

# GigaScience

## Chromosome-scale genome assembly of kiwifruit *Actinidia eriantha* with single-molecule sequencing and chromatin conformation capture

--Manuscript Draft--

<b>Manuscript Number:</b>	GIGA-D-18-00282	
<b>Full Title:</b>	Chromosome-scale genome assembly of kiwifruit <i>Actinidia eriantha</i> with single-molecule sequencing and chromatin conformation capture	
<b>Article Type:</b>	Data Note	
<b>Funding Information:</b>	National Natural Science Foundation of China (31471157)	Dr. Yongsheng Liu
	National Science Foundation (IOS-1339287)	Dr. Zhangjun Fei
<b>Abstract:</b>	<p>Background: Kiwifruit (<i>Actinidia</i> spp.) is a dioecious plant with fruits containing abundant vitamin C and minerals. A handful of kiwifruit species have been domesticated, among which the <i>A. eriantha</i> is increasingly favored in breeding due to its superior commercial traits. Recently, elite cultivars from <i>A. eriantha</i> have been successfully selected and further studies on their biology and breeding potential require genomic information which is currently unavailable.</p> <p>Findings: Here, we assembled a chromosome-scale genome sequence of <i>A. eriantha</i> cv. White using single-molecular sequencing and chromatin conformation capture. The assembly has a total size of 690.6 Mb and an N50 of 21.7 Mb. Approximately 99% of the assembly were in 29 pseudomolecules corresponding to the 29 kiwifruit chromosomes. Forty-three percent of the <i>A. eriantha</i> genome are repetitive sequences, and the non-repetitive part encodes 42,850 protein-coding genes, of which 39,075 have homologues from other plant species or contain protein domains. The divergence time between <i>A. eriantha</i> and its close relative <i>A. chinensis</i> is estimated to be 3.3 million years, and after diversification, 1,740 and 1,345 gene families are expanded or contracted in <i>A. eriantha</i>, respectively.</p> <p>Conclusions: We generate a high-quality reference genome of kiwifruit <i>A. eriantha</i>. This chromosome-scale genome assembly is substantially better than two published kiwifruit assemblies from <i>A. chinensis</i> in terms of genome contiguity and completeness. The availability of <i>A. eriantha</i> genome provides a valuable resource for facilitating kiwifruit breeding and the studies of kiwifruit biology.</p>	
<b>Corresponding Author:</b>	Zhangjun Fei Boyce Thompson Institute for Plant Research Ithaca, NY UNITED STATES	
<b>Corresponding Author Secondary Information:</b>		
<b>Corresponding Author's Institution:</b>	Boyce Thompson Institute for Plant Research	
<b>Corresponding Author's Secondary Institution:</b>		
<b>First Author:</b>	Zhangjun Fei	
<b>First Author Secondary Information:</b>		
<b>Order of Authors:</b>	Zhangjun Fei	
	Wei Tang	
	Xuepeng Sun	
	Junyang Yue	
	Xiaofeng Tang	
	Chen Jiao	
	Ying Yang	

	Xiangli Niu
	Min Miao
	Danfeng Zhang
	Shengxiong Huang
	Wei Shi
	Mingzhang Li
	Congbing Fang
	Yongsheng Liu
<b>Order of Authors Secondary Information:</b>	
<b>Additional Information:</b>	
<b>Question</b>	<b>Response</b>
Are you submitting this manuscript to a special series or article collection?	No
<p><b>Experimental design and statistics</b></p> <p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>. Information essential to interpreting the data presented should be made available in the figure legends.</p> <p>Have you included all the information requested in your manuscript?</p>	Yes
<p><b>Resources</b></p> <p>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 <a href="#">Research Resource Identifiers</a> (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	Yes
<b>Availability of data and materials</b>	Yes

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

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

1  
2  
3  
4 **Chromosome-scale genome assembly of kiwifruit *Actinidia eriantha* with single-molecule**  
5  
6 **sequencing and chromatin conformation capture**  
7  
8  
9

