRID:SCR_014597) [27], StringTie v1.3.3b (StringTie, RRID:SCR_016323) [28] and Trinity v2.0.6 (Trinity, RRID:SCR_013048) [29]. De novo assembly was generated using Trinity, then, transcriptome assemblies were combined and further refined using CD-HIT v4.6 [30], and finally, 192,169 unique transcripts were gained. The summary of the transcriptome assemblies is shown in Table S5.

AUGUSTUS v3.2.3 (Augustus, RRID:SCR_008417) [31] was employed for ab initio gene prediction, using model training based on coding sequences from Arabidopsis thaliana and S. miltiorrhiza (with two sets of proteins from independent genome annotation [8, 9]). Then, transcripts from RNA sequencing were aligned to the repeat-masked reference genome assembly with BlastN and TblastX from BLAST v2.2.28+ (NCBI BLAST, RRID:SCR_004870) [32] (E-value cutoff of 10^-5), and
protein sequences from A. thaliana and S. miltiorrhiza were aligned to the repeat-masked reference genome assembly with BlastX (E-value cutoff of $10^{-5}$). After optimization with Exonerate v2.4.0 (https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate) [33], gene model predictions were finalized prepared using the MAKER package v2.31.9 (MAKER, RRID:SCR_005309) [34] provided within AUGUSTUS. To assess the quality of the gene prediction, AED (Annotation Edit Distance) scores were generated for each of the predicted genes as part of the MAKER pipeline. Putative function for each identified gene was assessed by performing a BLAT (BLAT, RRID:SCR_011919) [35] search of the peptide sequences against the Uniprot database (UniProt, RRID:SCR_002380) [36], Protein annotation against PFAM (Pfam, RRID:SCR_004726) [37] and InterProScan (InterProScan, RRID:SCR_005829) [38] ID were also conducted using the scripts provided in the MAKER package. Completeness of gene annotation was checked using BUSCO (1,440 single copy orthologs of the Viridiplantae database) with a BLAST E-value cutoff of $10^{-5}$.

54,008 genes could be predicted, with average lengths of gene regions, genes (including 5’, 3’ UTRs, exons and introns), CDS and exons of 3,430.43 bp, 1696.34 bp, 1293.62 bp and 265.94 bp, respectively (Table S6). The comparisons among genomes from related species regarding lengths of genes, exons, and introns are shown in Fig. 2. The distribution of AED tagged by MAKER is shown in Fig. S3, in
which about 97% of the annotated genes (52,338 genes) had an AED < 0.5 (Table S6), thus indicating that the annotation is well supported. The result from BUSCO assessment of the quality of the genome assembly and annotation is shown in Table S7. 92.08 % of the universal single-copy genes (1,326 genes out of the total 1,440 genes) were identified, supporting the high quality of the genome assembly. Among the 1,326 BUSCO conserved single-copy genes detected in the scarlet genome, 466 genes were found single-copy, while 860 genes were duplicated (Table S7).

The predicted genes were annotated against several functional databases, including: (1) the NCBI non-redundant protein database (Nr; http://www.ncbi.nlm.nih.gov), (2) Swiss-Prot protein database (http://www.expasy.ch/sprot) [36], (3) Translated EMBL-Bank (part of the International Nucleotide Sequence Database Collaboration, TrEMBL, http://www.ebi.ac.uk/uniprot) [36], (4) the protein families database (Pfam; http://pfam.xfam.org/), (5) Cluster of Orthologous Groups for eukaryotic complete genomes (KOG) database (http://genome.jgi-psf.org/help/kogbrowser.jsf), (6) KO (the Kyoto Encyclopedia of Genes and Genomes, Orthology) database (http://www.genome.jp/kegg/ko.html) [39], and (7) Gene ontology (GO) (http://www.geneontology.org) [40]. 94.67 % of all predicted genes could be annotated with the following protein related databases: NR (94.60 %), Swiss-Prot (63.40 %), TrEMBL (93.50 %), Pfam (82.10 %), KOG (90.05 %), KO (37.40 %), and GO (78.80 %) (Table S8).
Identification of orthologous genes and phylogenetic inference

To analyze gene families, we downloaded the protein sequences of 15 additional species (Salvia miltiorrhiza [8, 9], Fraxinus excelsior [41], Olea europaea [42], Mimulus guttatus [43], Utricularia gibba [44], Sesamum indicum [45], Coffea canephora [46], Solanum lycopersicum [47], Daucus carota [48], Vitis vinifera [49], Arabidopsis thaliana [50], Populus trichocarpa [51], Oryza sativa [52] and Beta vulgaris [53]) (Table S9). Orthologous and paralogous gene clusters were identified among species using OrthoMCL v2.0.9 [54]. Recommended settings were used for all-against-all BLASTP comparisons (Blast+ v2.3.056) [32] and OrthoMCL[22] analyses.

