

GigaScience

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

--Manuscript Draft--

Manuscript Number:	GIGA-D-18-00028	
Full Title:	High quality assembly of the reference genome for scarlet sage, <i>Salvia splendens</i> , an economically important ornamental plant	
Article Type:	Data Note	
Funding Information:	Beijing Key Laboratory of Green Plants Breeding	Dr. Ri-Chen Cong
	Fundamental Research Funds for the Central Universities (YX2013-41)	Mr. Jian-Feng Mao
Abstract:	<p>Background: <i>Salvia splendens</i> 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.</p> <p>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 <i>S. splendens</i> and <i>S. miltiorrhiza</i> 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.</p> <p>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.</p>	
Corresponding Author:	Ri-Chen Cong, Ph.D Beijing Institute of Landscape Architecture Beijing, CHINA	
Corresponding Author Secondary Information:		
Corresponding Author's Institution:	Beijing Institute of Landscape Architecture	
Corresponding Author's Secondary Institution:		
First Author:	Jian-Feng Mao, Ph.D.	
First Author Secondary Information:		
Order of Authors:	Jian-Feng Mao, Ph.D.	
	Ai-Xiang Dong	
	Hai-Bo Xin	
	Zi-Jing Li	
	Hui Liu	
	Yan-Qiang Sun	
	Shuai Nie	

	Zheng-Nan Zhao
	Rong-Feng Cui
	Hua-Li Zhang
	Ren-Gang Zhang
	Quan-Zheng Yun
	Fatemeh Maghuly
	Ilga Porth
	Ri-Chen Cong
Order of Authors Secondary Information:	
Opposed Reviewers:	
Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
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. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist ?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or	Yes

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 have met the above requirement as detailed in our [Minimum Standards Reporting Checklist](#)?

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

3
4 Ai-Xiang Dong^{1‡}, Hai-Bo Xin^{1‡}, Zi-Jing Li^{1‡}, Hui Liu^{2‡}, Yan-Qiang Sun^{2‡}, Shuai Nie²,
5 Zheng-Nan Zhao¹, Rong-Feng Cui¹, Hua-Li Zhang¹, Ren-Gang Zhang³, Quan-Zheng Yun³,
6 Fatemeh Maghuly⁴, Ilga Porth⁵, Ri-Chen Cong^{1*}, Jian-Feng Mao^{2*}

7
8 ¹ Beijing Key Laboratory of Green Plants Breeding, Beijing Institute of Landscape
9 Architecture, Beijing, 10020, China.

10 ² Beijing Advanced Innovation Center for Tree Breeding by Molecular Design,
11 National Engineering Laboratory for Tree Breeding, Key Laboratory of Genetics and
12 Breeding in Forest Trees and Ornamental Plants, Ministry of Education, College of
13 Biological Sciences and Technology, Beijing Forestry University, Beijing, 100083,
14 China.

15 ³ Beijing Ori-Gene Science and Technology Co. Ltd, Beijing, 10226, China.

16 ⁴ Plant Biotechnology Unit (PBU), Dept. Biotechnology, BOKU-VIBT, University of
17 Natural Resources and Life Sciences, Muthgasse 18, 1190 Vienna, Austria.

18 ⁵ Département des sciences du bois et de la forêt, Pavillon Charles-Eugène-Marchand,
19 1030, Avenue de la Médecine, Université Laval, Québec (Québec) G1V 0A6, Canada.

20 ‡These authors contributed equally to this paper.

21 *Correspondence to: hardhopee@163.com (RCC); jianfeng.mao@bjfu.edu.cn (JFM)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 22 **Abstract**

2
3 23 **Background:** *Salvia splendens* Ker-Gawler, scarlet or tropical sage, is a tender

4
5
6 24 herbaceous perennial widely introduced and seen in public gardens all over the world.

7
8
9 25 With few molecular resources, breeding is still restricted to traditional phenotypic

10
11
12 26 selection, and the genetic mechanisms underlying phenotypic variation still remain

13
14
15 27 unknown. Hence, a high quality reference genome will be very valuable for marker

16
17
18 28 assisted breeding, genome editing or molecular genetics.

19
20 29 **Findings:** We generated 66 gigabases (Gb) and 37 Gb of raw DNA sequences,

21
22
23 30 respectively, from whole-genome sequencing of a largely homozygous scarlet sage

24
25
26 31 inbred line using PacBio Single-Molecule Real-Time (SMRT) and Illumina HiSeq

27
28
29 32 sequencing platforms. PacBio *de novo* assembly yielded a final genome with a scaffold

30
31
32 33 N50 size of 3.12 megabases (Mb), and a total length of 808 Mb. The repetitive

33
34
35 34 sequences identified accounted for 57.52% of the genome sequence and 54,008 protein-

36
37
38 35 coding genes were predicted collectively with *ab initio* and homology-based gene

39
40
41 36 prediction from the masked genome. The divergence time between *S. splendens* and *S.*

42
43
44 37 *miltiorrhiza* was estimated with 28.21 million years ago (Mya). Moreover, 3,797

45
46
47 38 species-specific genes and 1,187 expanded gene families were identified for the scarlet

48
49
50 39 sage genome.

51
52 40 **Conclusions:** We provide the first genome sequence and gene annotation for the scarlet

53
54
55 41 sage. The availability of these resources will be of great importance for further breeding

56
57
58 42 strategies, genome editing and also for comparative genomics among related species.

59
60
61
62
63
64
65

1 43 **Keywords:** annotation, evolution, reference genome, *Salvia splendens*, scarlet sage

2
3 44 **Data description**

4
5
6 45 **Background information**

7
8
9 46 *Salvia* L., with nearly 1,000 species of shrubs, herbaceous perennials, and annuals, is
10
11
12 47 the largest genus in the mint family (Lamiaceae: Nepetoideae: Mentheae: Salviinae) [1-
13
14
15 48 4]. The genus is widely distributed throughout the world. Many species of this genus
16
17
18 49 are extensively used for culinary purposes, essential oil production and Chinese herbal
19
20
21 50 remedies such as the two species *S. officinalis* [3] and *S. miltiorrhiza* (Danshen).
22
23
24 51 Additionally, they are used as ornamental plants valued for their flowers or for their
25
26
27 52 aromatic foliage such as *S. splendens* (**Fig. 1 a-k**).

28
29
30 53 *S. splendens*, scarlet or tropical sage, is a herbaceous perennial species, which is
31
32
33 54 native to Brazil. While it is a perennial in warmer climate zones, it grows as an annual
34
35
36 55 in cooler areas. *S. splendens* is a very popular bedding plant, and is widely introduced
37
38
39 56 in public gardens all over the world [3,5], characterized by its dense flowers, and wide
40
41
42 57 variation of colours (scarlet, purple, pink, blue, lavender, salmon, yellow green, white
43
44
45 58 and bicolor), as well as long lasting flowering (3-9 weeks or even longer). Additionally,
46
47
48 59 *S. splendens* can provide outstanding visual effects when grown in beds, borders and
49
50
51 60 containers with long-lasting lifespans ranging from late spring to first frost occurrence.
52
53
54 61 Furthermore, the flower is easy to maintain and fairly free of pests and diseases due to
55
56
57 62 Lamiaceae's characteristic insect repellent fragrance content [6]. The plant blends
58
59
60 63 nicely with other annuals or perennial plants for the best visual effects in an ensemble
61
62
63
64
65

1 64 setting; in addition this plant requires little deadheading as well it attracts various
2
3
4 65 butterfly species. *S. splendens* is a prolific and durable bloomer, thrives in full sun, and
5
6
7 66 survives in a large range of soil moisture regimes.

