A draft reference genome sequence for *Scutellaria baicalensis* Georgi

Qing Zhao[1,2,6], Jun Yang[1,2,6], Jie Liu[1], Meng-Ying Cui[1], Yuming Fang[1], Wengqing Qiu[3], Huiwen Shang[4], Zhicheng Xu[4], Yukun Wei[1], Lei Yang[1], Yonghong Hu[1], Xiao-Ya Chen[1,2] and Cathie Martin[1,5]*

[1]Shanghai Key Laboratory of Plant Functional Genomics and Resources, Shanghai Chenshan Botanical Garden, Shanghai Chenshan Plant Science Research Center, Chinese Academy of Sciences, Shanghai, China

[2]State Key Laboratory of Plant Molecular Genetics, CAS Center for Excellence in Molecular Plant Sciences, Shanghai Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai, China

[3]Key Laboratory of Metabolism and Molecular Medicine, Department of Biochemistry and Molecular Biology, School of Basic Medical Sciences, Fudan University, Shanghai, China

[4]Novogene Bioinformatics Institute, Beijing, China

[5]Department of Metabolic Biology, John Innes Centre, Norwich, NR4 7UH UK

[6]These authors contributed equally to this work

*Author for correspondence is Cathie Martin: email, Cathie.Martin@jic.ac.uk.

Abstract

Scutellaria baicalensis Georgi is an important medicinal plant used worldwide. Information about the genome of this species is important for scientists studying the metabolic pathways that synthesise the bioactive compounds in this plant. Here, we report a draft reference genome sequence for *S. baicalensis* obtained by a combination of Illumina and PacBio sequencing, which was assembled using 10 X Genomics and Hi-C technologies. We assembled 386.63 Mb of the 408.14 Mb genome, amounting to about 94.73% of the total genome size, and the sequences were anchored onto 9 pseudochromosomes with a super-N50 of 33.2 Mb. The reference genome sequence of *S. baicalensis* offers an important foundation for understanding the biosynthetic pathways for bioactive compounds in this medicinal plant and for its improvement through molecular breeding.

Introduction

Scutellaria baicalensis Georgi or Chinese Skullcap is a well-known medicinal plant that originated in East Asia and is cultivated in many European countries for its therapeutic properties (Shang et al., 2010). The dried root of *S. baicalensis* has been used as a traditional medicine for more than 2000 years in China, where it has the name Huang-Qin (黄芩) (Zhao et al., 2016). Huang-Qin is used in traditional Chinese medicine (TCM) for treatment of bitter, cold, liver and lung problems as recorded by *Shennong Bencaojing* (*The Divine Farmer's Materia Medica*), an ancient book on medicine and agriculture, written between about 200 and 250 AD (Ma, 2013). Recent scientific studies have reported the pharmacological activities of preparations of Scutellaria, particularly those of flavonoids that accumulate at high levels in the roots and rhizomes of *S. baicalensis* (Li-Weber, 2009; Shang et al., 2010; Tu et al., 2016). More than 132 compounds have been found in Huang-Qin, including flavones, flavanones, phenylethanoids and their glycosides (Qiao et al., 2016). The flavonoids in roots of Scutellaria have been reported to have various beneficial bioactivities including antibacterial, antiviral, antioxidant, anti-cancer, hepatoprotective and neuroprotective properties (Gao et al., 2011; Yang et al., 2012). Preparations of Scutellaria roots or their extracts are used as ingredients in many drugs for cold and liver problems and as adjunct treatments for lung and liver cancers in both Asian and Western countries (Wen, 2007). Despite the commercial importance and increasing demand for *Scutellaria*, improvements through breeding have been almost non-existent. The absence of genome information for this historically important medicinal plant has limited the understanding of how its flavonoid bioactives are made and prevented the improvement of productivity through genetic selection. Understanding the genes responsible for biosynthesis of the various flavonoids made in *S. baicalensis* and their regulation, will lay a foundation for biosynthesis and molecular breeding for improved productivity

