

GigaScience

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

--Manuscript Draft--

Manuscript Number:	GIGA-D-18-00028R1	
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 Greening Plants Breeding (Z201605)	Dr. Hai-Bo Xin
	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:	Ai-Xiang Dong	
First Author Secondary Information:		
Order of Authors:	Ai-Xiang Dong	
	Hai-Bo Xin	
	Zi-Jing Li	
	Hui Liu	
	Yan-Qiang Sun	
	Shuai Nie	

	Zheng-Nan Zhao
	Rong-Feng Cui
	Ren-Gang Zhang
	Quan-Zheng Yun
	Xin-Ning Wang
	Fatemeh Maghuly
	Ilga Porth
	Ri-Chen Cong
	Jian-Feng Mao, Ph.D.
Order of Authors Secondary Information:	
Response to Reviewers:	<p>We provide our comments directly underneath the points raised by you and within the three reviewers' reports as follows:</p> <p>AE: Please pay particular attention to reviewer #2's comment number 3: "Since the genomes of <i>Salvia miltiorrhiza</i> (Zhang et al. and Xu et al.) and <i>Mentha longifolia</i> have been published, a more detailed analysis about differences between <i>Salvia splendens</i> and the other two plants should be conducted, so as to highlight the importance of <i>Salvia splendens</i>. "</p> <p>R: We provided synteny analyses among detected metabolic gene cluster between the <i>Salvia</i> genomes. One section of comparative genomics was added (also see Figure S7 for synteny blocks). However, even <i>mentha</i> 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 <i>mentha</i> genome was not included in our comparative genomic studies.</p> <p>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 <i>Salvia</i> and <i>Mentha</i> genomes - what are similarities and differences, and what are unique features of <i>Salvia splendens</i>.</p> <p>R: please see answer provided above.</p> <p>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.</p> <p>R: Please see answers to Reviewer 3.</p> <p>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.</p> <p>R: Revised. Now we have 3 co-first authors.</p> <p>Reviewer #1: The authors of "High quality assembly of the reference genome for scarlet sage, <i>Salvia splendens</i>, an economically important ornamental plant" describe their efforts in generating a reference sequence for the plant <i>Salvia splendens</i> 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.</p> <p>R: Thank you.</p> <p>Reviewer #2: This manuscript described the construction of genome sequence and annotation for <i>Salvia splendens</i> 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</p>

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

	<p>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.</p> <p>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.</p> <p>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.</p> <p>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.</p>
Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics	Yes
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?	
Resources	Yes
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.	

<p>Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?</p>	
<p>Availability of data and materials</p> <p>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.</p> <p>Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?</p>	<p>Yes</p>

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^{1,2‡}, Zi-Jing Li^{1‡}, Hui Liu², Yan-Qiang Sun², Shuai Nie²,
5 Zheng-Nan Zhao¹, Rong-Feng Cui¹, Ren-Gang Zhang³, Quan-Zheng Yun³, Xin-Ning Wang³,
6 Fatemeh Maghuly⁴, Ilga Porth⁵, Ri-Chen Cong^{1*}, Jian-Feng Mao^{2*}

7
8 ¹ Beijing Key Laboratory of Greening 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. ORCID:
18 0000-0001-5433-0070.

19 ⁵ Département des sciences du bois et de la forêt, Pavillon Charles-Eugène-Marchand,
20 1030, Avenue de la Médecine, Université Laval, Québec (Québec) G1V 0A6, Canada.
21 ORCID: 0000-0002-9344-6348

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

22 ‡These authors contributed equally to this paper.

23 *Correspondence to: hardhopeee@163.com (RCC), ORCID: 0000-0002-4619-6120;

24 jianfeng.mao@bjfu.edu.cn (JFM), ORCID: 0000-0001-9735-8516

25 **Abstract**

26 **Background:** *Salvia splendens* Ker-Gawler, scarlet or tropical sage, is a tender
27 herbaceous perennial widely introduced and seen in public gardens all over the world.

28 With few molecular resources, breeding is still restricted to traditional phenotypic
29 selection, and the genetic mechanisms underlying phenotypic variation still remain
30 unknown. Hence, a high quality reference genome will be very valuable for marker
31 assisted breeding, genome editing or molecular genetics.

32 **Findings:** We generated 66 gigabases (Gb) and 37 Gb of raw DNA sequences,
33 respectively, from whole-genome sequencing of a largely homozygous scarlet sage
34 inbred line using PacBio Single-Molecule Real-Time (SMRT) and Illumina HiSeq
35 sequencing platforms. PacBio *de novo* assembly yielded a final genome with a
36 scaffold N50 size of 3.12 megabases (Mb), and a total length of 808 Mb. The
37 repetitive sequences identified accounted for 57.52% of the genome sequence and
38 54,008 protein-coding genes were predicted collectively with *ab initio* and
39 homology-based gene prediction from the masked genome. The divergence time
40 between *S. splendens* and *S. miltiorrhiza* was estimated with 28.21 million years ago
41 (Mya). Moreover, 3,797 species-specific genes and 1,187 expanded gene families
42 were identified for the scarlet sage genome.

1 43 **Conclusions:** We provide the first genome sequence and gene annotation for the
2
3
4 44 scarlet sage. The availability of these resources will be of great importance for further
5
6 45 breeding strategies, genome editing and also for comparative genomics among related
7
8
9 46 species.

10
11
12 47 **Keywords:** annotation, evolution, reference genome, *Salvia splendens*, scarlet sage
13
14

15 48 **Data description**

16 17 18 49 **Background information**

19
20
21 50 *Salvia* L., with nearly 1,000 species of shrubs, herbaceous perennials, and annuals, is
22
23
24 51 the largest genus in the mint family (Lamiaceae: Nepetoideae: Mentheae: Salviinae)
25
26
27 52 [1-4]. The genus is widely distributed throughout the world. Many species of this
28
29
30 53 genus are extensively used for culinary purposes, essential oil production and Chinese
31
32
33 54 herbal remedies such as the two species *S. officinalis* [3] and *S. miltiorrhiza*
34
35
36 55 (*Danshen*). Additionally, they are used as ornamental plants valued for their flowers or
37
38
39 56 for their aromatic foliage such as *S. splendens* (**Fig. 1 a-k**).

40
41 57 *S. splendens* (NCBI taxon ID:180675), scarlet or tropical sage, is a herbaceous
42
43
44 58 perennial species, which is native to Brazil. While it is a perennial in warmer climate
45
46
47 59 zones, it grows as an annual in cooler areas. *S. splendens* is a very popular bedding
48
49
50 60 plant, and is widely cultivated in public gardens all over the world [3, 5],
51
52
53 61 characterized by its dense flowers, and wide variation of colours (scarlet, purple, pink,
54
55
56 62 blue, lavender, salmon, yellow green, white and bicolor), as well as long lasting
57
58
59 63 flowering (3-9 weeks or even longer). Additionally, *S. splendens* can provide
60
61
62
63
64
65

1 64 outstanding visual effects when grown in beds, borders and containers with
2
3
4 65 long-lasting lifespans ranging from late spring to the occurrence first frost.
5
6 66 Furthermore, the flower is easy to maintain and fairly free of pests and diseases due to
7
8
9 67 Lamiaceae's characteristic insect repellent fragrance content [6]. The plant blends
10
11
12 68 nicely with other annuals or perennial plants for the best visual effects in an ensemble
13
14
15 69 setting; in addition this plant requires little deadheading as well it attracts various
16
17
18 70 butterfly species. *S. splendens* is a prolific and durable bloomer, thrives in full sun,
19
20
21 71 and survives in a large range of soil moisture regimes.

22
23 72 Traditional breeding activities using phenotypic selection as well as performing
24
25
26 73 targeted variety hybridizations between elite cultivars have resulted in a large number
27
28
29 74 of new cultivars with different performances regarding flowering characters (related
30
31
32 75 to colour, flowering time, flowering period amongst others), individual growth
33
34
35 76 performance, height, and/or tolerance to moisture or temperature extremes. However,
36
37
38 77 little is known about the molecular mechanisms underlying such economically
39
40
41 78 important characteristics for ornamental varieties. To date, only few genetic markers
42
43
44 79 [7] are available for marker assistant breeding or genetic modification.

45
46
47 80 In the current study, we present the first high quality genome assembly for *S.*
48
49
50 81 *splendens* with a hybrid assembly strategy using PacBio Single-Molecule Real-Time
51
52
53 82 and Illumina HiSeq short-read sequencing platforms. The genome assembly, its
54
55
56 83 structural and functional annotation, provide a valuable reference for the genomic
57
58
59 84 dissection of the phenotypic variation in *Salvia*, and new breeding strategies. This
60
61
62
63
64
65

1 85 reference genome could also be used in comparative genomics with the recently
2
3
4 86 released *Salvia* genome (*S. miltiorrhiza*) [8, 9] and the mint genome (*Mentha*
5
6 87 *longifolia*) [10] to study the biosynthesis of important fragrant and medicinal
7
8
9 88 compounds.