8
9 67 Traditional breeding activities using phenotypic selection as well as performing
10
11
12 68 targeted variety hybridizations between elite cultivars have resulted in a large number
13
14
15 69 of new cultivars with different performances regarding flowering characters (related to
16
17
18 70 colour, flowering time, flowering period amongst others), individual growth
19
20
21 71 performance, height, and/or tolerance to moisture or temperature extremes. However,
22
23
24 72 little is known about the molecular mechanisms underlying such economically
25
26
27 73 important characteristics for ornamental varieties. To date, only few genetic markers
28
29
30 74 [7] are available for marker assistant breeding or genetic modification.

31
32 75 In the current study, we present the first high quality genome assembly for *S.*
33
34
35 76 *splendens* with a hybrid assembly strategy using PacBio Single-Molecule Real-Time
36
37
38 77 and Illumina HiSeq short-read sequencing platforms. The genome assembly, its
39
40
41 78 structural and functional annotation, provide a valuable reference for the genomic
42
43
44 79 dissection of the phenotypic variation in *Salvia*, and new breeding strategies. This
45
46
47 80 reference genome could also be used in comparative genomics with the recently
48
49
50 81 released *Salvia* genome (*S. miltiorrhiza*) [8,9] and the mint genome (*Mentha longifolia*)
51
52
53 82 [10] to study the biosynthesis of important fragrant and medicinal compounds.

54
55 83

56
57
58 84 **Plant material**

59
60
61
62
63
64
65

1 85 We chose the elite variety *S. splendens*, "Aoyunshenghuo (Olympic flame)" (**Fig. 1 a-**
2
3
4 86 **b)** for whole genome sequencing, which was originally developed by multiple rounds
5
6 87 of selection/selfing of one hybrid to obtain this inbred line. This cultivar is characterized
7
8
9 88 by resistance to drought, high temperature, and improved performance related to a
10
11
12 89 longer flowering period; it is well adapted to climate conditions across North China,
13
14
15 90 and therefore grows well in Beijing. Because of the high homozygosity obtained due to
16
17
18 91 advanced generation selfing, this cultivar shows no phenotypic segregation, a
19
20
21 92 characteristic of important commercial value. Seeds of this cultivar were provided by
22
23
24 93 the Beijing Institute of Landscape Architecture germplasm bank.
25

26 94

27 28 29 95 **PacBio SMRT sequencing**

30
31
32 96 High quality high molecular weight genomic DNA was extracted from leaves of two
33
34
35 97 soil-grown seedlings (huo1 and huo1_1) following ~20 kb SMRTbell™ Libraries”
36
37
38 98 protocol ([http://www.pacb.com/wp-content/uploads/2015/09/Shared-Protocol-](http://www.pacb.com/wp-content/uploads/2015/09/Shared-Protocol-Preparing-Arabidopsis-DNA-for-20-kb-SMRTbell-Libraries.pdf)
39
40
41 99 [Preparing-Arabidopsis-DNA-for-20-kb-SMRTbell-Libraries.pdf](http://www.pacb.com/wp-content/uploads/2015/09/Shared-Protocol-Preparing-Arabidopsis-DNA-for-20-kb-SMRTbell-Libraries.pdf)). Plants for DNA
42
43
44 100 extraction have been placed in the dark for 48 h before harvesting the leaf material.
45
46
47 101 DNA was purified with Mobio PowerClean® Pro DNA Clean-Up Kit and quality was
48
49
50 102 assessed by standard agarose gel electrophoresis and Thermo Fisher Scientific Qubit
51
52
53 103 Fluorometry. Genomic DNA was sheared to a size range of 15–40 kb using either
54
55
56 104 AMPure beads (Beckman Coulte) or g-TUBE (Covaris), and enzymatically repaired
57
58
59 105 and converted into SMRTbell template libraries as recommended by Pacific
60
61
62
63
64
65

1 106 Biosciences. Following this procedure, hairpin adapters were ligated following
2
3
4 107 exonuclease-based digestion (of the remaining damaged DNA fragments and those
5
6 108 fragments without adapters at both ends). Subsequently, the resulting SMRTbell
7
8
9 109 templates were size-selected by Blue Pippin electrophoresis (Sage Sciences) and
10
11
12 110 templates ranging from 15 to 50 kb were sequenced on a PacBio RS II instrument using
13
14
15 111 P6-C4 sequencing chemistry (25 Single-Molecule Real-Time (SMRT) cells for
16
17
18 112 individual huo1) and on a PacBio Sequel instrument using S/P2-C2 sequencing
19
20
21 113 chemistry (8 SMRT cells for the other individual, huo1_1). A total of 8,858,116 PacBio
22
23
24 114 post-filtered reads were generated. This produced a total of 65,962,079,028 bp (roughly
25
26
27 115 82x of the assembled genome) of single-molecule sequencing data, with an average
28
29
30 116 read length of 7,446 bp (**Fig. S1** and **Table S1**).

31
32 117

33 34 35 118 **Illumina short-read sequencing**

36
37
38 119 DNA was extracted from leaf tissue of the same soil-grown seedlings (huo1 and
39
40
41 120 huo1_1) using the Qiagen DNeasy Plant Mini Kit. Two 500 bp paired-end (PE) libraries
42
43
44 121 (huo1 and huo1_1) were prepared using the NEBNext Ultra DNA Library Prep Kit for
45
46
47 122 Illumina sequencing with an Illumina HiSeq X Ten machine. Short reads were
48
49
50 123 processed with Trimmomatic (v0.33) [11] and Cutadapt (v1.13) [12] to remove adapter
51
52
53 124 sequences and leading and trailing bases with a quality score below 20 and reads with
54
55
56 125 an average per-base-quality of 20 over a 4 bp sliding window. Reads < 70 nucleotides
57
58
59 126 in length after trimming were removed from further analysis. A total of 265.53 million

1 127 reads were generated. This produced a total of 36.83 Gb (roughly 40x of the assembled
2
3
4 128 genome) of raw sequencing data, with an average cleaned read length of 137 bp (**Table**
5
6 129 **S1**).

7
8
9 130

10 11 131 **Estimation of genome size, heterozygosity, and repeat content**

12
13
14
15 132 All generated PacBio reads were filtered and corrected with Canu [13], thereafter,
16
17
18 133 Jellyfish [14] was used to count occurrence of k-mers based on the processed data.
19
20
21 134 Finally, gce [15] was employed to estimate the overall characteristics of the genome
22
23
24 135 such as genome size, repeat contents and heterozygous rate. In this study, a total of
25
26
27 136 22,117,819,357 k-mers were generated and the peak k-mer depth was 31 (**Fig. S2**). The
28
29
30 137 genome size was estimated to be approximately 711 Mb (**Table S2**) and the final
31
32
33 138 cleaned data corresponded to the coverage of about 33-fold. Repeat and error rates were
34
35
36 139 estimated to be 47.99% and 0.27%, respectively, and heterozygosity rate was 0.06%.