Results

The DNA for genome sequencing of *Scutellaria baicalensis* Georgi came from a single plant maintained in Shanghai Chenshan Botanical Garden. DNA was extracted and sequenced by Illumina and PacBio sequencing strategy. We obtained 48.02 Gb PacBio reads, amounting to 117.66 X coverage of the 408.14 Mb, a genome size estimated by k-mer distribution analysis (Supplementary Figure 1a and supplementary table 1). We also measured the genome size to be 392 Mb by flow cytometry, which is close to the value given by k-mer method (Supplementary Figure 1b). After interactive error-correction among the PacBio reads, assembly was carried out using FALCON to obtain primary contigs. To avoid problems of heterozygosity from outcrossing, the contigs were phased using FALCON-Unzip, then the updated primary contigs and haplotigs were polished with Quiver. The final contigs were error-corrected with the 67.96 Gb (166.51X) of short reads obtained from Illumina HiSeq X Ten sequencing (Supplementary table 1). The consensus sequences were further assembled using the reference of 86.72 G (212.485 X coverage) from 10 X genomics sequencing (Supplementary table 1). All the contigs were extended using FragScaff to generate an assembly with a total scaffold length of 386.6 Mb (94.73% of the genome) and N50 of 1.33 Mb (Supplementary table 2). To facilitate genome annotation and to obtain expression profiles of the *S. baicalensis* genes, we performed transcriptome sequencing from RNA samples extracted from flowers, flower buds, leaves, stems, roots, and JA-treated roots. RNA samples from each tissue were sequenced in triplicate using the HiSeq 2000 platform.

A Hi-C (in vivo fixation of chromosomes) library was then employed to refine the first version of the reference genome. This method sorted 475 of the 578 scaffolds into 114 super-scaffolds, accounting for 98.04% of the original 386.63 Mb assembly (Table 1 and Supplementary table 3). All the super-scaffolds could be located on 9 groups (Supplementary figure 2). The super-scaffold N50 reached 33.2 Mb, with the longest one being 87.96 Mb (Table 1 and Supplementary table 4). The groups are hereafter referred to as pseudochromosomes, and this number corresponds well to the number of chromosomes reported in previous studies (1n=9, 2n=18) (Cheng et al., 2010). The genome of *S. baicalensis* has a normal A/T/C/G content, and its GC content is 34.24%, with N comprising 0.6% (Supplementary table 5). SNP calling based on the genome sequence shows a heterozygosity rate of 0.31% (Supplementary table 6).

To test the coverage of the genome, we mapped the short reads generated from Illumina sequencing, and 96.5% of these reads could be mapped to the scaffolds with a 99.78% overall coverage (Supplementary table 7). EST sequences generated from transcriptome sequencing were also mapped to the assembly and we found that 86.12% ESTs had more than 90% of their sequences in one scaffold, 97.51% ESTs had more than 50% of their sequences in one scaffold (Supplementary table 8). CEGMA and BUSCO evaluations of the genome sequence indicated 96.37% and 94% completeness, respectively (Supplemtal table 9 and Supplental table 10). All data suggest a good quality assembly.

A pipeline combined with *de novo*, homolog and RNA-seq was used to construct gene models for the *S. baicalensis* genome. A total of 28,930 genes were generated this way, with an average length

of 2,980 bp and an average CDS length of 1,122 bp (Table 1 and Supplementary table 11), among which, 23,027 genes (79.6%) were supported by RNA-sequencing data. A total of 20,234 genes (69.9%) were supported by all three methods (RNA-seq, de novo and homolog), therefore these genes have been annotated with high-confidence. The resulting protein models were then compared to protein sequences in four protein databases, namely NR, Swiss-Prot, KEGG and InterPro. We found that 28524 (98.6%) gene products could be annotated by at least one of the databases. All the gene loci were named according to the nomenclature used for Arabidopsis, which indicated clearly the relative positions of genes on the pseudochromosomes.

Table 1. Overview of *S. baicalensis* draft reference genome assembly.