10
11
12 89

13 14 15 90 **Plant material**

16
17
18 91 We chose the elite variety *S. splendens*, "Aoyunshenghuo (Olympic flame)" (**Fig. 1**
19
20
21 92 **a-b**) for whole genome sequencing, which was originally developed by multiple
22
23
24 93 rounds of selection/selfing of one hybrid to obtain this inbred line. This cultivar is
25
26
27 94 characterized by resistance to drought, high temperature, and improved performance
28
29
30 95 related to a longer flowering period; it is well adapted to climate conditions across
31
32
33 96 North China, and therefore grows well in Beijing. Because of the high homozygosity
34
35
36 97 obtained due to advanced generation selfing, this cultivar shows no phenotypic
37
38
39 98 segregation, a characteristic of important commercial value. Seeds of this cultivar
40
41
42 99 were provided by the Beijing Institute of Landscape Architecture germplasm bank.

43
44 100

45 46 47 101 **PacBio SMRT sequencing**

48
49
50 102 High quality high molecular weight genomic DNA was extracted from leaves of two
51
52
53 103 soil-grown seedlings (huo1 and huo1_1) following ~20 kb SMRTbell™ Libraries”
54
55
56 104 protocol
57
58 105 (<http://www.pacb.com/wp-content/uploads/2015/09/Shared-Protocol-Preparing-Arabi>

59
60
61
62
63
64
65

1 106 dopsis-DNA-for-20-kb-SMRTbell-Libraries.pdf). Plants for DNA extraction have
2
3
4 107 been placed in the dark for 48 h before harvesting the leaf material. DNA was purified
5
6 108 with Mobio PowerClean® Pro DNA Clean-Up Kit and quality was assessed by
7
8
9 109 standard agarose gel electrophoresis and Thermo Fisher Scientific Qubit Fluorometry.
10
11 110 Genomic DNA was sheared to a size range of 15–40 kb using either AMPure beads
12
13 111 (Beckman Coulte) or g-TUBE (Covaris), and enzymatically repaired and converted
14
15
16 112 into SMRTbell template libraries as recommended by Pacific Biosciences. Following
17
18 113 this procedure, hairpin adapters were ligated following exonucleasese-based digestion
19
20
21 114 (of the remaining damaged DNA fragments and those fragments without adapters at
22
23
24 115 both ends). Subsequently, the resulting SMRTbell templates were size-selected by
25
26
27 116 Blue Pippin electrophoresis (Sage Sciences) and templates ranging from 15 to 50 kb
28
29
30 117 were sequenced on a PacBio RS II instrument using P6-C4 sequencing chemistry (25
31
32
33 118 Single-Molecule Real-Time (SMRT) cells for individual huo1) and on a PacBio
34
35
36 119 Sequel instrument using S/P2-C2 sequencing chemistry (8 SMRT cells for the other
37
38
39 120 individual, huo1_1). A total of 8,858,116 PacBio post-filtered reads were generated.
40
41
42 121 This produced a total of 65,962,079,028 bp (roughly 82x of the assembled genome) of
43
44
45 122 single-molecule sequencing data, with an average read length of 7,446 bp (**Fig. S1**
46
47
48 123 and **Table S1**).

51
52
53 124

54 55 125 **Illumina short-read sequencing**

56
57
58 126 DNA was extracted from leaf tissue of the same soil-grown seedlings (huo1 and
59
60
61
62
63
64
65

1 127 huo1_1) using the Qiagen DNeasy Plant Mini Kit. Two 500 bp paired-end (PE)
2
3
4 128 libraries (huo1 and huo1_1) were prepared using the NEBNext Ultra DNA Library
5
6 129 Prep Kit for Illumina sequencing with an Illumina HiSeq X Ten machine. Short reads
7
8
9 130 were processed with Trimmomatic v0.33 (Trimmomatic, RRID:SCR_011848) [11,
10
11
12 131 12] and Cutadapt v1.13 (cutadapt, RRID:SCR_011841) [13, 14] to remove adapter
13
14
15 132 sequences and leading and trailing bases with a quality score below 20 and reads with
16
17
18 133 an average per-base-quality of 20 over a 4 bp sliding window. Reads < 70 nucleotides
19
20
21 134 in length after trimming were removed from further analysis. A total of 265.53 million
22
23
24 135 reads were generated. This produced a total of 36.83 Gb (roughly 40x of the
25
26
27 136 assembled genome) of raw sequencing data, with an average cleaned read length of
28
29
30 137 137 bp (**Table S1**).

31
32 138

33 34 35 139 **Estimation of genome size, heterozygosity, and repeat content**

36
37
38 140 All generated PacBio reads were filtered and corrected with Canu v1.5 (Canu,
39
40
41 141 RRID:SCR_015880) [15], thereafter, Jellyfish (Jellyfish, RRID:SCR_005491) [16]
42
43
44 142 was used to count occurrence of k-mers based on the processed data. Finally, gce
45
46
47 143 1.0.0 [17] was employed to estimate the overall characteristics of the genome such as
48
49
50 144 genome size, repeat contents and heterozygous rate. In this study, a total of
51
52
53 145 22,117,819,357 k-mers were generated and the peak k-mer depth was 31 (**Fig. S2**).
54
55
56 146 The genome size was estimated to be approximately 711 Mb (**Table S2**) and the final
57
58
59 147 cleaned data corresponded to the coverage of about 33-fold. Repeat and error rates

1 148 were estimated to be 47.99% and 0.27%, respectively, and heterozygosity rate was
2
3
4 149 0.06%.

5
6 150

7
8
9 151 ***De novo* genome assembly**

10
11
12 152 *De novo* assembly was conducted as follows in a progressive manner. Firstly, primary
13
14
15 153 assemblies were generated from PacBio long reads of the 31 Gb from the ‘huo1’
16
17
18 154 sequenced individual by four different Overlap-Layout-Consensus (OLC) based
19
20
21 155 assemblers, Canu (produced assembly v0.1), MECAT 1.1 (assembly v0.2) [18],
22
23
24 156 FALCON v0.7 (Falcon, RRID:SCR_016089) [19]
25
26
27 157 (<https://github.com/PacificBiosciences/FALCON/>) after Canu correction (v0.3) and
28
29
30 158 SMARTdenovo 1.0.0 (<https://github.com/ruanjue/smardtenovo>) after Canu correction
31
32
33 159 (v0.4) (**Table S3**). Based on the size of the assembled genome, the total number of
34
35
36 160 assembled contigs, N50, the L50, maximum length of the contigs, and also the
37
38
39 161 completeness of the genome assembly as assessed by using BUSCO criteria v2.0.1
40
41
42 162 (BUSCO, RRID:SCR_015008)[20] (1,440 single copy orthologs of the Viridiplantae
43
44
45 163 database) with the BLAST E-value cutoff of 10^{-5} , assembly (v0.1) from Canu was
46
47
48 164 chosen for further polishing and scaffolding. In this selected primary assembly, the
49
50
51 165 assembled genome size was 808 Mb distributed across 2,306 contigs with N50 of 2.06
52
53
54 166 Mb, L50 of 109 and maximum contig length of 8.88 Mb. We also confirmed on
55
56
57 167 average 92.1% gene completeness in this assembly (**Table S3**). In the following steps,
58
59
60 168 the arrow algorithm v2.2.1

1 169 (<https://github.com/PacificBiosciences/GenomicConsensus>) was used to further
2
3
4 170 improve the assembly based on PacBio long reads (v1.0), after which
5
6 171 SSPACE-LongRead 1.1 [21] and SSPACE-standard 3.0 (SSPACE,
7
8
9 172 RRID:SCR_005056) [22] were used for subsequent scaffold assembly based on
10
11
12 173 PacBio long reads of 35 Gb from the second sequenced individual 'hou1_1' and
13
14
15 174 Illumina short reads, respectively. Finally, after scaffold processing and subsequent
16
17
18 175 gap filling with SOAPdenovo and GapCloser (GapCloser, RRID:SCR_015026) [23]
19
20
21 176 (v1.1), arrow v2.2.1 algorithm (based on PacBio long reads) and pilon (Pilon,
22
23
24 177 RRID:SCR_014731) (based on Illumina short reads, and run two times, parameters
25
26
27 178 for Pilon: --changes --diploid --dumpreads.), we got the final genome assembly (v1.2).
28
29
30 179 Mapping of Illumina reads was done by using Bowtie2 v2.3.0 (Bowtie,
31
32
33 180 RRID:SCR_005476) [24]. We detected 400,170 SNPs, 96,854 insertions and 62,637
34
35
36 181 deletions, respectively, for the first pilon run, while, subsequently, a greatly decreased
37
38
39 182 number of variants for the second pilon run (40,465 SNPs, 6,935 insertions and 9,976
40
41
42 183 deletions, respectively). In this final assembly, we gained an assembled genome size
43
44
45 184 of 808 Mb characterized by 2,204 contigs and 1,525 scaffolds (with contig N50 of
46
47
48 185 2.27 Mb and scaffold N50 of 3.12 Mb), and by gene completeness of 92.2% (**Table 1**
49
50
51 186 **and Table S3**). This assembly represents the highest continuity and completeness
52
53
54 187 among the recently released genome assemblies for the *Salvia* genus [8, 9] and the
55
56
57 188 mint [10], as it was examined by length distribution plotting of contigs and scaffolds
58
59
60 189 as shown in **Fig. 2a, b**.