10  
11 Wei Tang<sup>1,2,3\*</sup>, Xuepeng Sun<sup>4\*</sup>, Junyang Yue<sup>1,3\*</sup>, Xiaofeng Tang<sup>1,3</sup>, Chen Jiao<sup>4</sup>, Ying Yang<sup>1</sup>,  
12  
13 Xiangli Niu<sup>1,3</sup>, Min Miao<sup>1,3</sup>, Danfeng Zhang<sup>3</sup>, Shengxiong Huang<sup>3</sup>, Wei Shi<sup>3</sup>, Mingzhang Li<sup>5</sup>,  
14  
15  
16 Congbing Fang<sup>1</sup>, Zhangjun Fei<sup>4,6\*</sup>, Yongsheng Liu<sup>1,2,3\*</sup>  
17  
18  
19  
20

21 <sup>1</sup>School of Horticulture, Anhui Agricultural University, Hefei 230036, China  
22

23 <sup>2</sup>Ministry of Education Key Laboratory for Bio-resource and Eco-environment, College of Life  
24  
25 Science, State Key Laboratory of Hydraulics and Mountain River Engineering, Sichuan University,  
26  
27  
28 Chengdu 610064, China  
29  
30

31 <sup>3</sup>School of Food Science and Engineering, Hefei University of Technology, Hefei 230009, China  
32

33 <sup>4</sup>Boyce Thompson Institute, Cornell University, Ithaca NY 14853, USA  
34  
35

36 <sup>5</sup>Sichuan Academy of Natural Resource Sciences, Chengdu 610015, China  
37

38 <sup>6</sup>U.S. Department of Agriculture-Agricultural Research Service, Robert W. Holley Center for  
39  
40  
41 Agriculture and Health, Ithaca, NY 14853, USA  
42  
43  
44

45 \*W.T., X.S. and J.Y. contributed equally to this work  
46  
47

48 \*Correspondence authors: Dr. Zhangjun Fei, email: [zf25@cornell.edu](mailto:zf25@cornell.edu) or Dr. Yongsheng Liu,  
49  
50  
51 email: [liuyongsheng1122@hfut.edu.cn](mailto:liuyongsheng1122@hfut.edu.cn)  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 **Abstract**  
5

6 **Background:** Kiwifruit (*Actinidia* spp.) is a dioecious plant with fruits containing abundant  
7 vitamin C and minerals. A handful of kiwifruit species have been domesticated, among which the  
8 *A. eriantha* is increasingly favored in breeding due to its superior commercial traits. Recently, elite  
9 cultivars from *A. eriantha* have been successfully selected and further studies on their biology and  
10 breeding potential require genomic information which is currently unavailable.  
11  
12  
13  
14  
15  
16  
17

18 **Findings:** Here, we assembled a chromosome-scale genome sequence of *A. eriantha* cv. White  
19 using single-molecular sequencing and chromatin conformation capture. The assembly has a total  
20 size of 690.6 Mb and an N50 of 21.7 Mb. Approximately 99% of the assembly were in 29  
21 pseudomolecules corresponding to the 29 kiwifruit chromosomes. Forty-three percent of the *A.*  
22 *eriantha* genome are repetitive sequences, and the non-repetitive part encodes 42,850 protein-  
23 coding genes, of which 39,075 have homologues from other plant species or contain protein  
24 domains. The divergence time between *A. eriantha* and its close relative *A. chinensis* is estimated  
25 to be 3.3 million years, and after diversification, 1,740 and 1,345 gene families are expanded or  
26 contracted in *A. eriantha*, respectively.  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39

40 **Conclusions:** We generate a high-quality reference genome of kiwifruit *A. eriantha*. This  
41 chromosome-scale genome assembly is substantially better than two published kiwifruit  
42 assemblies from *A. chinensis* in terms of genome contiguity and completeness. The availability of  
43 *A. eriantha* genome provides a valuable resource for facilitating kiwifruit breeding and the studies  
44 of kiwifruit biology.  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54