Assembly feature	Statistic
Estimated genome size (k-mer analysis)	408.14 Mb
Assembly length	386.63 Mb
Length loaded on pseudochromosomes	379.07 Mb
Number of pseudochromosomes	9
Number of super-scaffolds	114
N50 of super-scaffolds	33.2 Mb
Longest super-scaffolds	87.96 Mb
Assembly % of genome	94.73
Repeat region of % assembly	53.95
Predicted gene models	28 930
Average coding sequence length	1 122.48 bp
Average exons per gene	5.07

The assembled *S. baicalensis* reference genome contains 55.15% repetitive sequences. Tandem duplications (small satellites and micro-satellites) and interspersed repeats account for 1.2% and 53.95% of the genome, respectively. Long terminal repeats of retroelements (LTR) are the most abundant interspersed repeat, occupying 34.4% of the genome, followed by DNA transposable elements at 15.4% (Supplementary table 12). Genes encoding non-coding RNAs; 1218 miRNAs, 517 tRNAs, 1846 rRNAs and 512 snRNAs were annotated in the genome (Supplementary table 13). An overview of the genes, repeats, non-coding RNA densities and all detected gene duplications are shown in Fig. 1.

Discussion

Members of the family Lamiaceae are commonly known as mint plants. The family has a worldwide distribution, and includes about 6,900 to 7,200 species with the largest genera being Salvia (900 species), followed by Scutellaria (360 species) and Stachys (300 species) (Raymond M. Harley, 2004). The extremely high degree diversity in specialised metabolites makes this family famous as medicinal plants and herbs, including mint, rosemary, basil, savory, perilla, salvia and skullcaps (Scutellaria) and so on. Terpenoids, phenolic acids and flavonoids are mainly responsible for these bioactivities (Lei et al., 2016; Wu et al., 2016; Zhao et al., 2016). Genome sequencing provides a key resource to study the molecular basis of diversity of the specialised metabolites in members of the family Lamiaceae and how their biosynthetic pathways evolved. To date, only the genomes of *S. miltiorrhiza* and *S. splendens* from the family Lamiaceae have been reported (Xu et al., 2016; Dong et al., 2018). We report here a high quality draft reference genome sequence of *S. baicalensis* of 386.63 Mb (about 94%). The genomic information reported offers a foundation for comparative genomic analysis between members of the family Lamiaceae. Furthermore, the genome sequence makes gene isolation and characterization, elucidation of metabolic pathways, as well as molecular breeding, easier for this important medicinal plant

Methods:

Genome Sequencing

Genomic DNA was extracted from leaves of a single *Scutellaria baicalensis* plant maintained in Shanghai Chenshan Botanical Garden, using a modified CTAB method (Tel-Zur et al., 1999). Quality control was done using a Sage Science Pippin pulse electrophoresis system. Genomic DNA with a length of around 150 kb was sheared using a Megaruptor DNA system and the resulting fragments of 30-50 kb were collected for the following steps. A SMRTbell DNA library was constructed with Sequel 2.0 reagent according to the manufacturer's instructions (Pacific Bioscience, www.pacb.com). SMRTbell sequencing primers together with P4 DNA polymerase were used for sequencing reactions on SMART Cells. The genomic DNA was sequenced on the PacBio Sequel system. This work produced 48.02 GB of single molecule data, about 117.66 X of the genome (Supplemental Table 1).

For short-insert library construction, DNA was extracted from the same plant using a DNA secure Plant Kit (TIANGEN, http://www.tiangen.com/) according to the manufacturer's instructions. The DNA was sheared and fragments with sizes of 200-300 bp were retrieved from agarose gels. The fragments were ligated to adaptors and were selected for RNA amplification for templates. The library was then sequenced on HiSeq X Ten. The Illumina sequencing produced 67.96 GB short reads data, amounting to about 166.51 X of the genome.

For 10 X genomic sequencing, we extracted DNA samples using the modified CTAB method (Tel-Zur et al., 1999). The GemCode Instrument (10 X Genomics) was employed for DNA indexing and barcoding. GEM reactions were carried out based on about 1 ng, 50 kb single DNA molecules and 16-bp barcodes were ligated to the molecules in droplets. The intermediate DNA was extracted from the droplets and sheared into 500 bp fragments (Zheng et al., 2016). The fragments were ligated to P7 adaptors, which were then sequenced on Illumina HiSeq X Ten platform (Mostovoy et al., 2016). We obtained 86.72 GB data from 10 X genomics sequencing, which was then used for genome assembly.