1 190

2
3
4 191 **DNA repeats annotation**

5
6 192 RepeatModeler v1.0.10 (RepeatModeler, RRID:SCR_015027)

7
8
9 193 (<http://www.repeatmasker.org/RepeatModeler.html>) was employed to *de novo* identify

10
11
12 194 and classify repeat families in the genome assembly. Subsequently, the outputs from

13
14
15 195 RepeatModeler and RepBase [25] library were combined and used as repeat library

16
17
18 196 for subsequent RepeatMasker (RepeatMasker, RRID:SCR_012954) (v4.0.7,

19
20
21 197 rmblast-2.2.28) (<http://www.repeatmasker.org/>) analyses, which was used to fully

22
23
24 198 discover and identify repeats within the assembled genome. In summary, 57.52% of

25
26
27 199 the genome was annotated as repeats among which we found 1.08% simple repeats

28
29
30 200 and 40.35% known transposable elements (TE). Long terminal repeats (LTRs)

31
32
33 201 constituted the greatest proportion, 26.49% of the genome, and DNA TE made up

34
35
36 202 11.91% of the genome. Gypsy (18.15% of the genome) and Copia (7.92%) TEs were

37
38
39 203 the largest components of LTRs. The results of repeat annotations are summarized in

40
41 204 **Table S4.**

42
43
44 205

45
46
47 206 **RNA sequencing, transcriptome assembly and functional annotation**

48
49
50 207 RNA was extracted from the two cultivated lines with different flower colours (red

51
52
53 208 and purple) using tissue obtained from, roots, shoots, leaves, calyxes and corollas.

54
55
56 209 Frozen tissue from all samples was ground manually using mortar and pestil, and

57
58
59 210 RNA was isolated using the NEBNext Poly(A) mRNA Magnetic Isolation Module.

60
61
62
63
64
65

1 211 RNA quality was assessed using an Agilent 2100 BioAnalyzer. Sequencing libraries
2
3
4 212 were prepared using the NEBNext Ultra RNA Library Prep Kit for Illumina. 150 bp
5
6 213 PE sequencing was performed using an Illumina HiSeq X Ten.

7
8
9 214 1,344 million raw reads from RNA sequencing were processed by Trimmomatic
10
11
12 215 and Cutadapt and aligned to the genome assembly with HiSat2 v2.1.0 (HiSat2,
13
14
15 216 RRID:SCR_015530) [26]. Base quality was checked with FastQC (FastQC,
16
17
18 217 RRID:SCR_014583) (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>)
19
20
21 218 before and after data cleaning, and respective statistics of RNA sequencing data are
22
23
24 219 shown in **Table S1**. Reference genome guided transcriptome assemblies were
25
26
27 220 independently prepared with Cufflinks v2.1.1 (Cufflinks, RRID:SCR_014597) [27],
28
29
30 221 StringTie v1.3.3b (StringTie, RRID:SCR_016323)[28] and Trinity v2.0.6 (Trinity,
31
32
33 222 RRID:SCR_013048) [29]. *De novo* assembly was generated using Trinity, then,
34
35
36 223 transcriptome assemblies were combined and further refined using CD-HIT v4.6 [30],
37
38
39 224 and finally, 192,169 unique transcripts were gained. The summary of the
40
41
42 225 transcriptome assemblies is shown in **Table S5**.

43
44 226 AUGUSTUS v3.2.3 (Augustus, RRID:SCR_008417) [31] was employed for *ab*
45
46
47 227 *initio* gene prediction, using model training based on coding sequences from
48
49
50 228 *Arabidopsis thaliana* and *S. miltiorrhiza* (with two sets of proteins from independent
51
52
53 229 genome annotation [8, 9]). Then, transcripts from RNA sequencing were aligned to
54
55
56 230 the repeat-masked reference genome assembly with BlastN and TblastX from BLAST
57
58
59 231 v2.2.28+ (NCBI BLAST, RRID:SCR_004870) [32] (E-value cutoff of 10^{-5}), and
60
61
62
63
64
65

1 232 protein sequences from *A. thaliana* and *S. miltiorrhiza* were aligned to the
2
3
4 233 repeat-masked reference genome assembly with BlastX (E-value cutoff of 10^{-5}). After
5
6 234 optimization with Exonerate v2.4.0
7
8
9 235 (<https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate>) [33], gene
10
11
12 236 model predictions were finalized prepared using the MAKER package v2.31.9
13
14
15 237 (MAKER, RRID:SCR_005309) [34] provided within AUGUSTUS. To assess the
16
17
18 238 quality of the gene prediction, AED (Annotation Edit Distance) scores were generated
19
20
21 239 for each of the predicted genes as part of the MAKER pipeline. Putative function for
22
23
24 240 each identified gene was assessed by performing a BLAT (BLAT,
25
26
27 241 RRID:SCR_011919) [35] search of the peptide sequences against the Uniprot
28
29
30 242 database (UniProt, RRID:SCR_002380) [36], Protein annotation against PFAM
31
32
33 243 (Pfam, RRID:SCR_004726) [37] and InterProScan (InterProScan,
34
35
36 244 RRID:SCR_005829) [38] ID were also conducted using the scripts provided in the
37
38
39 245 MAKER package. Completeness of gene annotation was checked using BUSCO
40
41
42 246 (1,440 single copy orthologs of the Viridiplantae database) with a BLAST E-value
43
44 247 cutoff of 10^{-5} .

46 248 54,008 genes could be predicted, with average lengths of gene regions, genes
48
49
50 249 (including 5', 3' UTRs, exons and introns), CDS and exons of 3,430.43 bp, 1696.34
51
52
53 250 bp, 1293.62 bp and 265.94 bp, respectively (**Table S6**). The comparisons among
54
55
56 251 genomes from related species regarding lengths of genes, exons, and introns are
57
58
59 252 shown in **Fig. 2**. The distribution of AED tagged by MAKER is shown in **Fig. S3**, in

1 253 which about 97% of the annotated genes (52,338 genes) had an AED < 0.5 (**Table**
2
3
4 254 **S6**), thus indicating that the annotation is well supported. The result from BUSCO
5
6 255 assessment of the quality of the genome assembly and annotation is shown in **Table**
7
8
9 256 **S7**. 92.08 % of the universal single-copy genes (1,326 genes out of the total 1,440
10
11
12 257 genes) were identified, supporting the high quality of the genome assembly. Among
13
14
15 258 the 1,326 BUSCO conserved single-copy genes detected in the scarlet genome, 466
16
17
18 259 genes were found single-copy, while 860 genes were duplicated (**Table S7**).

20
21 260 The predicted genes were annotated against several functional databases,
22
23
24 261 including: (1) the NCBI non-redundant protein database (Nr;
25
26 262 <http://www.ncbi.nlm.nih.gov>), (2) Swiss-Prot protein database
27
28
29 263 (<http://www.expasy.ch/sprot>) [36], (3) Translated EMBL-Bank (part of the
30
31
32 264 International Nucleotide Sequence Database Collaboration, TrEMBL,
33
34
35 265 <http://www.ebi.ac.uk/uniprot>) [36], (4) the protein families database (Pfam;
36
37
38 266 <http://pfam.xfam.org/>), (5) Cluster of Orthologous Groups for eukaryotic complete
39
40
41 267 genomes (KOG) database (<http://genome.jgi-psf.org/help/kogbrowser.jsf>), (6) KO (the
42
43
44 268 Kyoto Encyclopedia of Genes and Genomes, Orthology) database
45
46
47 269 (<http://www.genome.jp/kegg/ko.html>) [39], and (7) Gene ontology (GO)
48
49
50 270 (<http://www.geneontology.org>) [40]. 94.67 % of all predicted genes could be
51
52
53 271 annotated with the following protein related databases: NR (94.60 %), Swiss-Prot
54
55
56 272 (63.40 %), TrEMBL (93.50 %), Pfam (82.10 %), KOG (90.05 %), KO (37.40 %), and
57
58
59 273 GO (78.80 %) (**Table S8**).