37
38 140

39 40 41 141 ***De novo* genome assembly**

42
43
44 142 *De novo* assembly was conducted as follows in a progressive manner. Firstly, primary
45
46
47 143 assemblies were generated from PacBio long reads by four different Overlap-Layout-
48
49
50 144 Consensus (OLC) based assemblers, Canu (produced assembly v0.1) [13], MECAT
51
52
53 145 (assembly v0.2) [16], FALCON [17]
54
55 146 (<https://github.com/PacificBiosciences/FALCON/>) after Canu correction (v0.3) and
56
57
58 147 SMARTdenovo (<https://github.com/ruanjue/smarddenovo>) after Canu correction (v0.4)

1 148 (**Table S3**). Based on the size of the assembled genome, the total number of assembled
2
3
4 149 contigs, N50, the L50, maximum length of the contigs, and also the completeness of
5
6 150 the genome assembly as assessed by using BUSCO criteria [18] (956 single copy
7
8
9 151 orthologs of the Viridiplantae database) with the BLAST E-value cutoff of 10^{-5} ,
10
11
12 152 assembly (v0.1) from Canu was chosen for further polishing and scaffolding. In this
13
14
15 153 selected primary assembly, the assembled genome size was 808 Mb distributed across
16
17
18 154 2,306 contigs with N50 of 2.06 Mb, L50 of 109 and maximum contig length of 8.88
19
20
21 155 Mb. We also confirmed on average 92.1% gene completeness in this assembly (**Table**
22
23
24 156 **S3**). In the following steps, the arrow algorithm
25
26
27 157 (<https://github.com/PacificBiosciences/GenomicConsensus>) was used to further
28
29
30 158 improve the assembly based on PacBio long reads (v1.0), after which SSPACE-
31
32
33 159 LongRead [19] and SSPACE-standard [20] were used for subsequent scaffold assembly
34
35
36 160 based on PacBio long reads and Illumina short reads, respectively. Finally, after scaffold
37
38
39 161 processing and subsequent gap filling with SOAPdenovo and GapCloser [21] (v1.1),
40
41
42 162 arrow algorithm (based on PacBio long reads) and pilon (based on Illumina short reads,
43
44
45 163 and run two times), we got the final genome assembly (v1.2). In this final assembly, we
46
47
48 164 gained an assembled genome size of 808 Mb characterized by 2,204 contigs and 1,525
49
50
51 165 scaffolds (with contig N50 of 2.27 Mb and scaffold N50 of 3.12 Mb), and by gene
52
53
54 166 completeness of 92.2% (**Table 1 and Table S3**). This assembly represents the highest
55
56
57 167 continuity and completeness among the recently released genome assemblies for the
58
59
60 168 *Salvia* genus [8,9] and the mint family [10], as it was examined by length distribution
61
62
63
64
65

1 169 plotting of contigs and scaffolds as shown in **Fig. 2a, b.**

2
3
4 170

5
6 171 **DNA repeats annotation**

7
8
9 172 RepeatModeler (v1.0.10) (<http://www.repeatmasker.org/RepeatModeler.html>) was

10
11
12 173 employed to *de novo* identify and classify repeat families in the genome assembly.

13
14
15 174 Subsequently, the outputs from RepeatModeler and RepBase [22] library were

16
17
18 175 combined and used as repeat library for subsequent RepeatMasker (v4.0.7, rmblast-

19
20
21 176 2.2.28) (<http://www.repeatmasker.org/>) analyses, which was used to fully discover and

22
23
24 177 identify repeats within the assembled genome. In summary, 57.52% of the genome was

25
26
27 178 annotated as repeats among which we found 1.08% simple repeats and 40.35% known

28
29
30 179 transposable elements (TE). Long terminal repeats (LTRs) constituted the greatest

31
32
33 180 proportion, 26.49% of the genome, and DNA TE made up 11.91% of the genome.

34
35
36 181 Gypsy (18.15% of the genome) and Copia (7.92%) TEs were the largest components

37
38
39 182 of LTRs. The results of repeat annotations are summarized in **Table S4.**

40
41 183

42
43
44 184 **RNA sequencing, transcriptome assembly and functional annotation**

45
46
47 185 RNA was extracted from the two cultivated lines with different flower colours (red and

48
49
50 186 purple) using tissue obtained from, roots, shoots, leaves, calyxes and corollas. Frozen

51
52
53 187 tissue from all samples was ground manually using mortar and pestil, and RNA was

54
55
56 188 isolated using the NEBNext Poly(A) mRNA Magnetic Isolation Module. RNA quality

57
58
59 189 was assessed using an Agilent 2100 BioAnalyzer. Sequencing libraries were prepared

60
61
62
63
64
65

1 190 using the NEBNext Ultra RNA Library Prep Kit for Illumina. 150 bp PE sequencing
2
3
4 191 was performed using an Illumina HiSeq X Ten.
5

6 192 1,344 million raw reads from RNA sequencing were processed by Trimmomatic
7
8
9 193 and Cutadapt and aligned to the genome assembly with HiSat2 [23]. Base quality was
10
11
12 194 checked with FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>)
13
14
15 195 before and after data cleaning, and respective statistics of RNA sequencing data are
16
17
18 196 shown in **Table S1**. Reference genome guided transcriptome assemblies were
19
20
21 197 independently prepared with Cufflinks [24], StringTie [25] and Trinity [26]. *De novo*
22
23
24 198 assembly was generated using Trinity, then, transcriptome assemblies were combined
25
26
27 199 and further refined using CD-HIT [27], and finally, 192,169 unique transcripts were
28
29
30 200 gained. The summary of the transcriptome assemblies is shown in **Table S5**.
31

32 201 AUGUSTUS [28] was employed for *ab initio* gene prediction, using model training
33
34
35 202 based on coding sequences from *Arabidopsis thaliana* and *S. miltiorrhiza* (with two
36
37
38 203 sets of proteins from independent genome annotation [8, 9]). Then, transcripts from
39
40
41 204 RNA sequencing were aligned to the repeat-masked reference genome assembly with
42
43
44 205 BlastN and TblastX [29] (E-value cutoff of 10^{-5}), and protein sequences from *A.*
45
46
47 206 *thaliana* and *S. miltiorrhiza* were aligned to the repeat-masked reference genome
48
49
50 207 assembly with BlastX (E-value cutoff of 10^{-5}). After optimization with Exonerate
51
52
53 208 (<https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate>) [30], gene
54
55
56 209 model predictions were finalized prepared using the MAKER package [31] provided
57
58
59 210 within AUGUSTUS. To assess the quality of the gene prediction, AED (Annotation
60
61
62
63
64
65

1 211 Edit Distance) scores were generated for each of the predicted genes as part of the
2
3
4 212 MAKER pipeline. Putative function for each identified gene was assessed by
5
6 213 performing a BLAT [32] search of the peptide sequences against the Uniprot database
7
8
9 214 [33], Protein annotation against PFAM [34] and InterProScan [35] ID were also
10
11
12 215 conducted using the scripts provided in the MAKER package. Completeness of gene
13
14
15 216 annotation was checked using BUSCO (956 single copy orthologs of the Viridiplantae
16
17
18 217 database) with a BLAST E-value cutoff of 10^{-5} .

20
21 218 54,008 genes could be predicted, with average lengths of gene regions, genes
22
23
24 219 (including 5', 3' UTRs, exons and introns), CDS and exons of 3,430.43 bp, 1696.34 bp,
25
26
27 220 1293.62 bp and 265.94 bp, respectively (**Table S6**). The comparisons among genomes
28
29
30 221 from related species regarding lengths of genes, exons, and introns are shown in **Fig. 2**.
31
32
33 222 The distribution of AED tagged by MAKER is shown in **Fig. S3**, in which about 97%
34
35
36 223 of the annotated genes (52,338 genes) had an AED < 0.5 (**Table S6**), thus indicating
37
38
39 224 that the annotation is well supported. The result from BUSCO assessment of the quality
40
41
42 225 of the genome assembly and annotation is shown in **Table S7**. 92.08 % of the universal
43
44
45 226 single-copy genes (1,326 genes out of the total 1,440 genes) were identified, supporting
46
47
48 227 the high quality of the genome assembly. Among the 1,326 BUSCO conserved single-
49
50
51 228 copy genes detected in the scarlet genome, 466 genes were found single-copy, while
52
53
54 229 860 genes were duplicated (**Table S7**).