DNA from young leaves, collected from *Scutellaria baicalensis*, was used as starting material for the Hi-C library. Formaldehyde was used for fixing chromatin. The leaf cells were lysed and DpnII endonuclease was used for digesting the fixed chromatin. The 5' overhangs of the DNA were recovered with biotin-labelled nucleotides and the resulting blunt ends were ligated with each other using DNA ligase. Proteins were removed with protease to release the DNA molecules from the cross-links. The purified DNA was sheared into 350 bp fragments and ligated with adaptors (Yaffe and Tanay, 2011). The fragments labelled with biotin were extracted using streptavidin beads and, after PCR enrichment, the libraries were sequenced on Illumina Hiseq PE150.

Estimation of genome size

The genome size was measured by flow cytometry according to the protocol described by Dolezel (Doležel and Bartoš, 2005). Briefly, young leaves of *S. baicalensis* were chopped with a sharp razor blade for 60 seconds in nuclear isolation buffer (200 mM TRIS; 4 mM MgCl2.6H2O; 0.5 % (v/v) Triton

X-100; pH 7.5). Samples were incubated for 3 minutes, filtered through 50 µm CellTrics^{R} filters. Plant cell nuclei were stained by adding 2 ml buffer containing propidium iodide and RNAse A in the dark for 2 minutes. The relative Nuclear Genome Size was analysed on a flow cytometer (BD FACSAria III). This analysis gave us an estimate of genome size of 392 Mb compared with the tomato genome (Consortium, 2012). We further evaluated the genome size by k-mer frequency analysis based on Illumina short reads using the K-mer Analysis Toolkit (http://www.earlham.ac.uk/kat-tools).

Genome assembly

De novo assembly was carried out based on PacBio reads. Error correction was conducted by mapping the seed reads in FALCON according to the manufacturer's instructions (https://github.com/PacificBiosciences/FALCON/wiki/Manual) with the following parameters: --max_diff 100; --max_cov 100; --min_cov 2; --min_len 5000. Contaminations were removed with PacBio's whitelisting pipeline. The resulting primary assembly was phased using FALCON- Unzip (default parameters). Heterozygosity of the contigs was analysed with FALCON-Unzip, which was then phased according to the differences. The phased sequences were assembled into haplotigs for a diploid. The contigs were polished with Quiver (http://pbsmrtpipe.readthedocs.io/en/master/getting_started.html) with the parameters: pbsmrtpipe.options.chunk_mode: True pbsmrtpipe.options.max_nchunks: 50, to produce primary contigs, which were further corrected with reference to the Illumina reads with Pilon (https://github.com/broadinstitute/pilon/wiki).

RNA sequencing and analysis

Six tissues were harvested from 3-month-old *Scutellaria baicalensis* plants, namely, flower buds, flowers, leaves, roots, JA-treated roots (100 µM MeJA treated for 24 hours), and stems. Each tissue was collected for 3 biological replicates. Total RNA was extracted from these tissues using the RNAprep pure Plant Kit (Tiangen). After removing DNA, mRNA was isolated using oligodT beads. The mRNA was harvested and broken into short fragments, which were then used as templates for cDNA synthesis. After end repair, a single nucleotide A (adenine) was added and adapters were ligated to the cDNA. Fragments of 200-300 bp were separated for PCR amplification. An Agilent 2100 Bioanaylzer and ABI StepOnePlus Real-Time PCR System were used for quality control of the libraries, which were then sequenced on Illumina HiSeq™ 2000. Raw reads produced by sequencing were stored in Fastq format. Raw reads with adaptors and unknown nucleotides and low quality reads with more than 20% low quality base calling (base quality ≤10) were removed to leave clean reads. Trinity (Haas et al., 2013) was employed for *de novo* assembly based on the clean reads to produce Unigenes.