1 274

2
3
4 275 **Identification of orthologous genes and phylogenetic inference**

5
6 276 To analyze gene families, we downloaded the protein sequences of 15 additional
7
8
9 277 species (*Salvia miltiorrhiza* [8, 9], *Fraxinus excelsior* [41], *Olea europaea* [42],
10
11 278 *Mimulus guttatus* [43], *Utricularia gibba* [44], *Sesamum indicum* [45], *Coffea*
12
13 279 *canephora* [46], *Solanum lycopersicum* [47], *Daucus carota* [48], *Vitis vinifera* [49],
14
15 280 *Arabidopsis thaliana* [50], *Populus trichocarpa* [51], *Oryza sativa* [52] and
16
17 281 *Beta_vulgaris* [53]) (**Table S9**). Orthologous and paralogous gene clusters were
18
19 282 identified among species using OrthoMCL v2.0.9 [54]. Recommended settings were
20
21 283 used for all-against-all BLASTP comparisons (Blast+ v2.3.056) [32] and
22
23 284 OrthoMCL[22] analyses.

24
25
26
27
28
29
30
31
32 285 A total of 35,808 OrthoMCL families were built based on effective database sizes
33
34
35 286 of all versus all BLASTP with an E-value of 10^{-5} and a Markov Chain Clustering
36
37 287 default inflation parameter. We identified 1,306 gene families (3,797 genes) that were
38
39 288 specific to the scarlet sage genome when comparing with the other 15 genomes
40
41 289 (**Table S10**), and we detected 10,770 gene families that have expanded in the scarlet
42
43 290 sage lineage, using CAFE v4.0 [55, 56] (**Fig. 2c**). The expanded gene families were
44
45 291 enriched for 60 significant ($q < 0.05$) GO-terms of three different functional categories,
46
47 292 i.e. BP, CC, and MF (**Table S11**) and one KEGG pathway (amino acid metabolism)
48
49 293 (**Table S12**) significant at $q < 0.05$. Also, 3,579 genes and 78 gene families were
50
51 294 detected to be contracted and found to have rapidly evolved within the scarlet sage
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 295 genome (**Fig. 2c**). Subsequently, 134 orthologous proteins among the 16 analyzed
2
3
4 296 genomes were acquired and aligned with MUSCLE v3.8.31 (MUSCLE,
5
6 297 RRID:SCR_011812) [57] employing default settings. A maximum likelihood
8
9 298 phylogenetic tree was then generated using the concatenated amino acid sequences in
10
11
12 299 PhyML v3.0 (PhyML, RRID:SCR_014629) [58] with GTR+G+I model. The
13
14
15 300 divergence time was estimated with r8s v1.81 [59] and calibrated against the timing of
16
17
18 301 divergence between *A. thaliana* and *V. vinifera* (124 Mya) [60] as well as against *A.*
19
20
21 302 *thaliana* and *P. trichocarpa* divergence time (90 Mya) [61]. The phylogenetic analysis
22
23
24 303 identified the close relationship among the three *Salvia* genomes and their divergence
25
26
27 304 time was estimated with about 28.21 Mya (**Fig. 2c**).

28
29
30 305

306 **Secondary metabolic pathways: gene annotations, gene clusters and comparative** 33 34 35 307 **genomics**

36
37
38 308 The mint family is recognized as providing promising sources of bioactive secondary
39
40
41 309 metabolites [62]. In fact, a diverse variety of bioactive secondary metabolites can be
42
43
44 310 found with a wide range of pharmacological activities: antimicrobial, antispasmodic,
45
46
47 311 carminative, antioxidant, antiulcer, cytoprotective, hepatoprotective, cholagogue,
48
49
50 312 chemo-preventive, anti-inflammatory, antidiabetogenic etc. Here, we obtained
51
52
53 313 enzymatic annotations for coding genes by employing the E2P2 package v3.1
54
55
56 314 (<https://gitlab.com/rhee-lab/E2P2/tree/master>). Then, we mapped genes to flavonoid
57
58
59 315 and menthol biosynthesis pathways by querying the Plant Metabolic Network (PMN
60
61
62
63
64
65

1 316 v12.5, <https://www.plantcyc.org/> [63]. Regarding the flavonoid biosynthesis pathway,
2
3
4 317 we found an abundance of genes encoding annotated enzymes in this pathway,
5
6 318 especially of note the 41 genes for flavanone synthase I (EC: 1.14.11.9) (**Figure S5**
7
8
9 319 and **Supplementary_File_1**). With respect to menthol biosynthesis, certain genes are
10
11
12 320 still lacking annotations for enzymes such as (+)-pulegone reductase (EC: 1.3.1.81),
13
14
15 321 (-)-isopiperitenone reductase (EC: 1.3.1.82) or menthol-dehydrogenase (lacking EC
16
17
18 322 number) (**Figure S6** and **Supplementary_File_1**). However, this pathway mapping
19
20
21 323 analysis provides a highly valuable reference for the genetic dissection of key
22
23
24 324 metabolic genes for the scarlet sage.

25
26 325 The presence of metabolic gene clusters for secondary metabolic pathways are
27
28
29 326 common in bacteria and filamentous fungi, and are also widely reported in plants
30
31
32 327 [64-66]. Using the newly created and robust computational toolkit, plantSMASH [67],
33
34
35 328 we identified 85 gene clusters potentially related to secondary metabolic biosynthesis
36
37
38 329 in the scarlet sage genome as reported here, and 23 gene clusters in the *S. miltiorrhiza*
39
40
41 330 genome [8]. Genomic position, gene composition, functional annotation of the
42
43
44 331 identified gene clusters were summarized in **Table S13, Supplementary_File_2** and
45
46
47 332 **Supplementary_File_3**. The gene clusters were found to be potentially related to the
48
49
50 333 biosynthesis of alkaloids, saccharides, polyketides, terpenes, and lignans. It was
51
52
53 334 previously reported that physical clustering of terpene synthase genes (TPS) and
54
55
56 335 cytochrome P450 mono-oxygenase genes is frequently associated with consecutive
57
58
59 336 enzymatic actions in terpenoid biosynthesis [68]. Interestingly, we detected eight such
60
61
62
63
64
65

1 337 gene clusters within the scarlet sage genome, but none in the *S. miltiorrhiza* genome,
2
3
4 338 which could partially be due to the draft status of the genome assembly for *S.*
5
6 339 *miltiorrhiza*. Furthermore, significant gene co-expression across different organs was
7
8
9 340 detected for one TPS gene and two out of four P450 genes located in a single gene
10
11
12 341 cluster (i.e. Cluster 63; **Table S13** and **Supplementary_File_2**). Evidence for
13
14
15 342 moderate or significant co-expression among clustered genes was revealed and shown
16
17
18 343 in **Supplementary_File_2**.

19
20
21 344 Based on the collinearity elucidated by former OrthoMCL analyses, a
22
23
24 345 comparative genomic study between the scarlet sage and *S. miltiorrhiza* genomes
25
26
27 346 revealed six pairs of gene clusters which share synteny between these two congeneric
28
29
30 347 plants, and two blocks from the scarlet sage share synteny with one block from *S.*
31
32 348 *miltiorrhiza* (**Figure S7**). Among the shared synteny blocks, four could be related to
33
34
35 349 saccharide, one to lignan and another to polyketide biosynthesis. The smaller number
36
37
38 350 of gene clusters detected for *S. miltiorrhiza* and subsequently, fewer shared synteny
39
40
41 351 blocks of metabolic gene cluster between these two species may be partially attributed
42
43
44 352 to the present state of the *S. miltiorrhiza* genome assembly which is hundred times
45
46
47 353 more fragmented than that of the scarlet sage. Thus, here, we provided a starting point
48
49
50 354 for comparative genomics among plant species within the mint family.

51
52 355
53
54
55 356 In summary, we presented the draft assembly for the scarlet sage genome using a
56
57
58 357 PacBio long-read dominated strategy, which was responsible for obtaining the high
59
60
61
62
63
64
65

1 358 sequence assembly quality. Also, the almost complete homozygosity within the
2
3
4 359 sequenced inbred line's genome was a key factor for the high continuity gained in this
5
6 360 study. The novel genome data generated in the present study will provide a valuable
7
8
9 361 resource for studying the molecular underpinnings of the various phenotypic variation
10
11
12 362 found within *Salvia sp.*, and sets the foundation for molecular-informed breeding
13
14
15 363 strategies and genome editing approaches for this valued ornamental flowering plant.
16
17
18 364 Moreover, this genome assembly is useful for comparative genomic studies among
19
20
21 365 related species.