55
56 230 The predicted genes were annotated against several functional databases,
57
58
59 231 including: (1) the NCBI non-redundant protein database (Nr;
60
61
62
63
64
65

1 232 <http://www.ncbi.nlm.nih.gov>), (2) Swiss-Prot protein database
2
3
4 233 (<http://www.expasy.ch/sprot>) [33], (3) Translated EMBL-Bank (part of the
5
6 234 International Nucleotide Sequence Database Collaboration, TrEMBL,
7
8
9 235 <http://www.ebi.ac.uk/uniprot>) [33], (4) the protein families database (Pfam;
10
11
12 236 <http://pfam.xfam.org/>), (5) Cluster of Orthologous Groups for eukaryotic complete
13
14
15 237 genomes (KOG) database (<http://genome.jgi-psf.org/help/kogbrowser.jsf>), (6) KO (the
16
17
18 238 Kyoto Encyclopedia of Genes and Genomes, Orthology) database
19
20
21 239 (<http://www.genome.jp/kegg/ko.html>) [36], and (7) Gene ontology (GO)
22
23
24 240 (<http://www.geneontology.org>) [37]. 94.67 % of all predicted genes could be annotated
25
26
27 241 with the following protein related databases: NR (94.60 %), Swiss-Prot (63.40 %),
28
29
30 242 TrEMBL (93.50 %), Pfam (82.10 %), KOG (90.05 %), KO (37.40 %), and GO (78.80
31
32 243 %) (**Table S8**).

33
34
35 244

36 37 38 245 **Identification of orthologous genes and phylogenetic inference**

39
40
41 246 To analyze gene families, we downloaded the protein sequences of 15 additional species
42
43
44 247 (*Salvia miltiorrhiza* [8, 9], *Fraxinus excelsior* [38], *Olea europaea* [39], *Mimulus*
45
46
47 248 *guttatus* [40], *Utricularia gibba* [41], *Sesamum indicum* [42], *Coffea canephora* [43],
48
49
50 249 *Solanum lycopersicum* [44], *Daucus carota* [45], *Vitis vinifera* [46], *Arabidopsis*
51
52
53 250 *thaliana* [47], *Populus trichocarpa* [48], *Oryza sativa* [49] and *Beta vulgaris* [50])
54
55
56 251 (**Table S9**). Orthologous and paralogous gene clusters were identified among species
57
58
59 252 using OrthoMCL [51]. Recommended settings were used for all-against-all BLASTP
60
61
62
63
64
65

1 253 comparisons (Blast+ v2.3.056) [29] and OrthoMCL [51] analyses.

2
3 254 A total of 35,808 OrthoMCL families were built based on effective database sizes
4
5
6 255 of all versus all BLASTP with an E-value of 10^{-5} and a Markov Chain Clustering
7
8
9 256 default inflation parameter. We identified 1,306 gene families (3,797 genes) that were
10
11
12 257 specific to the scarlet sage genome when comparing with the other 15 genomes (**Table**
13
14
15 258 **S10**), and we detected 10,770 gene families that have expanded in the scarlet sage
16
17
18 259 lineage, using CAFE [52] (**Fig. 2c**). The expanded gene families were enriched for 60
19
20
21 260 significant ($q < 0.05$) GO-terms of three different functional categories, i.e. BP, CC, and
22
23
24 261 MF (**Table S11**) and one KEGG pathway (amino acid metabolism) (**Table S12**)
25
26
27 262 significant at $q < 0.05$. Also, 3,579 genes and 78 gene families were detected to be
28
29
30 263 contracted and found to have rapidly evolved within the scarlet sage genome (**Fig. 2c**).
31
32 264 Subsequently, 134 orthologous proteins among the 16 analyzed genomes were acquired
33
34
35 265 and aligned with MUSCLE v3.8.31 [53] employing default settings. A maximum
36
37
38 266 likelihood phylogenetic tree was then generated using the concatenated amino acid
39
40
41 267 sequences in PhyML 3.0 [54] with GTR+G+I model. The divergence time was
42
43
44 268 estimated with r8s [55] and calibrated against the timing of divergence between *A.*
45
46
47 269 *thaliana* and *V. vinifera* (124 Mya) [56] as well as against *A. thaliana* and *P. trichocarpa*
48
49
50 270 divergence time (90 Mya) [57]. The phylogenetic analysis identified the close
51
52
53 271 relationship among the three *Salvia* genomes and their divergence time was estimated
54
55
56 272 with about 28.21 Mya (**Fig. 2c**)

57
58 273 In summary, we presented the draft assembly for the scarlet sage genome using a
59
60
61
62
63
64
65

1 274 PacBio long-read dominated strategy, which was responsible for obtaining the high
2
3
4 275 sequence assembly quality. Also, the almost complete homozygosity within the
5
6 276 sequenced inbred line's genome was a key factor for the high continuity gained in this
7
8
9 277 study. The novel genome data generated in the present study will provide a valuable
10
11
12 278 resource for studying the molecular underpinnings of the various phenotypic variation
13
14
15 279 found within *Salvia sp.*, and sets the foundation for molecular-informed breeding
16
17
18 280 strategies and genome editing approaches for this valued ornamental flowering plant.
19
20
21 281 Moreover, this genome assembly is useful for comparative genomic studies among
22
23
24 282 related species.

25
26
27 283

28 284 **Availability of supporting data**

29
30
31
32 285 The genome assembly, annotations, and other supporting data are available via the
33
34
35 286 GigaScience database GigaDB. The raw sequence data have been deposited in the Short
36
37
38 287 Read Archive (SRA) under NCBI BioProject ID PRJNA422035.

39
40
41 288

42 43 289 **Abbreviations**

44
45
46 290 bp: base pair; kb: kilobases; Mb: megabases; Gb: gigabases; TE: transposable element;
47
48
49 291 BUSCO: benchmarking universal single-copy orthologs; CDS: coding sequence.

50
51
52 292

53 54 55 293 **Acknowledgement**

56
57
58 294 This study was funded by Beijing Key Laboratory of Green Plants Breeding and
59
60
61
62
63
64
65

1 295 Fundamental Research Funds for the Central Universities (NO.YX2013-41).

2
3
4 296

5
6 297 **Author Contributions**

7
8
9 298 AXD, HBX, RCC, JFM, FM and IP conceived and designed the study; AXD, HBX,

10
11
12 299 ZJL, HL, YQS, SN, ZNZ, RFC, HLZ, RGZ and QZY prepared the materials and

13
14
15 300 conducted the experiments; JFM, HBX, FM, IP wrote the manuscript.

16
17
18 301

19
20
21 302 **Conflict of Interest**

22
23
24 303 The authors declare that they have no competing financial interests.

25
26
27 304

28
29 305 **References**

30
31 306 1. Drew BT, González-Gallegos JG, Xiang C-L, Kriebel R, Drummond CP,
32 307 Walker JB, et al. *Salvia* united: The greatest good for the greatest number.
33
34 308 *Taxon*. 2017;66 1:133-45. doi:10.12705/661.7.

35
36 309 2. Sutton J. *The Gardener's Guide to Growing Salvias*. David & Charles; 1999.

37
38 310 3. Clebsch B and Barner CD. *The New Book of Salvias: Sages for Every Garden*.
39 311 Timber Press; 2003.

40 312 4. Walker JB, Sytsma KJ, Treutlein J and Wink M. *Salvia* (Lamiaceae) is not
41
42 313 monophyletic: implications for the systematics, radiation, and ecological
43 314 specializations of *Salvia* and tribe Mentheae. *American Journal of Botany*.
44 315 2004;91 7:1115-25. doi:10.3732/ajb.91.7.1115.

45
46 316 5. Griffiths M and Society RH. *Index of Garden Plants*. Macmillan; 1994.