55 Key words: Kiwifruit; *Actinidia eriantha*; Genome assembly; single molecular sequencing; Hi-C  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

## Data description

### *Introduction*

Kiwifruit is well known as the king of fruits due to its remarkably high vitamin C content and abundant minerals [1, 2]. Native to China, kiwifruit belongs to the genus *Actinidia* which contains 54 species and 75 taxa [3]. All species in this genus are perennial, deciduous and dioecious with a climbing or scrambling growth habit, and they also have many common morphological features including the characteristic radiating arrangement of styles of the female flower and the structure of the fruit [4]. Despite rich germplasm resources in kiwifruit, only a few *Actinidia* species have been domesticated, such as *A. chinensis* var. *chinensis*, *A. chinensis* var. *deliciosa* and *A. eriantha*, whose fruit size are close to commercial standard [5-7].

Owing to its strong resistance to *Pseudomonas syringae* pv. *Actinidiae*, long shelf-life, enriched ascorbic acid and peelable skin [7-11], *A. eriantha* (2n=58) has been favored in kiwifruit breeding. Recently, new cultivars have been selected either from the wild germplasm of *A. eriantha* such as ‘White’ (Fig. 1) or from the interspecific hybridization between *A. eriantha* (♂) and *A. chinensis* (♀) such as ‘Jinyan’ [7, 12]. The ‘White’ has particularly large fruits (96 g on average) with green flesh and favorable flavor and has been widely cultivated in China [7].

*Actinidia eriantha* has also been used for genetic and genomic studies thanks to its high efficiency in genetic transformation and relatively short phase of juvenility [13]. The flowering and fruiting of *A. eriantha* can be accomplished within two years in green house conditions with a low requirement for winter chilling [13]. In addition, roots of *A. eriantha* which contain many bioactive compounds such as triterpenes and polysaccharides are employed as a traditional Chinese medicine for the treatment of gastric carcinoma, nasopharyngeal carcinoma, breast carcinoma, and hepatitis [12, 14].

1  
2  
3  
4 Previously, two kiwifruit genomes were published and both are from *A. chinensis* ('Hongyang'  
5 and 'Red 5') [15, 16]. These short-read based assemblies are very fragmented, possibly due to the  
6 high complexity and heterozygosity of the kiwifruit genomes as well as technical limitations. Here,  
7 we used single-molecular sequencing combined with the high-throughput chromosome  
8 conformation capture (Hi-C) technology to assemble the genome of the elite kiwifruit cultivar  
9 'White' of *A. eriantha*. The availability of this high-quality chromosome-scale genome sequence  
10 not only provides fundamental knowledge regarding kiwifruit biology but also presents a valuable  
11 resource for kiwifruit breeding programs.  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22