22
23
24 366

25 26 367 **Availability of supporting data**

27
28
29 368 The genome assembly, annotations, and other supporting data are available via the
30
31
32 369 *GigaScience* database GigaDB [69]. The raw sequence data have been deposited in
33
34
35 370 the Short Read Archive (SRA) under NCBI BioProject ID PRJNA422035.

36
37
38 371

39 40 41 372 **Abbreviations**

42
43
44 373 AED: Annotation Edit Distance; bp: base pair; BUSCO: benchmarking universal
45
46
47 374 single-copy orthologs; CDS: coding sequence; Gb: gigabases; kb: kilobases; LTR:
48
49
50 375 Long terminal repeats; Mb: megabases; Mya: million years ago; PE: paired end;
51
52
53 376 SMRT: Single-Molecule Real-Time; SNP: Single Nucleotide Polymorphism; TE:
54
55
56 377 TPS: transposable element; terpene synthase genes.

57
58 378
59
60
61
62
63
64
65

1 379 **Acknowledgement**

2
3
4 380 This study was funded by Beijing Key Laboratory of Greening Plants Breeding (NO.
5
6 381 Z201605) and Fundamental Research Funds for the Central Universities (NO.
7
8
9 382 YX2013-41).

10
11
12 383

13
14
15 384 **Author Contributions**

16
17
18 385 AXD, HBX, RCC, JFM, FM and IP conceived and designed the study; AXD, HBX,
19
20
21 386 ZJL, HL, YQS, SN, ZNZ, RFC, HLZ, RGZ and QZY prepared the materials and
22
23
24 387 conducted the experiments; JFM, HBX, FM, IP wrote the manuscript.

25
26 388

27
28
29 389 **Conflict of Interest**

30
31
32 390 The authors declare that they have no competing financial interests.

33
34
35 391

36
37
38 392 **References**

- 39
40 393 1. Drew BT, González-Gallegos JG, Xiang C-L, Kriebel R, Drummond CP,
41 394 Walker JB, et al. *Salvia* united: The greatest good for the greatest number.
42 395 *Taxon*. 2017;66 1:133-45.
43 396 2. Sutton J. *The Gardener's Guide to Growing Salvias*. David & Charles; 1999.
44 397 3. Clebsch B and Barner CD. *The New Book of Salvias: Sages for Every*
45 398 *Garden*. Timber Press; 2003.
46 399 4. Walker JB, Sytsma KJ, Treutlein J and Wink M. *Salvia* (Lamiaceae) is not
47 400 monophyletic: implications for the systematics, radiation, and ecological
48 401 specializations of *Salvia* and tribe Mentheae. *American Journal of Botany*.
49 402 2004;91 7:1115-25.
50 403 5. Griffiths M and Society RH. *Index of Garden Plants*. Macmillan; 1994.
51 404 6. Regnault-Roger C. The potential of botanical essential oils for insect pest
52 405 control. *Integrated Pest Management Reviews*. 1997;2 1:25-34.
53 406 7. Ge X, Chen H, Wang H, Shi A and Liu K. *De Novo Assembly and Annotation*

60
61
62
63
64
65

- 407 of *Salvia splendens* Transcriptome Using the Illumina Platform. PLoS One.
408 2014;9 3:e87693.
- 409 8. Zhang G, Tian Y, Zhang J, Shu L, Yang S, Wang W, et al. Hybrid de novo
410 genome assembly of the Chinese herbal plant danshen (*Salvia miltiorrhiza*
411 Bunge). GigaScience. 2015;4 1:62.
- 412 9. Xu H, Song J, Luo H, Zhang Y, Li Q, Zhu Y, et al. Analysis of the Genome
413 Sequence of the Medicinal Plant *Salvia miltiorrhiza*. Molecular Plant. 2016;9
414 6:949-52.
- 415 10. Vining KJ, Johnson SR, Ahkami A, Lange I, Parrish AN, Trapp SC, et al. Draft
416 Genome Sequence of *Mentha longifolia* and Development of Resources for
417 Mint Cultivar Improvement. Molecular Plant. 2017;10 2:323-39.
- 418 11. Bolger AM, Lohse M and Usadel B. Trimmomatic: a flexible trimmer for
419 Illumina sequence data. Bioinformatics. 2014;30 15:2114-20.
- 420 12. Alberto CM, Sanso AM and Xifreda CC. Chromosomal studies in species of
421 *Salvia* (Lamiaceae) from Argentina. Botanical Journal of the Linnean Society.
422 2003;141 4:483-90.
- 423 13. Martin M. Cutadapt removes adapter sequences from high-throughput
424 sequencing reads. EMBnetjournal. 2011;17 1: 10-12.
- 425 14. The Gene Ontology (GO) database and informatics resource. Nucleic Acids
426 Research. 2004;32 Database issue:D258-D61.
- 427 15. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH and Phillippy AM.
428 Canu: scalable and accurate long-read assembly via adaptive k-mer weighting
429 and repeat separation. Genome Research. 2017; 722-736.
- 430 16. Marçais G and Kingsford C. A fast, lock-free approach for efficient parallel
431 counting of occurrences of k-mers. Bioinformatics. 2011;27 6:764-70.
- 432 17. Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, et al. Estimation of genomic
433 characteristics by analyzing k-mer frequency in de novo genome projects.
434 arXiv preprint arXiv:13082012. 2013.
- 435 18. Xiao C-L, Chen Y, Xie S-Q, Chen K-N, Wang Y, Han Y, et al. MECAT: fast
436 mapping, error correction, and de novo assembly for single-molecule
437 sequencing reads. Nature Methods. 2017;14:1072.
- 438 19. Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al.
439 Phased diploid genome assembly with single-molecule real-time sequencing.
440 Nature Methods. 2016;13 12:1050-4.
- 441 20. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV and Zdobnov EM.
442 BUSCO: assessing genome assembly and annotation completeness with
443 single-copy orthologs. Bioinformatics. 2015;31 19:3210-2.
- 444 21. Boetzer M and Pirovano W. SSPACE-LongRead: scaffolding bacterial draft
445 genomes using long read sequence information. BMC Bioinformatics. 2014;15
446 1:211.
- 447 22. Boetzer M, Henkel CV, Jansen HJ, Butler D and Pirovano W. Scaffolding
448 pre-assembled contigs using SSPACE. Bioinformatics. 2011;27 4:578-9.

- 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
- 449 23. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an
450 empirically improved memory-efficient short-read de novo assembler.
451 GigaScience. 2012;1:18.
 - 452 24. Langmead B and Salzberg SL. Fast gapped-read alignment with Bowtie 2.
453 Nature Methods. 2012;9 4:357-9.
 - 454 25. Bao W, Kojima KK and Kohany O. Repbase Update, a database of repetitive
455 elements in eukaryotic genomes. Mobile DNA. 2015;6 1:11.
 - 456 26. Kim D, Langmead B and Salzberg SL. HISAT: a fast spliced aligner with low
457 memory requirements. Nature Methods. 2015;12 4:357-60.
 - 458 27. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et
459 al. Transcript assembly and quantification by RNA-Seq reveals unannotated
460 transcripts and isoform switching during cell differentiation. Nature
461 Biotechnology. 2010;28 5:511-5.
 - 462 28. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT and Salzberg
463 SL. StringTie enables improved reconstruction of a transcriptome from
464 RNA-seq reads. Nature Biotechnology. 2015;33 3:290-5.
 - 465 29. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al.
466 Full-length transcriptome assembly from RNA-Seq data without a reference
467 genome. Nature Biotechnology. 2011;29:644.
 - 468 30. Fu L, Niu B, Zhu Z, Wu S and Li W. CD-HIT: accelerated for clustering the
469 next-generation sequencing data. Bioinformatics. 2012;28 23:3150-2.
 - 470 31. Stanke M, Diekhans M, Baertsch R and Haussler D. Using native and
471 syntenically mapped cDNA alignments to improve de novo gene finding.
472 Bioinformatics. 2008;24 5:637-44.
 - 473 32. Boratyn GM, Schäffer AA, Agarwala R, Altschul SF, Lipman DJ and Madden
474 TL. Domain enhanced lookup time accelerated BLAST. Biology Direct.
475 2012;7 1:12.
 - 476 33. Slater GSC and Birney E. Automated generation of heuristics for biological
477 sequence comparison. BMC Bioinformatics. 2005; 6 1:1-11.
 - 478 34. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, et al. MAKER: an
479 easy-to-use annotation pipeline designed for emerging model organism
480 genomes. Genome Research. 2008;18 1:188-96.
 - 481 35. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Research. 2002;12
482 4:656-64.
 - 483 36. Bairoch A and Apweiler R. The SWISS-PROT protein sequence database and
484 its supplement TrEMBL in 2000. Nucleic Acids Research. 2000;28 1:45-8.
 - 485 37. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, et al. The
486 Pfam Protein Families Database. Nucleic Acids Research. 2002;30 1:276-80.
 - 487 38. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al.
488 InterProScan: protein domains identifier. Nucleic Acids Research. 2005;33
489 Web Server issue:W116-W20.
 - 490 39. Kanehisa M and Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes.