47
48 317 6. Regnault-Roger C. The potential of botanical essential oils for insect pest
49 318 control. *Integrated Pest Management Reviews*. 1997;2 1:25-34.
50 319 doi:10.1023/a:1018472227889.

51
52 320 7. Ge X, Chen H, Wang H, Shi A and Liu K. De novo assembly and annotation of
53 321 *salvia splendens* transcriptome using the Illumina platform. *PLoS One*. 2014;9
54 322 3:e87693. doi:10.1371/journal.pone.0087693.

55
56 323 8. Zhang G, Tian Y, Zhang J, Shu L, Yang S, Wang W, et al. Hybrid de novo
57 324 genome assembly of the Chinese herbal plant danshen (*Salvia miltiorrhiza*
58 325 Bunge). *GigaScience*. 2015;4 1:62. doi:10.1186/s13742-015-0104-3.

60
61
62
63
64
65

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
- 326 9. Xu H, Song J, Luo H, Zhang Y, Li Q, Zhu Y, et al. Analysis of the Genome
327 Sequence of the Medicinal Plant *Salvia miltiorrhiza*. *Molecular Plant*. 2016;9
328 6:949-52. doi:10.1016/j.molp.2016.03.010.
 - 329 10. Vining KJ, Johnson SR, Ahkami A, Lange I, Parrish AN, Trapp SC, et al. Draft
330 Genome Sequence of *Mentha longifolia* and Development of Resources for
331 Mint Cultivar Improvement. *Molecular Plant*. 2017;10 2:323-39.
332 doi:10.1016/j.molp.2016.10.018.
 - 333 11. Bolger AM, Lohse M and Usadel B. Trimmomatic: a flexible trimmer for
334 Illumina sequence data. *Bioinformatics*. 2014;30 15:2114-20.
335 doi:10.1093/bioinformatics/btu170.
 - 336 12. Martin M. Cutadapt removes adapter sequences from high-throughput
337 sequencing reads. *EMBnetjournal*. 2011;17 1 doi:10.14806/ej.17.1.200 pp. 10-
338 12.
 - 339 13. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH and Phillippy AM.
340 Canu: scalable and accurate long-read assembly via adaptive k-mer weighting
341 and repeat separation. *Genome Research*. 2017; doi:10.1101/gr.215087.116.
 - 342 14. Marçais G and Kingsford C. A fast, lock-free approach for efficient parallel
343 counting of occurrences of k-mers. *Bioinformatics*. 2011;27 6:764-70.
344 doi:10.1093/bioinformatics/btr011.
 - 345 15. Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, et al. Estimation of genomic
346 characteristics by analyzing k-mer frequency in *de novo* genome projects. arXiv
347 preprint arXiv:13082012. 2013.
 - 348 16. Xiao C-L, Chen Y, Xie S-Q, Chen K-N, Wang Y, Han Y, et al. MECAT: fast
349 mapping, error correction, and de novo assembly for single-molecule
350 sequencing reads. *Nature Methods*. 2017;14:1072. doi:10.1038/nmeth.4432.
 - 351 17. Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al.
352 Phased diploid genome assembly with single-molecule real-time sequencing.
353 *Nature Methods*. 2016;13 12:1050-4. doi:10.1038/nmeth.4035.
 - 354 18. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV and Zdobnov EM.
355 BUSCO: assessing genome assembly and annotation completeness with single-
356 copy orthologs. *Bioinformatics*. 2015;31 19:3210-2.
 - 357 19. Boetzer M and Pirovano W. SSPACE-LongRead: scaffolding bacterial draft
358 genomes using long read sequence information. *BMC Bioinformatics*. 2014;15
359 1:211. doi:10.1186/1471-2105-15-211.
 - 360 20. Boetzer M, Henkel CV, Jansen HJ, Butler D and Pirovano W. Scaffolding pre-
361 assembled contigs using SSPACE. *Bioinformatics*. 2011;27 4:578-9.
362 doi:10.1093/bioinformatics/btq683.
 - 363 21. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an
364 empirically improved memory-efficient short-read *de novo* assembler.
365 *GigaScience*. 2012;1:18-. doi:10.1186/2047-217X-1-18.
 - 366 22. Bao W, Kojima KK and Kohany O. Repbase Update, a database of repetitive
367 elements in eukaryotic genomes. *Mobile DNA*. 2015;6 1:11.

368 doi:10.1186/s13100-015-0041-9.

- 1 369 23. Kim D, Langmead B and Salzberg SL. HISAT: a fast spliced aligner with low
2 370 memory requirements. *Nature Methods*. 2015;12 4:357-60.
- 3 371 24. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al.
4 372 Transcript assembly and quantification by RNA-Seq reveals unannotated
5 373 transcripts and isoform switching during cell differentiation. *Nature*
6 374 *Biotechnology*. 2010;28 5:511-5. doi:10.1038/nbt.1621.
- 7 375 25. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT and Salzberg SL.
8 376 StringTie enables improved reconstruction of a transcriptome from RNA-seq
9 377 reads. *Nature Biotechnology*. 2015;33 3:290-5.
- 10 378 26. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al.
11 379 Full-length transcriptome assembly from RNA-Seq data without a reference
12 380 genome. *Nature Biotechnology*. 2011;29:644. doi:10.1038/nbt.1883.
- 13 381 27. Fu L, Niu B, Zhu Z, Wu S and Li W. CD-HIT: accelerated for clustering the
14 382 next-generation sequencing data. *Bioinformatics*. 2012;28 23:3150-2.
- 15 383 28. Stanke M, Diekhans M, Baertsch R and Haussler D. Using native and
16 384 syntenically mapped cDNA alignments to improve *de novo* gene finding.
17 385 *Bioinformatics*. 2008;24 5:637-44. doi:10.1093/bioinformatics/btn013.
- 18 386 29. Boratyn GM, Schäffer AA, Agarwala R, Altschul SF, Lipman DJ and Madden
19 387 TL. Domain enhanced lookup time accelerated BLAST. *Biology Direct*. 2012;7
20 388 1:12. doi:10.1186/1745-6150-7-12.
- 21 389 30. Slater GSC and Birney E. Automated generation of heuristics for biological
22 390 sequence comparison. *BMC Bioinformatics*. 2005. doi:10.1186/1471-2105-6-
23 391 31.
- 24 392 31. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, et al. MAKER: an
25 393 easy-to-use annotation pipeline designed for emerging model organism
26 394 genomes. *Genome Res*. 2008;18 1:188-96. doi:10.1101/gr.6743907.
- 27 395 32. Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res*. 2002;12 4:656-
28 396 64. doi:10.1101/gr.229202.
- 29 397 33. Bairoch A and Apweiler R. The SWISS-PROT protein sequence database and
30 398 its supplement TrEMBL in 2000. *Nucleic Acids Research*. 2000;28 1:45-8.
- 31 399 34. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, et al. The Pfam
32 400 Protein Families Database. *Nucleic Acids Research*. 2002;30 1:276-80.
- 33 401 35. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al.
34 402 InterProScan: protein domains identifier. *Nucleic Acids Research*. 2005;33 Web
35 403 Server issue:W116-W20. doi:10.1093/nar/gki442.
- 36 404 36. Kanehisa M and Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes.
37 405 *Nucleic Acids Research*. 2000;28 1:27-30.
- 38 406 37. The Gene Ontology C, Ashburner M, Ball CA, Blake JA, Botstein D, Butler H,
39 407 et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*.
40 408 2000;25 1:25-9. doi:10.1038/75556.
- 41 409 38. Sollars ESA, Harper AL, Kelly LJ, Sambles CM, Ramirez-Gonzalez RH,

1 410 Swarbreck D, et al. Genome sequence and genetic diversity of European ash
2 411 trees. *Nature*. 2016;541:212. doi:10.1038/nature20786.