To annotate the Unigenes, blast (blastX and blastn) searches were carried out against NR, Swiss-Prot, KEGG, COG and NT databases (e-value<0.00001). Annotations of the proteins with the highest similarity to each Unigene were retrieved. Then GO (Gene Ontology) annotations based on the similarity were assigned to each Unigene using Blast2GO (Conesa et al., 2005). Expression of Unigenes

was calculated using the FPKM method (Fragments Per Kb per Million fragments) (Mortazavi et al., 2008). The formula is: $FPKM=10^6C/NL/10^3$, where C is the number of fragments that aligned specifically to a certain Unigene; N is the total number of fragments that aligned to all Unigenes; L is the base number in the CDS of the Unigene.

Evaluation of genome quality

To evaluate the coverage of the assembly, all the paired-end Illumina short reads were mapped to the assembly using BWA (http://bio-bwa.sourceforge.net/) (Li and Durbin, 2009). Gene completeness was evaluated using the ESTs generated from RNA sequencing. The ESTs were mapped to the assembly using BLAT (http://genome.ucsc.edu/goldenpath/help/blatSpec.html). For CEGMA (Core Eukaryotic Genes Mapping Approach, http://korflab.ucdavis.edu/dataseda/cegma/) evaluation, we build a set of highly reliable conserved protein families that occur in a range of model eukaryotes (Parra et al., 2007). Then we mapped the 248 core eukaryotic genes to the genome. The genome was also assessed using BUSCO (Benchmarking Universal Single-Copy Orthologs: http://busco.ezlab.org/) gene set analysis (Simão et al., 2015), which includes 956 single-copy orthologous genes.

Genome annotation

Repeat elements were annotated using a combined strategy. Alignment searches were undertaken against the RepBase database (http://www.girinst.org/repbase), then Repeatproteinmask searches (http://www.repeatmasker.org/) were used for prediction of homologs (Jurka et al., 2005). For *de novo* annotation of repeat elements, LTR_FINDER (http://tlife.fudan.edu.cn/ltr_finder/), Piler (http://www.drive5.com/piler/), RepeatScout (http://www.repeatmasker.org/) and RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html) were used to construct a *de novo* library, then annotation was carried out with Repeatmasker (Price et al., 2005). Non-coding RNA was annotated using tRNAscan-SE (http://lowelab.ucsc.edu/tRNAscan-SE/) (for tRNA) or INFERNAL (http://infernal.janelia.org/) (for miRNA and snRNA). Since rRNA sequences are highly conserved among plants, rRNA from *S. baicalensis* was screened by blast searches.

Gene structure screening was carried out through a combination of homology, de novo and EST based predictions. A gene set including protein coding sequences from *Arabidopsis thaliana*, *Oryza brachyantha*, *Populus trichocarpa*, *Salvia miltiorrhiza*, *Sesamum indicum*, *Solanum lycopersicum* and *Utricularia gibba* was mapped to the assembly of *S. baicalensis* using blast (http://blast.ncbi.nlm.nih.gov/Blast.cgi) (E value$\leq 1e^{-5}$) and gene structure of each hit was predicted through Genewise (http://www.ebi.ac.uk/~birney/wise2/) (Birney et al., 2004). *De novo* gene structure identification was performed using Augustus (http://bioinf.uni-greifswald.de/augustus/) and GlimmerHMM (http://ccb.jhu.edu/software/glimmerhmm/). Gene structure was also screened by mapping ESTs to the assembly through Blat (http://genome.ucsc.edu/cgi-bin/hgBlat). All the resulting genes were corrected by reference to the transcriptome data and integrated into a non- redundant gene

set using EVidenceModeler (EVM, http://evidencemodeler.sourceforge.net/) (Haas et al., 2008). The genes were further corrected with PASA (Program to Assemble Spliced Alignment, http://pasa.sourceforge.net/) to predict UTRs and alternative splicing (Haas et al., 2008), which generated 28,930 gene models.