- 491 Nucleic Acids Research. 2000;28 1:27-30.
- 492 40. The Gene Ontology C, Ashburner M, Ball CA, Blake JA, Botstein D, Butler
493 H, et al. Gene Ontology: tool for the unification of biology. Nature Genetics.
494 2000;25 1:25-9.
- 495 41. Sollars ESA, Harper AL, Kelly LJ, Sambles CM, Ramirez-Gonzalez RH,
496 Swarbreck D, et al. Genome sequence and genetic diversity of European ash
497 trees. Nature. 2016;541:212.
- 498 42. Unver T, Wu Z, Sterck L, Turktas M, Lohaus R, Li Z, et al. Genome of wild
499 olive and the evolution of oil biosynthesis. Proceedings of the National
500 Academy of Sciences. 2017;114 44:E9413-E22.
- 501 43. Hellsten U, Wright KM, Jenkins J, Shu S, Yuan Y, Wessler SR, et al.
502 Fine-scale variation in meiotic recombination in *Mimulus* inferred from
503 population shotgun sequencing. Proceedings of the National Academy of
504 Sciences. 2013;110 48:19478-82.
- 505 44. Lan T, Renner T, Ibarra-Laclette E, Farr KM, Chang T-H, Cervantes-Pérez SA,
506 et al. Long-read sequencing uncovers the adaptive topography of a
507 carnivorous plant genome. Proceedings of the National Academy of Sciences.
508 2017;114 22:E4435-E41.
- 509 45. Wang L, Yu S, Tong C, Zhao Y, Liu Y, Song C, et al. Genome sequencing of
510 the high oil crop sesame provides insight into oil biosynthesis. Genome
511 Biology. 2014;15 2:R39.
- 512 46. Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, et
513 al. The coffee genome provides insight into the convergent evolution of
514 caffeine biosynthesis. Science. 2014;345 6201:1181-4.
- 515 47. Tomato Genome Consortium. The tomato genome sequence provides insights
516 into fleshy fruit evolution. Nature. 2012;485:635.
- 517 48. Iorizzo M, Ellison S, Senalik D, Zeng P, Satapoomin P, Huang J, et al. A
518 high-quality carrot genome assembly provides new insights into carotenoid
519 accumulation and asterid genome evolution. Nature Genetics. 2016;48:657.
- 520 49. The French-Italian Public Consortium for Grapevine Genome Consortium for
521 Grapevine Genome Characterization. The grapevine genome sequence
522 suggests ancestral hexaploidization in major angiosperm phyla. Nature.
523 2007;449:463.
- 524 50. Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S and
525 Town CD. Araport11: a complete reannotation of the *Arabidopsis thaliana*
526 reference genome. The Plant Journal. 2017;89 4:789-804.
- 527 51. Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, et al.
528 The Genome of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray).
529 Science. 2006;313 5793:1596-604.
- 530 52. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, et al. The TIGR
531 Rice Genome Annotation Resource: improvements and new features. Nucleic
532 Acids Research. 2007;35 suppl_1:D883-D7.

- 1 533 53. Dohm JC, Minoche AE, Holtgräwe D, Capella-Gutiérrez S, Zakrzewski F,
2 534 Tafer H, et al. The genome of the recently domesticated crop plant sugar beet
3 535 (*Beta vulgaris*). *Nature*. 2013;505:546.
- 4 536 54. Li L, Stoeckert CJ and Roos DS. OrthoMCL: Identification of Ortholog
5 537 Groups for Eukaryotic Genomes. *Genome Research*. 2003;13 9:2178-89.
- 6 538 55. De Bie T, Cristianini N, Demuth JP and Hahn MW. CAFE: a computational
7 539 tool for the study of gene family evolution. *Bioinformatics*. 2006;22
8 540 10:1269-71.
- 9 541 56. Badouin H, Gouzy J, Grassa CJ, Murat F, Staton SE, Cottret L, et al. The
10 542 sunflower genome provides insights into oil metabolism, flowering and
11 543 Asterid evolution. *Nature*. 2017;546:148.
- 12 544 57. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and
13 545 high throughput. *Nucleic Acids Research*. 2004;32 5:1792-7.
- 14 546 58. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W and Gascuel O.
15 547 New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies:
16 548 Assessing the Performance of PhyML 3.0. *Systematic Biology*. 2010;59
17 549 3:307-21.
- 18 550 59. Sanderson MJ. r8s: inferring absolute rates of molecular evolution and
19 551 divergence times in the absence of a molecular clock. *Bioinformatics*. 2003;19
20 552 2:301-2.
- 21 553 60. Doyle JA. Molecular and Fossil Evidence on the Origin of Angiosperms.
22 554 *Annual Review of Earth and Planetary Sciences*. 2012;40 1:301-26.
- 23 555 61. Wang H, Moore MJ, Soltis PS, Bell CD, Brockington SF, Alexandre R, et al.
24 556 Rosid radiation and the rapid rise of angiosperm-dominated forests.
25 557 *Proceedings of the National Academy of Sciences*. 2009;106 10:3853-8.
- 26 558 62. Mimica-Dukic N and Bozin B. *Mentha* L. Species (Lamiaceae) as Promising
27 559 Sources of Bioactive Secondary Metabolites. *Current Pharmaceutical Design*.
28 560 2008;14 29:3141-50.
- 29 561 63. Schläpfer P, Zhang P, Wang C, Kim T, Banf M, Chae L, et al. Genome-Wide
30 562 Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants.
31 563 *Plant Physiology*. 2017;173 4:2041-59.
- 32 564 64. Osbourn A. Secondary metabolic gene clusters: evolutionary toolkits for
33 565 chemical innovation. *Trends in Genetics*. 2010;26 10:449-57.
- 34 566 65. Nützmann H-W and Osbourn A. Gene clustering in plant specialized
35 567 metabolism. *Current Opinion in Biotechnology*. 2014;26:91-9.
- 36 568 66. Hans-Wilhelm N, Ancheng H and Anne O. Plant metabolic clusters – from
37 569 genetics to genomics. *New Phytologist*. 2016;211 3:771-89.
- 38 570 67. Kautsar SA, Suarez Duran HG, Blin K, Osbourn A and Medema MH.
39 571 plantiSMASH: automated identification, annotation and expression analysis of
40 572 plant biosynthetic gene clusters. *Nucleic Acids Research*. 2017;45
41 573 W1:W55-W63.
- 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
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

574 68. Boutanaev AM, Moses T, Zi J, Nelson DR, Mugford ST, Peters RJ, et al.
575 Investigation of terpene diversification across multiple sequenced plant
576 genomes. *Proceedings of the National Academy of Sciences*. 2015;112
577 1:E81-E8.

578 69. Dong, A; Xin, H; Li, Z; Liu, H; Sun, Y; Nie, S; Zhao, Z; Cui, R; Zhang, R;
579 Yun, Q; Wang, X; Maghuly, F; Porth, I; Cong, R; Mao, J (2018): Supporting
580 data for "High quality assembly of the reference genome for scarlet sage,
581 *Salvia splendens*, an economically important ornamental plant" GigaScience
582 Database. <http://dx.doi.org/10.5524/100463>.
583

1 584 **Supplementary Figures**

2
3 585 **Figure S1.** Length distribution of PacBio subreads.

4
5
6 586 **Figure S2.** K-mer frequency distribution at k-mer size of 17. A k-mer refers to an
7
8
9 587 artificial sequence division of K nucleotides. From k-mer frequency, genomic
10
11
12 588 characteristics (genome size, repeat structure and heterozygous rate) could be
13
14
15 589 estimated. Peaks at depths of 31 and 62 were annotated with dash lines.

16
17
18 590 **Figure S3.** Distribution of AED scores from gene prediction. AED, Annotation Edit
19
20
21 591 Distance, AED = 0 indicates perfect agreement between annotation and the evidence;
22
23
24 592 AED = 1 indicates no evidence support for annotation.

25
26
27 593 **Figure S4.** Length distribution of annotated genes, exons and introns. **a-c** for
28
29
30 594 annotated genes, exons and introns from different genome assemblies.

31
32 595 **Figure S5.** Flavonoid biosynthesis pathway. Flavonoid biosynthesis pathways by
33
34
35 596 querying the Plant Metabolic Network (<https://www.plantcyc.org/>), enzymatic coding
36
37
38 597 genes of the scarlet sage were shown for key reactions.