3 412 39. Unver T, Wu Z, Sterck L, Turktas M, Lohaus R, Li Z, et al. Genome of wild
4 413 olive and the evolution of oil biosynthesis. *Proceedings of the National*
5 414 *Academy of Sciences*. 2017;114 44:E9413-E22. doi:10.1073/pnas.1708621114.

6 415 40. Hellsten U, Wright KM, Jenkins J, Shu S, Yuan Y, Wessler SR, et al. Fine-scale
7 416 variation in meiotic recombination in *Mimulus* inferred from population
8 417 shotgun sequencing. *Proceedings of the National Academy of Sciences*.
9 418 2013;110 48:19478-82. doi:10.1073/pnas.1319032110.

10 419 41. Lan T, Renner T, Ibarra-Laclette E, Farr KM, Chang T-H, Cervantes-Pérez SA,
11 420 et al. Long-read sequencing uncovers the adaptive topography of a carnivorous
12 421 plant genome. *Proceedings of the National Academy of Sciences*. 2017;114
13 422 22:E4435-E41. doi:10.1073/pnas.1702072114.

14 423 42. Wang L, Yu S, Tong C, Zhao Y, Liu Y, Song C, et al. Genome sequencing of the
15 424 high oil crop sesame provides insight into oil biosynthesis. *Genome Biology*.
16 425 2014;15 2:R39. doi:10.1186/gb-2014-15-2-r39.

17 426 43. Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, et
18 427 al. The coffee genome provides insight into the convergent evolution of caffeine
19 428 biosynthesis. *Science*. 2014;345 6201:1181-4. doi:10.1126/science.1255274.

20 429 44. Consortium TTG. The tomato genome sequence provides insights into fleshy
21 430 fruit evolution. *Nature*. 2012;485:635. doi:10.1038/nature11119.

22 431 45. Iorizzo M, Ellison S, Senalik D, Zeng P, Satapoomin P, Huang J, et al. A high-
23 432 quality carrot genome assembly provides new insights into carotenoid
24 433 accumulation and asterid genome evolution. *Nature Genetics*. 2016;48:657.
25 434 doi:10.1038/ng.3565.

26 435 46. The French–Italian Public Consortium for Grapevine Genome C. The grapevine
27 436 genome sequence suggests ancestral hexaploidization in major angiosperm
28 437 phyla. *Nature*. 2007;449:463. doi:10.1038/nature06148.

29 438 47. Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S and Town
30 439 CD. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference
31 440 genome. *The Plant Journal*. 2017;89 4:789-804. doi:10.1111/tpj.13415.

32 441 48. Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, et al.
33 442 The Genome of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray).
34 443 *Science*. 2006;313 5793:1596-604. doi:10.1126/science.1128691.

35 444 49. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, et al. The TIGR
36 445 Rice Genome Annotation Resource: improvements and new features. *Nucleic*
37 446 *Acids Research*. 2007;35 suppl_1:D883-D7. doi:10.1093/nar/gkl976.

38 447 50. Dohm JC, Minoche AE, Holtgräwe D, Capella-Gutiérrez S, Zakrzewski F, Tafer
39 448 H, et al. The genome of the recently domesticated crop plant sugar beet (*Beta*
40 449 *vulgaris*). *Nature*. 2013;505:546. doi:10.1038/nature12817.

41 450 51. Li L, Stoeckert CJ and Roos DS. OrthoMCL: Identification of Ortholog Groups
42 451 for Eukaryotic Genomes. *Genome Research*. 2003;13 9:2178-89.

452 doi:10.1101/gr.1224503.

453 52. De Bie T, Cristianini N, Demuth JP and Hahn MW. CAFE: a computational tool
454 for the study of gene family evolution. *Bioinformatics*. 2006;22 10:1269-71.
455 doi:10.1093/bioinformatics/btl097.

456 53. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high
457 throughput. *Nucleic Acids Research*. 2004;32 5:1792-7.
458 doi:10.1093/nar/gkh340.

459 54. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W and Gascuel O.
460 New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies:
461 Assessing the Performance of PhyML 3.0. *Systematic Biology*. 2010;59 3:307-
462 21. doi:10.1093/sysbio/syq010.

463 55. Sanderson MJ. r8s: inferring absolute rates of molecular evolution and
464 divergence times in the absence of a molecular clock. *Bioinformatics*. 2003;19
465 2:301-2.

466 56. Doyle JA. Molecular and Fossil Evidence on the Origin of Angiosperms.
467 *Annual Review of Earth and Planetary Sciences*. 2012;40 1:301-26.
468 doi:10.1146/annurev-earth-042711-105313.

469 57. Wang H, Moore MJ, Soltis PS, Bell CD, Brockington SF, Alexandre R, et al.
470 Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proceedings*
471 *of the National Academy of Sciences*. 2009;106 10:3853-8.
472 doi:10.1073/pnas.0813376106.

473

474

475

476

477

478

479

480

481

482

483

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 484 **Supplemental Figures**

2
3 485 **Figure S1.** Length distribution of PacBio subreads.

4
5
6 486 **Figure S2.** K-mer frequency distribution at k-mer size of 17. A k-mer refers to an
7
8
9 487 artificial sequence division of K nucleotides. From k-mer frequency, genomic
10
11
12 488 characteristics (genome size, repeat structure and heterozygous rate) could be
13
14
15 489 estimated. Peaks at depths of 31 and 62 were annotated with dash lines.

16
17
18 490 **Figure S3.** Distribution of AED scores from gene prediction. AED, Annotation Edit
19
20
21 491 Distance, AED = 0 indicates perfect agreement between annotation and the evidence;
22
23
24 492 AED = 1 indicates no evidence support for annotation.

25
26
27 493 **Figure S4.** Length distribution of annotated genes, exons and introns. **a-c** for
28
29 494 annotated genes, exons and introns from different genome assemblies.

30
31
32 495

33
34
35 496 **Supplemental Tables**

36
37
38 497 **Table S1.** Summary of Pacbio and Illumina sequencing data generated in the present
39
40
41 498 study. IDs of the study, sample, library and accession in NCBI SRA and sequencing
42
43
44 499 platform, material origins of the sequenced DNA or RNA, the statistics of the raw and
45
46
47 500 cleaned data are shown.

48
49 501 **Table S2.** Estimation of genome characteristics based on 17-mer statistics.

50
51
52 502 **Table S3.** Statistics of the different versions of the genome assembly of the scarlet
53
54
55 503 sage. NA: data not available; * statistics for contigs/scaffolds.

56
57
58 504 **Table S4.** Summary of the annotated interspersed repeats in the genome assembly of
59
60
61
62
63
64
65

1 505 the scarlet sage.

2
3
4 506 **Table S5.** Summary of the transcriptome assemblies.

5
6 507 **Table S6.** Summary of the annotated genes. AED: Annotation Edit Distance; gene

7
8
9 508 regions (including UTRs, exons and introns); genes (including 5', 3' UTRs, exons and
10
11
12 509 introns).

13
14
15 510 **Table S7.** Summary of BUSCO evaluation of gene prediction.

16
17
18 511 **Table S8.** Summary of functional annotation of predicted genes.

19
20
21 512 **Table S9.** Genomic data used for gene families analyses. Origins, download links,
22
23
24 513 assembly versions, genome properties and references of 15 analyzed genomes are
25
26 514 shown.

27
28
29 515 **Table S10.** Summary of gene family analyses. Unique groups and genes, single-copy
30
31
32 516 and duplicated groups and genes are summarized for the 16 analyzed genomes of 15
33
34
35 517 plant species.