### 23 24 25 26 ***Sample collection and genome sequencing*** 27

28 Fresh young leaves were collected from a female individual of *A. eriantha* cv. White. High  
29 molecular weight (HMW) genomic DNA was extracted using the CTAB method as described in  
30 the protocol ([https://www.pacb.com/wp-content/uploads/2015/09/Shared-Protocol-Preparing-  
31 Arabidopsis-DNA-for-20-kb-SMRTbell-Libraries.pdf](https://www.pacb.com/wp-content/uploads/2015/09/Shared-Protocol-Preparing-Arabidopsis-DNA-for-20-kb-SMRTbell-Libraries.pdf)). To construct genomic libraries  
32 (SMRTbell libraries) for PacBio long-read sequencing, HMW genomic DNA was sheared into  
33 fragments of approximately 20 kb using a Covaris g-Tube (KBiosciences p/n520079),  
34 enzymatically repaired and converted to SMRTbell template following the manufacturer's  
35 instruction (DNA Template Prep Kit 1.0, PacBio p/n 100-259-100). The templates were size-  
36 selected using a BluePippin (SageScience, Inc.) to enrich large DNA fragments (> 10 kb) and then  
37 sequenced on a PacBio Sequel system. A total of nine SMRT cells were sequenced, yielding  
38 3,889,480 reads with a mean and median length of 10,065 and 15,661 bp, respectively, and a total  
39 of 39.1 Gb sequences, about 51.6× coverage of the kiwifruit genome with an estimated size of  
40 758 Mb based on the flow cytometry analysis [17] (Table S1).  
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  
2  
3  
4 Three paired-end Illumina libraries with insert sizes of 180, 220 and 500 bp, and seven  
5 mate-pair libraries with insert sizes of 3, 4, 5, 8, 10, 15, 17 kb, were prepared using Illumina's  
6 Genomic DNA Sample Preparation kit and the Nextera Mate Pair Sample Preparation kit (Illumina,  
7 San Diego, CA), respectively. All libraries were sequenced on an Illumina HiSeq 2500 system,  
8 which yielded about 80.1 and 97.3 Gb of raw sequence data for paired-end and mate-pair libraries,  
9 respectively (Table S1). The raw Illumina paired-end reads were processed to remove adaptors  
10 and low-quality bases using Trimmomatic [18] (v0.35), and the mate-pair reads were cleaned using  
11 NextClip [19] (v1.3.1) with default parameters. Finally, we obtained 76.6 and 46.2 Gb high-quality  
12 cleaned sequences for paired-end and mate-pair libraries, respectively (Table S1).  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25

26 To construct the Hi-C library, 'White' plants were grown in a greenhouse, and  
27 approximately 4~6 grams young leaves were then harvested and subsequently fixed in the  
28 formaldehyde (1% v/v) for 10 min at room temperature. The fixation was terminated by adding  
29 glycine to a final concentration of 0.125M. The fixed samples were ground into powder in liquid  
30 nitrogen and then lysed with the addition of Triton X-100 to a concentration of 1% (v/v). The  
31 nuclei were isolated and prepared for Hi-C library construction according to a previously published  
32 protocol [20]. The library was sequenced on an Illumina HiSeq 2500 system using the paired-end  
33 mode, which yield a total of approximately 118 million read pairs.  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47

### 48 ***Transcriptome sequencing***

49  
50 To improve gene prediction, we generated transcriptome sequences from a pool of mixed tissues  
51 of 'White' including root, stem, leaf, flower, and fruits at 7, 30, 60, 90 and 120 days after anthesis.  
52  
53 Total RNA was extracted from these tissues using an RNA extraction kit (BIOFIT, China), treated  
54 with DNase I and further purified with an RNA clean kit (Promega, USA). RNA-Seq libraries  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65



1  
2  
3  
4 were constructed with the NEBNext® Ultra™ RNA Library Prep Kit (Illumina, USA), and  
5  
6 sequenced on an Illumina HiSeq 2500 system using the paired-end mode. A total of ~19.5 million  
7  
8 raw read pairs were obtained, which were processed with Trimmomatic to remove adaptors. The  
9  
10 cleaned reads were assembled *de novo* with Trinity [21] (version 2.4.0). Mapping of RNA-Seq  
11  
12 reads to the genome assembly was performed with STAR [22] (version 020201), and read counting  
13  
14 on the coding regions was performed with HTSeq [23] (version 0.6.0.).  
15  
16  
17  
18  
19  
20

### 21 ***Chromosome-scale assembly of the A. eriantha genome***

22  
23 We employed a strategy which took into account the unique advantage of different assemblers to  
24  
25 construct the ‘White’ genome using PacBio long reads. First, PacBio long reads were corrected  
26  
27 and assembled using the Canu program [24] (v1.7), which is a modularized pipeline consisting of  
28  
29 three primary stages - read correction, trimming and assembly. The Canu-corrected reads were  
30  
31 also assembled independently with the wtdbg program (<https://github.com/ruanjue/wtdbg>), a fast  
32  
33 assembler for long noisy reads. Subsequently, the two independent assemblies (one with Canu and  
34  
35 another with wtdbg) were merged by Quickmerge [25] (v0.2) to improve the contiguity. The  
36  
37 merged assembly was further processed to correct errors using Pilon [26] (version 1.22) with high-  
38  
39 quality cleaned Illumina reads from all paired-end and mate-pair libraries representing a total  
40  
41 genome coverage of 171× (Table S1). This yielded 2,818,370 nucleotides, 2,495,388 insertions  
42  
43 and 1,691,495 deletions being corrected. The resulting final assembled *A. eriantha* cv. ‘White’  
44  
45 genome contained 4,076 contigs with a N50 length of 539,246 bp and a cumulative size of  
46  
47 690,376,929 bp (Table 1). The contiguity and completeness of this assembly far exceeds that of  
48  
49 two published kiwifruit *A. chinensis* genomes (Table 1).  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