A total of 35,808 OrthoMCL families were built based on effective database sizes of all versus all BLASTP with an E-value of $10^{-5}$ and a Markov Chain Clustering default inflation parameter. We identified 1,306 gene families (3,797 genes) that were specific to the scarlet sage genome when comparing with the other 15 genomes (Table S10), and we detected 10,770 gene families that have expanded in the scarlet sage lineage, using CAFE v4.0 [55, 56] (Fig. 2c). The expanded gene families were enriched for 60 significant ($q<0.05$) GO-terms of three different functional categories, i.e. BP, CC, and MF (Table S11) and one KEGG pathway (amino acid metabolism) (Table S12) significant at $q<0.05$. Also, 3,579 genes and 78 gene families were detected to be contracted and found to have rapidly evolved within the scarlet sage
Subsequently, 134 orthologous proteins among the 16 analyzed genomes were acquired and aligned with MUSCLE v3.8.31 (MUSCLE, RRID:SCR_011812) [57] employing default settings. A maximum likelihood phylogenetic tree was then generated using the concatenated amino acid sequences in PhyML v3.0 (PhyML, RRID:SCR_014629) [58] with GTR+G+I model. The divergence time was estimated with r8s v1.81 [59] and calibrated against the timing of divergence between A. thaliana and V. vinifera (124 Mya) [60] as well as against A. thaliana and P. trichocarpa divergence time (90 Mya) [61]. The phylogenetic analysis identified the close relationship among the three Salvia genomes and their divergence time was estimated with about 28.21 Mya (Fig. 2c).

Secondary metabolic pathways: gene annotations, gene clusters and comparative genomics

The mint family is recognized as providing promising sources of bioactive secondary metabolites [62]. In fact, a diverse variety of bioactive secondary metabolites can be found with a wide range of pharmacological activities: antimicrobial, antispasmodic, carminative, antioxidant, antiulcer, cytoprotective, heptoprotective, cholagogue, chemo-preventive, anti-inflammatory, antidiabetogenic etc. Here, we obtained enzymatic annotations for coding genes by employing the E2P2 package v3.1 (https://gitlab.com/rhee-lab/E2P2/tree/master). Then, we mapped genes to flavonoid and menthol biosynthesis pathways by querying the Plant Metabolic Network (PMN...
Regarding the flavonoid biosynthesis pathway, we found an abundance of genes encoding annotated enzymes in this pathway, especially of note the 41 genes for flavanone synthase I (EC: 1.14.11.9) (Figure S5 and Supplementary_File_1). With respect to menthol biosynthesis, certain genes are still lacking annotations for enzymes such as (+)-pulegone reductase (EC: 1.3.1.81), (-)-isopiperitenone reductase (EC: 1.3.1.82) or menthol-dehydrogenase (lacking EC number) (Figure S6 and Supplementary_File_1). However, this pathway mapping analysis provides a highly valuable reference for the genetic dissection of key metabolic genes for the scarlet sage.

The presence of metabolic gene clusters for secondary metabolic pathways are common in bacteria and filamentous fungi, and are also widely reported in plants [64-66]. Using the newly created and robust computational toolkit, plantSMASH [67], we identified 85 gene clusters potentially related to secondary metabolic biosynthesis in the scarlet sage genome as reported here, and 23 gene clusters in the S. miltiorrhiza genome [8]. Genomic position, gene composition, functional annotation of the identified gene clusters were summarized in Table S13, Supplementary_File_2 and Supplementary_File_3. The gene clusters were found to be potentially related to the biosynthesis of alkaloids, saccharides, polyketides, terpenes, and lignans. It was previously reported that physical clustering of terpene synthase genes (TPS) and cytochrome P450 mono-oxygenase genes is frequently associated with consecutive enzymatic actions in terpenoid biosynthesis [68]. Interestingly, we detected eight such
gene clusters within the scarlet sage genome, but none in the S. miltiorrhiza genome, which could partially be due to the draft status of the genome assembly for S. miltiorrhiza. Furthermore, significant gene co-expression across different organs was detected for one TPS gene and two out of four P450 genes located in a single gene cluster (i.e. Cluster 63; Table S13 and Supplementary_File_2). Evidence for moderate or significant co-expression among clustered genes was revealed and shown in Supplementary_File_2.