Gene function was annotated by performing BLASTP (E-value ≤ $1e^{-5}$) against the protein databases, SwissProt (http://www.uniprot.org/), TrEMBL (http://www.uniprot.org/) and KEGG (http://www.genome.jp/kegg/). InterPro (https://www.ebi.ac.uk/interpro/) and Pfam (http://pfam.xfam.org/) were used for screening the functional domains of the proteins. Then GO (Gene Ontology) terms based on the searches were assigned to each gene.

Data accessibility

Reference genome data are deposited in GenBank under project number PRJNA484052 and transcriptome sequence reads are deposited in the Sequence Read Archive (SRA) under accession number SRA accession SRP156996. The authors of the present manuscript reserve their priority for using these data for genome level analyses. By using these data, you agree not to use them for the publication of such genome level analyses without the written consent of the authors. If you agree to these terms, you can obtain the credentials for accessing the data by writing and/or email to Qing Zhao (Qing.Zhao@jic.ac.uk) at Shanghai Chenshan Botanical Garden.

Acknowledgements

This work was supported by National Natural Science Foundation of China (31870282, 31700268 and 31788103), Fund of Chinese Academy of Sciences (QYZDY-SSW-SMC026 and 153D31KYSB20160074), Chenshan Special Fund for Shanghai Landscaping Administration Bureau Program (G172402 and G162409) and by the CAS/JIC and Centre of Excellence for Plant and Microbial Sciences (CEPAMS) joint foundation for support to QZ, XYC and CM.

Authour Contributions

QZ and CM initiated the programme, coordinated the project and wrote the manuscript. LJ, M-Y C, Y-M F, Y-K W and LY maintained and prepared the samples. QZ, W-Q Q, Z-C X and H-W S designed the sequencing strategy and performed sequencing. JY, QZ, X-Y C, Y-H H and CM performed analysis.

Competing financial interests

The authors declare no competing financial interests.