36
37
38 518 **Table S11.** GO enrichment of expanded gene families. (A) 'Category' is the Gene

39
40
41 519 Ontology (GO) term ID; (B) 'p_value' is the over represented p-value indicating the

42
43
44 520 observed frequency of a given term among analyzed genes is equal to the expected

45
46
47 521 frequency based on the null distribution; i.e. lower p-values indicate stronger evidence

48
49
50 522 for overrepresentation; (C) 'q_value' is the Benjamini and Hochberg adjusted p-value,

51
52
53 523 (D) 'numEPInCat' is the number of expanded gene families in the corresponding GO

54
55
56 524 category; (E) 'numInCat' is the number of detected gene families in the corresponding

57
58
59 525 GO category; (F) 'Term' is the GO term; (G) 'Ontology' indicates which ontology the

60
61
62
63
64
65

1 526 term comes from. 60 significant ($q < 0.05$) GO-terms of three different functional
2
3
4 527 categories are indicated in bold.
5
6 528 **Table S12.** KEGG enrichment of expanded gene families. (A) 'KO category' is the
7
8
9 529 KEGG Orthology (KO) category ID; (B) 'p_value' is the over represented p-value
10
11
12 530 indicating the observed frequency of a given term among analyzed genes is equal to the
13
14
15 531 expected frequency based on the null distribution; i.e. lower p-values indicate stronger
16
17
18 532 evidence for overrepresentation; (C) 'q_value' is the Benjamini and Hochberg adjusted
19
20
21 533 p-value, (D) 'numEPInCat' is the number of expanded gene families in the
22
23
24 534 corresponding KO category; (E) 'numInCat' is the number of detected gene families in
25
26
27 535 the corresponding KO category; (F) 'Pathway' is the KEGG pathway; (G) 'Class'
28
29
30 536 indicates which KEGG class the pathway comes from. One significant ($q < 0.05$) KEGG
31
32
33 537 pathway is indicated in bold.

34
35 538

36
37
38 539

39
40
41 540

42
43
44 541

45
46
47 542

48
49
50 543

51
52
53 544

54
55
56 545

57
58
59 546

60
61
62
63
64
65

1 547 **Tables**

2
3
4 548 **Table 1.** Statistics of the final genome assembly of the scarlet sage.

	Contig		Scaffold	
	Size (bp)	Number	Size (bp)	Number
Total Size	807,514,799	-	809,159,598	-
Total Number	-	2,204	-	1,525
N10	6,529,455	10	8,157,631	9
N50	2,267,074	100	3,123,266	73
N90	265,262	456	433,303	324
Max.	10,812,588	-	12,944,193	-
Min.	500	-	9,495	-
Mean	366,386	-	530,596	-
Median	38,049	-	48,557	-
Gap	-	-	1,644,799 (0.2%)	679
GC Content	38.84%	-	38.76%	-

26 549

27
28
29 550

30
31
32 551

33
34
35 552

36
37
38 553

39
40
41 554

42
43
44 555

45
46
47 556

48
49
50 557

51
52
53 558

54
55
56 559

57
58
59 560

60
61
62
63
64
65

1 561 **Figures:**

2
3
4 562 **Fig. 1** Images of the scarlet sage, *Salvia splendens*.

5
6 563 **a-b**, flowers of the sequenced cultivar of *S. splendens*, "Aoyunshenghuo (Olympic
7
8
9 564 flame)"; **c**, the scarlet sage with different flower colors in bedding; **d-k**, the scarlet
10
11
12
13 565 sage with flowers of different pure colors or bi-colors.

14
15 566

16
17
18 567 **Fig. 2** Quality of scarlet sage genome assembly and the phylogenomic inferences.

19
20
21 568 Quality was assessed by comparing the scarlet genome with the recently released

22
23
24 569 genomes of related species. Length distribution of contigs (**a**) and scaffolds (**b**); **c**,

25
26
27 570 phylogenetic tree, divergence time, and profiles of gene families that underwent

28
29
30 571 expansion or contraction. *Salvia miltiorrhiza* Zhang [8] and *Salvia miltiorrhiza* Xu [9]

31
32 572 are two genome assemblies reported for *Salvia miltiorrhiza*.