Based on the collinearity elucidated by former OrthoMCL analyses, a comparative genomic study between the scarlet sage and S. miltiorrhiza genomes revealed six pairs of gene clusters which share synteny between these two congeneric plants, and two blocks from the scarlet sage share synteny with one block from S. miltiorrhiza (Figure S7). Among the shared synteny blocks, four could be related to saccharide, one to lignan and another to polyketide biosynthesis. The smaller number of gene clusters detected for S. miltiorrhiza and subsequently, fewer shared synteny blocks of metabolic gene cluster between these two species may be partially attributed to the present state of the S. miltiorrhiza genome assembly which is hundred times more fragmented than that of the scarlet sage. Thus, here, we provided a starting point for comparative genomics among plant species within the mint family.

In summary, we presented the draft assembly for the scarlet sage genome using a PacBio long-read dominated strategy, which was responsible for obtaining the high
sequence assembly quality. Also, the almost complete homozygosity within the sequenced inbred line’s genome was a key factor for the high continuity gained in this study. The novel genome data generated in the present study will provide a valuable resource for studying the molecular underpinnings of the various phenotypic variation found within \textit{Salvia sp.}, and sets the foundation for molecular-informed breeding strategies and genome editing approaches for this valued ornamental flowering plant. Moreover, this genome assembly is useful for comparative genomic studies among related species.

**Availability of supporting data**

The genome assembly, annotations, and other supporting data are available via the \textit{GigaScience} database GigaDB [69]. The raw sequence data have been deposited in the Short Read Archive (SRA) under NCBI BioProject ID PRJNA422035.

**Abbreviations**

AED: Annotation Edit Distance; bp: base pair; BUSCO: benchmarking universal single-copy orthologs; CDS: coding sequence; Gb: gigabases; kb: kilobases; LTR: Long terminal repeats; Mb: megabases; Mya: million years ago; PE: paired end; SMRT: Single-Molecule Real-Time; SNP: Single Nucleotide Polymorphism; TE: TPS: transposable element; terpene synthase genes.
Acknowledgement

This study was funded by Beijing Key Laboratory of Greening Plants Breeding (NO. Z201605) and Fundamental Research Funds for the Central Universities (NO. YX2013-41).

Author Contributions

AXD, HBX, RCC, JFM, FM and IP conceived and designed the study; AXD, HBX, ZJL, HL, YQS, SN, ZNZ, RFC, HLZ, RGZ and QZY prepared the materials and conducted the experiments; JFM, HBX, FM, IP wrote the manuscript.

Conflict of Interest

The authors declare that they have no competing financial interests.

References


69. Dong, A; Xin, H; Li, Z; Liu, H; Sun, Y; Nie, S; Zhao, Z; Cui, R; Zhang, R; Yun, Q; Wang, X; Maghuly, F; Porth, I; Cong, R; Mao, J (2018): Supporting data for "High quality assembly of the reference genome for scarlet sage, Salvia splendens, an economically important ornamental plant" GigaScience Database. http://dx.doi.org/10.5524/100463.
Supplementary Figures

Figure S1. Length distribution of PacBio subreads.

Figure S2. K-mer frequency distribution at k-mer size of 17. A k-mer refers to an artificial sequence division of K nucleotides. From k-mer frequency, genomic characteristics (genome size, repeat structure and heterozygous rate) could be estimated. Peaks at depths of 31 and 62 were annotated with dash lines.

Figure S3. Distribution of AED scores from gene prediction. AED, Annotation Edit Distance, AED = 0 indicates perfect agreement between annotation and the evidence; AED = 1 indicates no evidence support for annotation.

Figure S4. Length distribution of annotated genes, exons and introns. a-c for annotated genes, exons and introns from different genome assemblies.

Figure S5. Flavonoid biosynthesis pathway. Flavonoid biosynthesis pathways by querying the Plant Metabolic Network (https://www.plantcyc.org/), enzymatic coding genes of the scarlet sage were shown for key reactions.

Figure S6. Menthol biosynthesis pathway. Menthol biosynthesis pathways by querying the Plant Metabolic Network (https://www.plantcyc.org/), enzymatic coding genes of the scarlet sage were shown for key reactions.

Figure S7. Shared synten addressed for metabolic gene clusters between Salvia genomes. a-f: display of the different pairs of synteny blocks. Genes are colored along the contigs/scaffolds to compare between scarlet sage and *Salvia miltiorrhiza* Zhang [8], with metabolic genes highlighted with olive drab color, other homologous genes
are shown in grey.

Supplementary Tables

Table S1. Summary of Pacbio and Illumina sequencing data generated in the present study. IDs of the study, sample, library and accession in NCBI SRA and sequencing platform, material origins of the sequenced DNA or RNA, the statistics of the raw and cleaned data are shown.