Figure 1

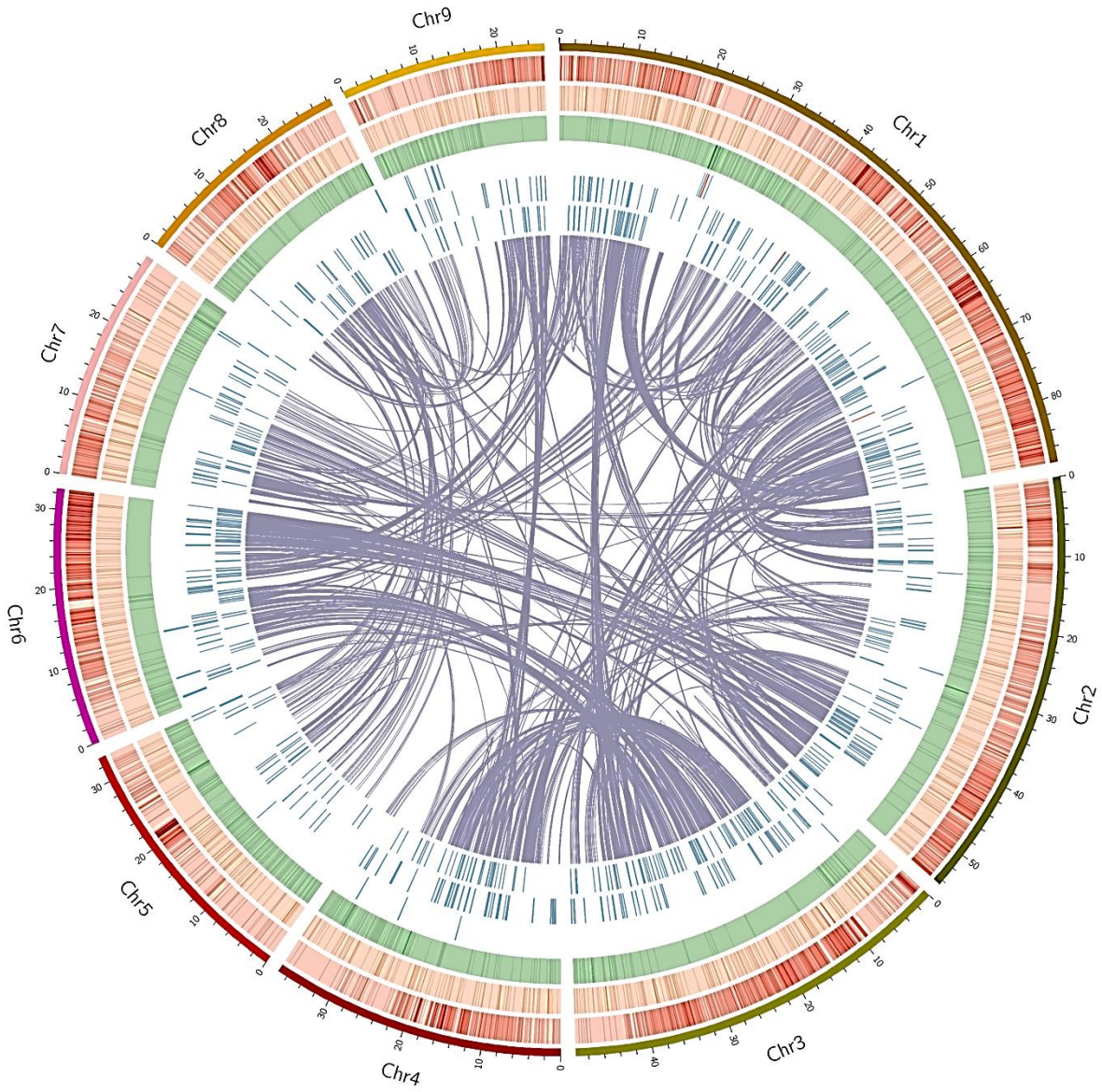

Fig. 1. Summary of current *S. baicalensis* draft genome assembly. The outer lines represent pseudochromosomes. The coloured bands summarise the density of genes (red), genes encoding miRNAs (orange), repeats (green), genes encoding rRNAs (blue), genes encoding snRNAs (blue) and genes encoding tRNAs (blue). All detected gene duplications are indicated with links inside the circles.

References

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. Genome Research **14,** 988.

Cheng, Z., Cai, M., Hao, D., Deng, R., and Yan, L. (2010). Karyotype Analysis and Meiotic Observations of Pollen Mother Cells in Scutellaria baicalensis Georgi. Chinese Wild Plant Resources.

Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics **21,** 3674-3676.

Consortium, T.G. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485,** 635.

Doležel, J., and Bartoš, J. (2005). Plant DNA Flow Cytometry and Estimation of Nuclear Genome Size. Annals of Botany **95,** 99-110.

Dong, A.X., Xin, H.B., Li, Z.J., Liu, H., Sun, Y.Q., Nie, S., Zhao, Z.N., Cui, R.F., Zhang, R.G., and Yun, Q.Z. (2018). High-quality assembly of the reference genome for scarlet sage,Salvia splendens, an economically important ornamental plant. Gigascience **7**.

Gao, J., Morgan, W.A., Sanchez-Medina, A., and Corcoran, O. (2011). The ethanol extract of Scutellaria baicalensis and the active compounds induce cell cycle arrest and apoptosis including upregulation of p53 and Bax in human lung cancer cells. Toxicology and applied pharmacology **254,** 221-228.

Haas, B.J., Salzberg, S.L., Wei, Z., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biology **9,** R7.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., and Lieber, M. (2013). De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nature Protocols **8,** 1494-1512.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic & Genome Research **110,** 462-467.

Lei, Y., Yang, C., Li, C., Zhao, Q., Ling, L., Xin, F., and Chen, X.Y. (2016). Recent advances in biosynthesis of bioactive compounds in traditional Chinese medicinal plants. Science bulletin **61,** 3.

Li-Weber, M. (2009). New therapeutic aspects of flavones: The anticancer properties of Scutellaria and its main active constituents Wogonin, Baicalein and Baicalin. Cancer treatment reviews **35,** 57-68.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. (Oxford University Press).

Ma, J.X. (2013). Explanatory notes to Shennong Bencaojing **3,** 140.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods **5,** 621-628.