39
40
41 598 **Figure S6.** Menthol biosynthesis pathway. Menthol biosynthesis pathways by
42
43
44 599 querying the Plant Metabolic Network (<https://www.plantcyc.org/>), enzymatic coding
45
46
47 600 genes of the scarlet sage were shown for key reactions.

48
49
50 601 **Figure S7.** Shared synteny addressed for metabolic gene clusters between *Salvia*
51
52
53 602 genomes. a-f: display of the different pairs of synteny blocks. Genes are colored along
54
55
56 603 the contigs/scaffolds to compare between scarlet sage and *Salvia miltiorrhiza* Zhang
57
58
59 604 [8], with metabolic genes highlighted with olive drab color, other homologous genes

1 605 are shown in grey.

2
3
4 606

5
6 607 **Supplementary Tables**

7
8
9 608 **Table S1.** Summary of Pacbio and Illumina sequencing data generated in the present
10
11
12 609 study. IDs of the study, sample, library and accession in NCBI SRA and sequencing
13
14
15 610 platform, material origins of the sequenced DNA or RNA, the statistics of the raw and
16
17
18 611 cleaned data are shown.

19
20
21 612 **Table S2.** Estimation of genome characteristics based on 17-mer statistics.

22
23
24 613 **Table S3.** Statistics of the different versions of the genome assembly of the scarlet
25
26
27 614 sage. NA: data not available; * statistics for contigs/scaffolds.

28
29
30 615 **Table S4.** Summary of the annotated interspersed repeats in the genome assembly of
31
32
33 616 the scarlet sage.

34
35 617 **Table S5.** Summary of the transcriptome assemblies.

36
37
38 618 **Table S6.** Summary of the annotated genes. AED: Annotation Edit Distance; gene
39
40
41 619 regions (including UTRs, exons and introns); genes (including 5', 3' UTRs, exons and
42
43
44 620 introns).

45
46
47 621 **Table S7.** Summary of BUSCO evaluation of gene prediction.

48
49
50 622 **Table S8.** Summary of functional annotation of predicted genes.

51
52
53 623 **Table S9.** Genomic data used for gene families analyses. Origins, download links,
54
55
56 624 assembly versions, genome properties and references of 15 analyzed genomes are
57
58
59 625 shown.

60
61
62
63
64
65

1 626 **Table S10.** Summary of gene family analyses. Unique groups and genes, single-copy
2
3
4 627 and duplicated groups and genes are summarized for the 16 analyzed genomes of 15
5
6 628 plant species.

7
8
9 629 **Table S11.** GO enrichment of expanded gene families. (A) 'Category' is the Gene
10
11
12 630 Ontology (GO) term ID; (B) 'p_value' is the over represented p-value indicating the
13
14
15 631 observed frequency of a given term among analyzed genes is equal to the expected
16
17
18 632 frequency based on the null distribution; i.e. lower p-values indicate stronger evidence
19
20
21 633 for overrepresentation; (C) 'q_value' is the Benjamini and Hochberg adjusted p-value,
22
23
24 634 (D) 'numEPInCat' is the number of expanded gene families in the corresponding GO
25
26
27 635 category; (E) 'numInCat' is the number of detected gene families in the corresponding
28
29
30 636 GO category; (F) 'Term' is the GO term; (G) 'Ontology' indicates which ontology the
31
32
33 637 term comes from. 60 significant ($q < 0.05$) GO-terms of three different functional
34
35
36 638 categories are indicated in bold.

37
38 639 **Table S12.** KEGG enrichment of expanded gene families. (A) 'KO category' is the
39
40
41 640 KEGG Orthology (KO) category ID; (B) 'p_value' is the over represented p-value
42
43
44 641 indicating the observed frequency of a given term among analyzed genes is equal to
45
46
47 642 the expected frequency based on the null distribution; i.e. lower p-values indicate
48
49
50 643 stronger evidence for overrepresentation; (C) 'q_value' is the Benjamini and
51
52
53 644 Hochberg adjusted p-value, (D) 'numEPInCat' is the number of expanded gene
54
55
56 645 families in the corresponding KO category; (E) 'numInCat' is the number of detected
57
58
59 646 gene families in the corresponding KO category; (F) 'Pathway' is the KEGG pathway;

1 647 (G) 'Class' indicates which KEGG class the pathway comes from. One significant
2
3
4 648 ($q < 0.05$) KEGG pathway is indicated in bold.

5
6 649 **Table S13.** Summary of metabolic gene clusters detected in genomes of *Salvia*
7
8
9 650 *miltiorrhiza* and *S. splendens*. (A) 'Genome' denotes the genome origination; (B)
10
11
12 651 'Cluster' is the code for a certain gene cluster detected; (C) 'Record' denotes the
13
14
15 652 contig/scaffold ID from where the gene cluster was detected; (D) 'Type' denotes the
16
17
18 653 functional assignment for the gene cluster; (E) 'From', 'To' and 'Size' denote the
19
20
21 654 genomic position and range of the gene cluster; (F) 'Core domains' denote the domain
22
23
24 655 annotation for the metabolic genes in the cluster; (G) 'CD-HIT Cluster' indicate the
25
26
27 656 number of genes in the cluster; (H) 'Gene cluster genes' is showing the ID of genes in
28
29
30 657 the cluster.

31
32 658

33 34 35 659 **Supplementary Files**

36
37
38 660 **Supplementary File 1.** Genes (Gene ID, name and EC number) mapped to flavonoid
39
40
41 661 and menthol biosynthesis pathways.

42
43
44 662 **Supplementary File 2.** Structure of a metabolic gene cluster (polyketide synthesis)
45
46
47 663 and gene expression patterns of *Salvia splendens*. Genomic position, gene
48
49
50 664 composition, functional annotation of gene cluster are shown, also including a
51
52
53 665 heatmap of tissue specific expression of the genes within the presented cluster is
54
55
56 666 shown. HG: root of red flower (individual); HJ: stem of red flower (individual); HY:
57
58
59 667 leave of red flower (individual); HE: calyx of red flower (individual); HHG: corolla

1 668 of red flower (individual); ZG: root of purple flower (individual); ZJ: stem of purple
2
3
4 669 flower (individual); ZY: leave of purple flower (individual); ZE: calyx of purple
5
6 670 flower (individual); ZHG: corolla of purple flower (individual).
7
8

9 671 **Supplementary File 3.** Structure of a metabolic gene cluster (alkaloid synthesis).
10

11 672 Genomic position, gene composition, functional annotation of gene cluster were
12
13
14
15 673 shown.
16

17 674
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

1 675 **Tables**

2
3
4 676 **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
27 677

28
29
30 678

31
32
33 679

34
35
36 680

37
38
39 681

40
41 682

42
43
44 683

45
46
47 684

48
49
50 685

51
52
53 686

54
55
56 687

57
58
59 688

60
61
62
63
64
65

1 689 **Figures:**

2
3
4 690 **Fig. 1** Images of the scarlet sage, *Salvia splendens*.

5
6 691 **a-b**, flowers of the sequenced cultivar of *S. splendens*, "Aoyunshenghuo (Olympic
7
8
9 692 flame)"; **c**, the scarlet sage with different flower colors in bedding; **d-k**, the scarlet
10
11
12 693 sage with flowers of different pure colors or bi-colors.

13
14
15 694

16
17
18 695 **Fig. 2** Quality of scarlet sage genome assembly and the phylogenomic inferences.

19
20
21 696 Quality was assessed by comparing the scarlet genome with the recently released
22
23
24 697 genomes of related species. Length distribution of contigs (**a**) and scaffolds (**b**); **c**,
25
26
27 698 phylogenetic tree, divergence time, and profiles of gene families that underwent
28
29
30 699 expansion or contraction. *Salvia miltiorrhiza* Zhang [14] and *Salvia miltiorrhiza* Xu
31
32 700 [14] are two genome assemblies reported for *Salvia miltiorrhiza*.