Table S2. Estimation of genome characteristics based on 17-mer statistics.

Table S3. Statistics of the different versions of the genome assembly of the scarlet sage. NA: data not available; * statistics for contigs/scaffolds.

Table S4. Summary of the annotated interspersed repeats in the genome assembly of the scarlet sage.

Table S5. Summary of the transcriptome assemblies.

Table S6. Summary of the annotated genes. AED: Annotation Edit Distance; gene regions (including UTRs, exons and introns); genes (including 5’, 3’ UTRs, exons and introns).

Table S7. Summary of BUSCO evaluation of gene prediction.

Table S8. Summary of functional annotation of predicted genes.

Table S9. Genomic data used for gene families analyses. Origins, download links, assembly versions, genome properties and references of 15 analyzed genomes are shown.
Table S10. Summary of gene family analyses. Unique groups and genes, single-copy and duplicated groups and genes are summarized for the 16 analyzed genomes of 15 plant species.

Table S11. GO enrichment of expanded gene families. (A) ‘Category’ is the Gene Ontology (GO) term ID; (B) ‘p_value’ is the over represented p-value indicating the observed frequency of a given term among analyzed genes is equal to the expected frequency based on the null distribution; i.e. lower p-values indicate stronger evidence for overrepresentation; (C) ‘q_value’ is the Benjamini and Hochberg adjusted p-value, (D) ‘numEPInCat’ is the number of expanded gene families in the corresponding GO category; (E) ‘numInCat’ is the number of detected gene families in the corresponding GO category; (F) ‘Term’ is the GO term; (G) ‘Ontology’ indicates which ontology the term comes from. 60 significant (q<0.05) GO-terms of three different functional categories are indicated in bold.

Table S12. KEGG enrichment of expanded gene families. (A) ‘KO category’ is the KEGG Orthology (KO) category ID; (B) ‘p_value’ is the over represented p-value indicating the observed frequency of a given term among analyzed genes is equal to the expected frequency based on the null distribution; i.e. lower p-values indicate stronger evidence for overrepresentation; (C) ‘q_value’ is the Benjamini and Hochberg adjusted p-value, (D) ‘numEPInCat’ is the number of expanded gene families in the corresponding KO category; (E) ‘numInCat’ is the number of detected gene families in the corresponding KO category; (F) ‘Pathway’ is the KEGG pathway;
‘Class’ indicates which KEGG class the pathway comes from. One significant (q<0.05) KEGG pathway is indicated in bold.

**Table S13.** Summary of metabolic gene clusters detected in genomes of *Salvia miltiorrhiza* and *S. splendens.* (A) ‘Genome’ denotes the genome origination; (B) ‘Cluster’ is the code for a certain gene cluster detected; (C) ‘Record’ denotes the contig/scaffold ID from where the gene cluster was detected; (D) ‘Type’ denotes the functional assignment for the gene cluster; (E) ‘From’, ‘To’ and ‘Size’ denote the genomic position and range of the gene cluster; (F) ‘Core domains’ denote the domain annotation for the metabolic genes in the cluster; (G) ‘CD-HIT Cluster’ indicate the number of genes in the cluster; (H) ‘Gene cluster genes’ is showing the ID of genes in the cluster.

**Supplementary Files**

**Supplementary File 1.** Genes (Gene ID, name and EC number) mapped to flavonoid and menthol biosynthesis pathways.

**Supplementary File 2.** Structure of a metabolic gene cluster (polyketide synthesis) and gene expression patterns of *Salvia splendens.* Genomic position, gene composition, functional annotation of gene cluster are shown, also including a heatmap of tissue specific expression of the genes within the presented cluster is shown. HG: root of red flower (individual); HJ: stem of red flower (individual); HY: leave of red flower (individual); HE: calyx of red flower (individual); HHG: corolla
of red flower (individual); ZG: root of purple flower (individual); ZJ: stem of purple
flower (individual); ZY: leave of purple flower (individual); ZE: calyx of purple
flower (individual); ZHG: corolla of purple flower (individual).

**Supplementary File 3.** Structure of a metabolic gene cluster (alkaloid synthesis).

Genomic position, gene composition, functional annotation of gene cluster were
shown.


Table 1. Statistics of the final genome assembly of the scarlet sage.