Mostovoy, Y., Levysakin, M., Lam, J., Lam, E.T., Hastie, A.R., Marks, P., Lee, J., Chu, C., Lin, C., and Džakula, Ž. (2016). A Hybrid Approach for de novo Human Genome Sequence Assembly and Phasing. Nature Methods **13,** 587-590.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics **23,** 1061-1067.

Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat families in large genomes. Bioinformatics **21 Suppl 1,** i351.

Qiao, X., Li, R., Song, W., Miao, W.J., Liu, J., Chen, H.B., Guo, D.A., and Ye, M. (2016). A targeted strategy to analyze untargeted mass spectral data: Rapid chemical profiling of Scutellaria baicalensis using ultra-high performance liquid chromatography coupled with hybrid quadrupole orbitrap mass spectrometry and key ion filtering. J Chromatogr A **1441,** 83-95.

Raymond M. Harley, S.A., Andrey L. Budantsev, Philip D. Cantino, Barry J. Conn, Renée J. Grayer, Madeline M. Harley, Rogier P.J. de Kok, Tatyana V. Krestovskaja, Ramón Morales, Alan J. Paton, and P. Olof Ryding. (2004). The Families and Genera of Vascular Plants volume VII. (Berlin; Heidelberg, Germany: Springer-Verlag).

Shang, X.F., He, X.R., He, X.Y., Li, M.X., Zhang, R.X., Fan, P.C., Zhang, Q.L., and Jia, Z.P. (2010). The genus Scutellaria an ethnopharmacological and phytochemical review. J Ethnopharmacol **128,** 279-313.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31,** 3210-3212.

Tel-Zur, N., Abbo, S., Myslabodski, D., and Mizrahi, Y. (1999). Modified CTAB Procedure for DNA Isolation from Epiphytic Cacti of the. Plant Molecular Biology Reporter **17,** 249-254.

Tu, B., Li, R.R., Liu, Z.J., Chen, Z.F., Ouyang, Y., and Hu, Y.J. (2016). Structure-activity relationship study between baicalein and wogonin by spectrometry, molecular docking and microcalorimetry. Food Chemistry **208,** 192-198.

Wen, J. (2007). Sho-saiko-to, A Clinically Documented Herbal Preperation for Treating Chronic Liver Disease. HerbalGram;Feb-Apr2007, Issue 73, p5 **59,** 34-43.

Wu, Y.B., Ni, Z.Y., Shi, Q.W., Dong, M., Kiyota, H., Gu, Y.C., and Cong, B. (2016). Constituents from Salvia species and their biological activities. Chemical Reviews **112,** 5967-6026.

Xu, H., Song, J., Luo, H., Zhang, Y., Li, Q., Zhu, Y., Xu, J., Li, Y., Song, C., Wang, B., Sun, W., Shen, G., Zhang, X., Qian, J., Ji, A., Xu, Z., Luo, X., He, L., Li, C., Sun, C., Yan, H., Cui, G., Li, X., Li, X., Wei, J., Liu, J., Wang, Y., Hayward, A., Nelson, D., Ning, Z., Peters, R.J., Qi, X., and Chen, S. (2016). Analysis of the Genome Sequence of the Medicinal Plant Salvia miltiorrhiza. Mol Plant **9,** 949-952.

Yaffe, E., and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nature Genetics **43,** 1059.

Yang, M.D., Chiang, Y.M., Higashiyama, R., Asahina, K., Mann, D.A., Mann, J., Wang, C.C.C., and Tsukamoto, H. (2012). Rosmarinic acid and baicalin epigenetically derepress peroxisomal proliferator-activated receptor gamma in hepatic stellate cells for their antifibrotic effect. Hepatology **55,** 1271-1281.

Zhao, Q., Chen, X.Y., and Martin, C. (2016). Scutellaria baicalensis, the golden herb from the garden of Chinese medicinal plants. Science bulletin **61,** 1391-1398.

Zheng, G.X., Lau, B.T., Schnall-Levin, M., Jarosz, M., Bell, J.M., Hindson, C.M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D.A., Merrill, L., and Terry, J.M. (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nature biotechnology **34,** 303-311.