33
34
35 701

36
37
38 702

39
40
41 703

42
43
44 704

45
46
47 705

48
49
50 706

51
52 707

53
54 708

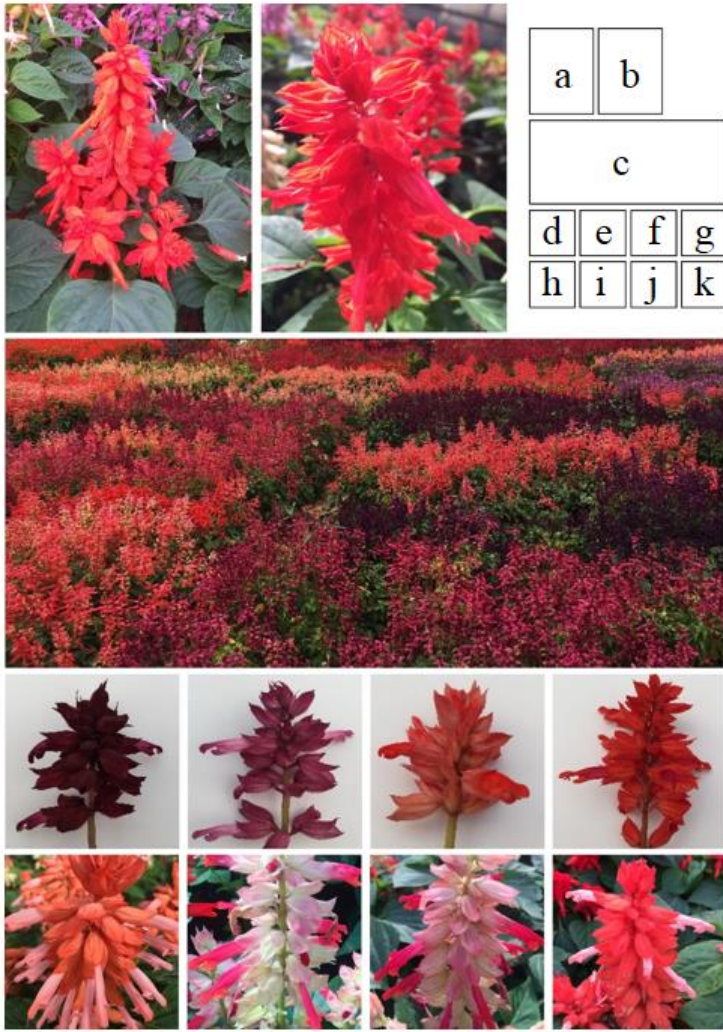
55 709

56 710

57 711

58
59
60
61
62
63
64
65

712 **Fig. 1**



713

714

715

716

717

718

719

720

721

722

723

724

725

726

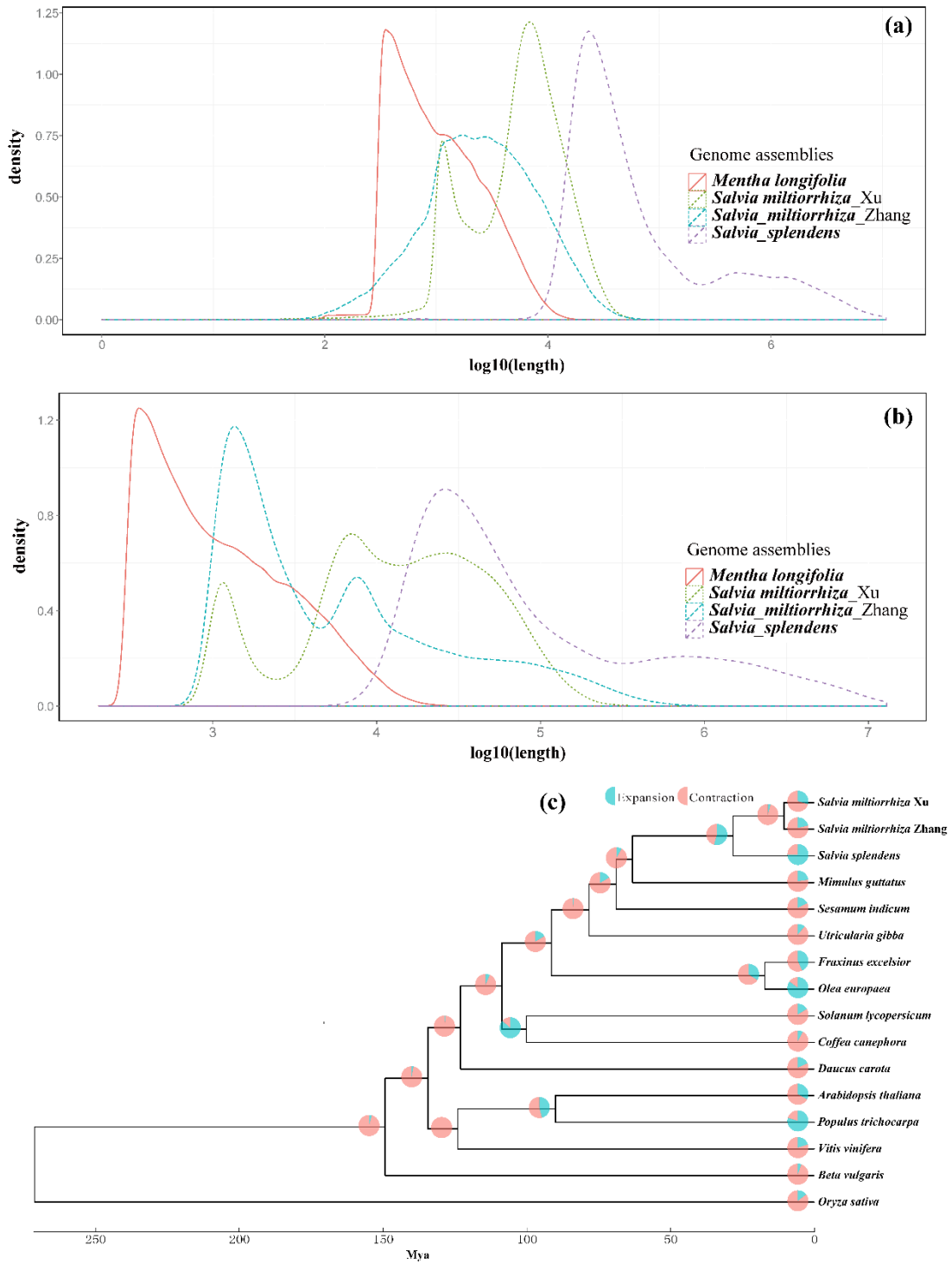
727

728

729

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

730 **Fig. 2**




731


732

733


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



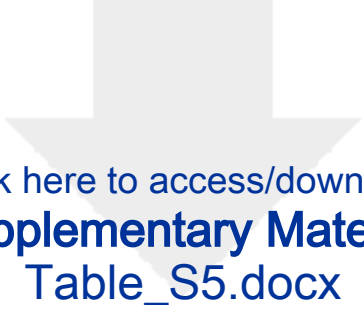
Click here to access/download
Supplementary Material
Table_S2.docx



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

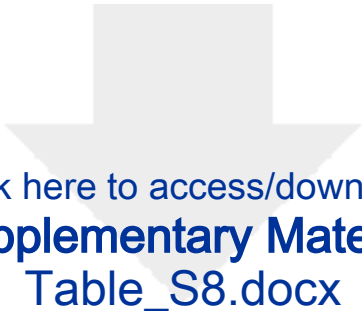





Click here to access/download
Supplementary Material
Table_S6.docx



Click here to access/download
Supplementary Material
Table_S7.docx




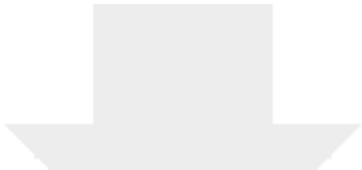
Click here to access/download
Supplementary Material
Table_S8.docx






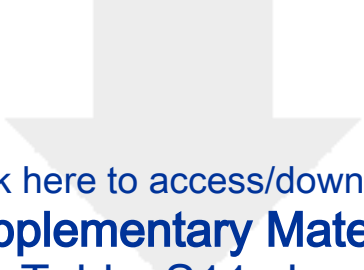
Click here to access/download
Supplementary Material
Table_S9.docx



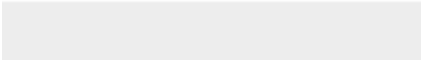



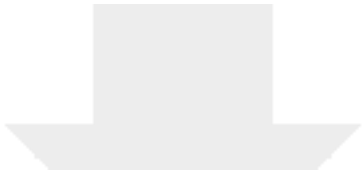
Click here to access/download
Supplementary Material
Table_S10.docx






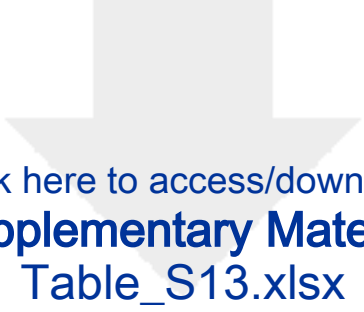
Click here to access/download
Supplementary Material
Table_S11.xlsx






Click here to access/download
Supplementary Material
Table_S12.docx





Click here to access/download
Supplementary Material
Table_S13.xlsx





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






Click here to access/download
Supplementary Material
Fig_S5.pdf



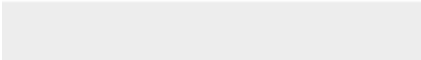



Click here to access/download
Supplementary Material
Fig_S6.pdf





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_3.docx