<table>
<thead>
<tr>
<th></th>
<th>Contig</th>
<th></th>
<th>Scaffold</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Size (bp)</td>
<td>Number</td>
<td>Size (bp)</td>
<td>Number</td>
</tr>
<tr>
<td><strong>Total Size</strong></td>
<td>807,514,799</td>
<td>-</td>
<td>809,159,598</td>
<td>-</td>
</tr>
<tr>
<td><strong>Total Number</strong></td>
<td>-</td>
<td>2,204</td>
<td>-</td>
<td>1,525</td>
</tr>
<tr>
<td><strong>N10</strong></td>
<td>6,529,455</td>
<td>10</td>
<td>8,157,631</td>
<td>9</td>
</tr>
<tr>
<td><strong>N50</strong></td>
<td>2,267,074</td>
<td>100</td>
<td>3,123,266</td>
<td>73</td>
</tr>
<tr>
<td><strong>N90</strong></td>
<td>265,262</td>
<td>456</td>
<td>433,303</td>
<td>324</td>
</tr>
<tr>
<td><strong>Max.</strong></td>
<td>10,812,588</td>
<td>-</td>
<td>12,944,193</td>
<td>-</td>
</tr>
<tr>
<td><strong>Min.</strong></td>
<td>500</td>
<td>-</td>
<td>9,495</td>
<td>-</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>366,386</td>
<td>-</td>
<td>530,596</td>
<td>-</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>38,049</td>
<td>-</td>
<td>48,557</td>
<td>-</td>
</tr>
<tr>
<td><strong>Gap</strong></td>
<td>-</td>
<td>-</td>
<td>1,644,799</td>
<td>679</td>
</tr>
<tr>
<td><strong>GC Content</strong></td>
<td>38.84%</td>
<td>-</td>
<td>38.76%</td>
<td>-</td>
</tr>
</tbody>
</table>
Figures:

Fig. 1 Images of the scarlet sage, *Salvia splendens*.

a-b, flowers of the sequenced cultivar of *S. splendens*, "Aoyunshenghuo (Olympic flame)"; c, the scarlet sage with different flower colors in bedding; d-k, the scarlet sage with flowers of different pure colors or bi-colors.

Fig. 2 Quality of scarlet sage genome assembly and the phylogenomic inferences.

Quality was assessed by comparing the scarlet genome with the recently released genomes of related species. Length distribution of contigs (a) and scaffolds (b); c, phylogenetic tree, divergence time, and profiles of gene families that underwent expansion or contraction. *Salvia miltiorrhiza* Zhang [14] and *Salvia miltiorrhiza* Xu [14] are two genome assemblies reported for *Salvia miltiorrhiza*.
Fig. 2

(a) Genome assemblies
- Mentha longifolia
- Salvia miltiorrhiza Xu
- Salvia miltiorrhiza Zhang
- Salvia splendens

(b) Genome assemblies
- Mentha longifolia
- Salvia miltiorrhiza Xu
- Salvia miltiorrhiza Zhang
- Salvia splendens

(c) Phylogenetic tree
- Expansion
- Contraction

Plant species:
- Salvia miltiorrhiza Xu
- Salvia miltiorrhiza Zhang
- Salvia splendens
- Vfrostden gossettii
- Scroenmone indicum
- Circaevia glabra
- Psenadinus acidthier
- Genta europae
- Scopaea fruticosum
- Coffea canephora
- Dacca canus
- Anabasaphi rayafanae
- Papedes trachyspermum
- Fitchi virgulifru
- Beta vulgaris
- Oryza sativa
Click here to access/download
**Supplementary Material**
Table_S1.xlsx
Click here to access/download
Supplementary Material
Table_S2.docx
Click here to access/download
Supplementary Material
Table_S3.docx
Click here to access/download
**Supplementary Material**
Table_S4.docx
Click here to access/download
Supplementary Material
Table_S6.docx
Click here to access/download
Supplementary Material
Table_S8.docx
Click here to access/download
Supplementary Material
Table_S9.docx
Click here to access/download
**Supplementary Material**
Table_S10.docx
Click here to access/download
Supplementary Material
Table_S11.xlsx
Click here to access/download
Supplementary Material
Table_S12.docx
Supplementary Material

Click here to access/download

**Supplementary Material**

**Table S13.xlsx**
Click here to access/download
Supplementary Material
Fig_S2.pdf
Click here to access/download
Supplementary Material
Fig_S3.pdf
Click here to access/download
Supplementary Material
Fig_S4.pdf
Click here to access/download

**Supplementary Material**

Fig_S5.pdf
Click here to access/download
Supplementary Material
Fig_S6.pdf
Click here to access/download
Supplementary Material
Fig_S7.pdf
Click here to access/download
Supplementary Material
Supplementary_File_1.docx
Click here to access/download
Supplementary Material
Supplementary_File_2.docx