33
34
35 573

36
37
38 574

39
40
41 575

42
43
44 576

45
46
47 577

48
49
50 578

51
52 579

53
54 580

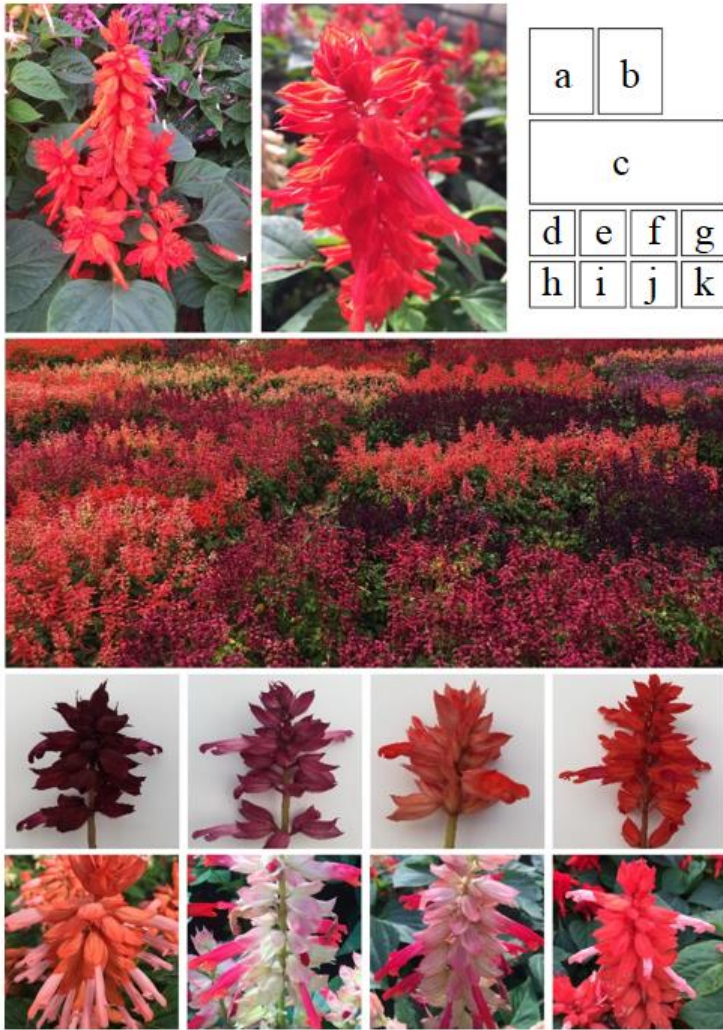
55 581

56 582

57 583

60
61
62
63
64
65

584 **Fig. 1**



585

586

587

588

589

590

591

592

593

594

595

596

597

598

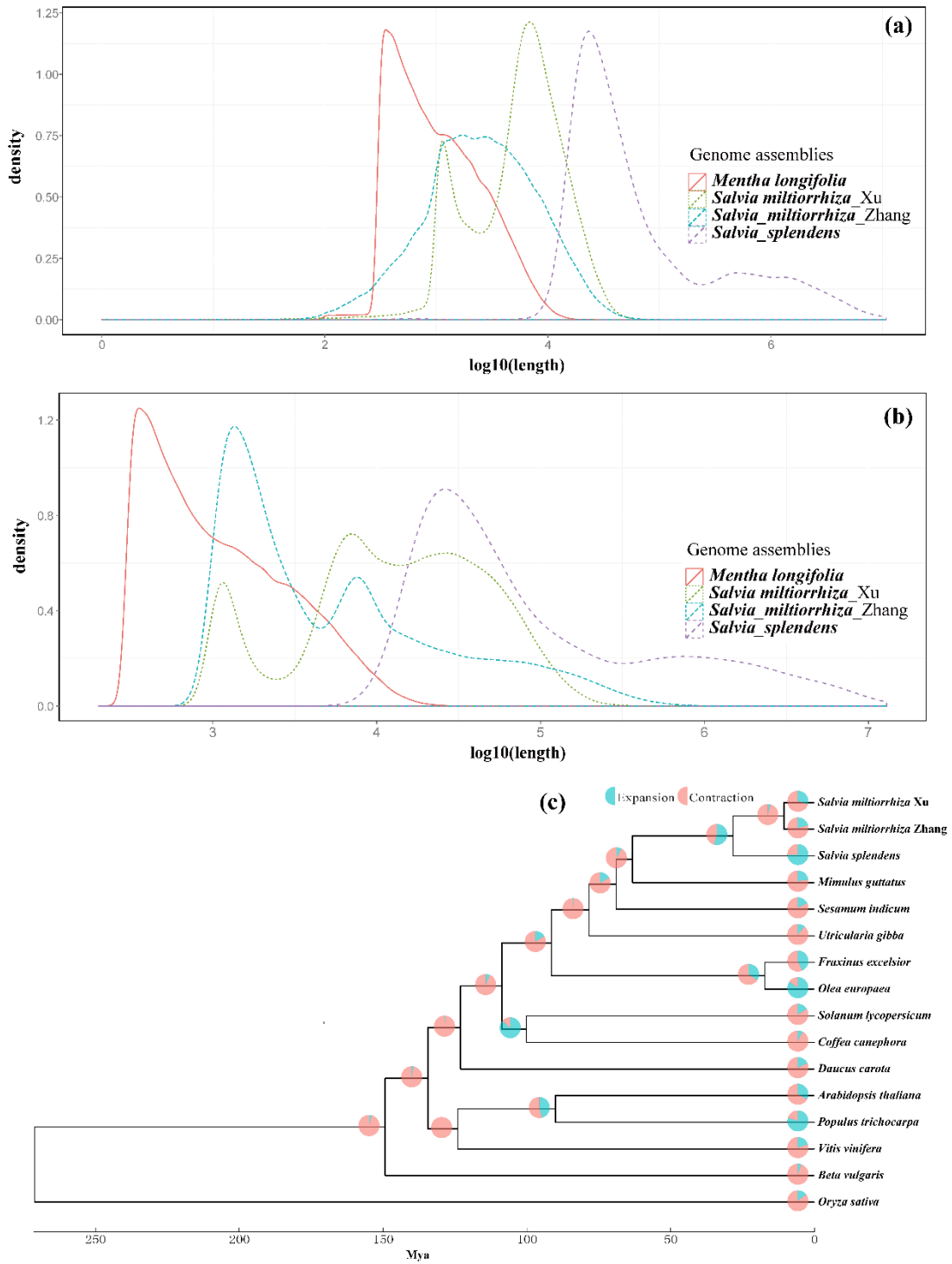
599

600

601


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

602 **Fig. 2**




603


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65



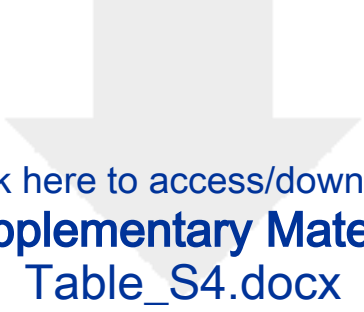
Click here to access/download
Supplementary Material
Table_S1.xlsx




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

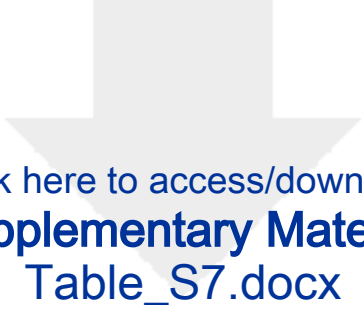




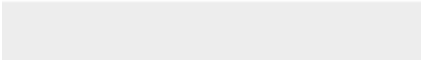

Click here to access/download
Supplementary Material
Table_S5.docx




Click here to access/download
Supplementary Material
Table_S6.docx




Click here to access/download
Supplementary Material
Table_S7.docx

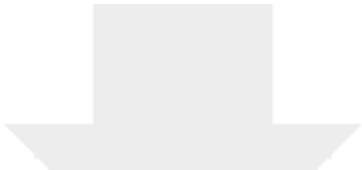





Click here to access/download
Supplementary Material
Table_S8.docx




Click here to access/download
Supplementary Material
Table_S9.docx

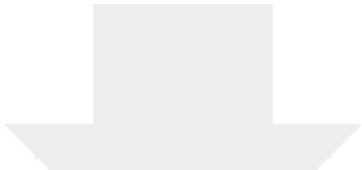


Click here to access/download
Supplementary Material
Table_S10.docx




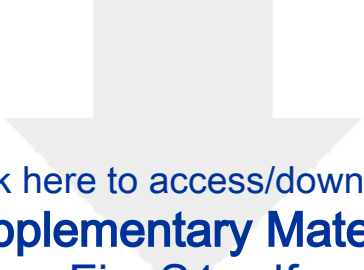


Click here to access/download
Supplementary Material
Table_S11.xlsx

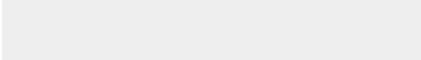




Click here to access/download
Supplementary Material
Table_S12.docx





Click here to access/download
Supplementary Material
Fig_S1.pdf



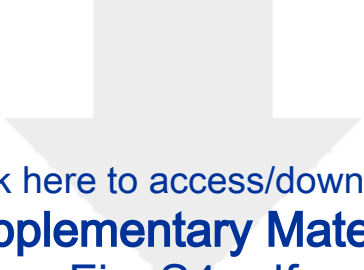


Click here to access/download
Supplementary Material
Fig_S2.pdf




Click here to access/download
Supplementary Material
Fig_S3.pdf





Click here to access/download
Supplementary Material
Fig_S4.pdf



January 26, 2018

Dr. Laurie Goodman, Editor of GigaScience

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

Dear Dr. Goodman,

Attached, please find the above-mentioned manuscript that my colleagues and I are submitting to **GigaScience** for possible publication as a **Data Note**.

The scarlet or tropical sage (*Salvia splendens*) is a tender herbaceous perennial widely introduced all over the world for ornamental purposes thus representing considerable economic value. Currently, few molecular resources exist for this species, and thus, improvement is still restricted to traditional phenotypic selection. In order to further improve selection for advantageous traits, the genetic mechanisms underlying phenotypic variation need to be further explored. Here, we provide a comprehensive new resource for *Salvia* genomics research based on: (1) PacBio Single-Molecule Real-Time (SMRT) and Illumina short-read sequencing providing long- and short-reads respectively, useful for *Salvia* high quality genome assembly, (2) genes and whole genome annotations with comprehensive bioinformatics computation (genetic diversity, DNA repeats annotations) and the help of a large set of RNA sequences obtained from multiple tissues, (3) gene family evolution characterisations (expansion; contraction) and phylogenomic analyses (estimated times of divergence) using a representative set of 14 additional angiosperm species. The availability of these resources will prove to be of great importance for further breeding strategies of *Salvia*, genome editing and also for comparative genomics among related species.

We attest that this manuscript has not been submitted to any other journal for publication. We also confirm that all the listed coauthors contributed to the study, and have read and approved the manuscript and are free from any conflicts of interest.

We look forward to hearing from you.

Sincerely Yours,
Jian-Feng Mao

Beijing Forestry University,
Qinghua East Road No 35.
Beijing, 100083,
P. R. China.