Table 1 Assembly statistics

	<i>A. eriantha</i>		<i>A. chinensis</i>	
	White	Hongyang	red5	
<b>Contigs</b>				
Total contig number (#)	4,076	26,721	39,868	
Total contig length (Mb)	690.4	604.2		
Contig N50 (kb)	539.2	58.9		
Contig N90 (kb)	50.7	11.6		
Longest contig length (kb)	3,260.20	423.5		
<b>Scaffolds</b>				
Total scaffold number (#)	1,735	7,698	3,887	
Total scaffold length (Mb)	690.6	616.1	550.5	
Scaffold N50 (kb)	23,583.9	646.8	623.8	
Scaffold N90 (kb)	20,112.1	122.7	140.7	
Longest scaffold length (Mb)	28.6	3.4	4.43	
Anchored to chromosome (Mb/%)	682.4 / 98.84	452.4 / 73.4	547.9 / 98.9	
Anchored with order and orientation (Mb/%)	634.4 / 91.90	333.6 / 54.1		

To scaffold the contigs based on chromatin interaction maps inferred from the Hi-C data, we first used HiC-Pro [27] to evaluate and filter the cleaned Hi-C reads. The Hi-C data usually contains a considerable part of invalid interaction read pairs which are non-informative and need to be filtered out beforehand. Among the 51 million read pairs that were uniquely aligned to the *A. eriantha* assembly, 33 million (64.1%) were valid interaction pairs and their insertion size spanned predominantly from dozens to hundreds of kilobases, therefore providing efficient information for scaffolding. As a part of error correction of the assembly, we also used valid Hi-C reads to identify potential misassembled contigs. In principle, a genuine contig should display a continuous Hi-C interaction map whereas the discrete distribution of an interaction map likely indicates a misassembly. We examined the interaction map for each contig and broke 51 that were possibly misassembled. Subsequently, the corrected PacBio assembly was used for scaffolding using the LACHESIS program [28] with parameters “CLUSTER\_MIN\_RE\_SITES=48, CLUSTER\_MAX\_LINK\_DENSITY=2, CLUSTER\_NONINFORMATIVE\_RATIO=2, ORDER\_MIN\_N\_RES\_IN\_TRUN=14, ORDER\_MIN\_N\_RES\_IN\_SHREDS=15”. LACHESIS

1  
2  
3  
4 assigned 3,666 contigs with a total size of 682,355,494 bp (98.84% of the assembly) into 29 groups  
5  
6 corresponding to the 29 kiwifruit chromosomes (Fig. 2 and 3a), among which 634,430,648 bp  
7  
8 (91.90%) had defined order and orientation (Table 1 and S2). The final chromosome-scale  
9  
10 assembly had a total length of 690,781,529 bp and an N50 of 23,583,865 bp.  
11  
12  
13  
14  
15

### 16 ***Evaluation of the genome assembly***

17  
18 We first evaluated the quality of the assembled *A. eriantha* ‘White’ genome by mapping Illumina  
19  
20 genomic and RNA-Seq reads to the assembly. Reads from the three paired-end genomic libraries  
21  
22 had very high mapping rates, ranging from 98.6% to 98.8%, and the properly paired read mapping  
23  
24 rates were between 76.9% and 90.4%. For the RNA-Seq reads, 91.7% could be mapped to the  
25  
26 genome and 87.1% were uniquely mapped. The high mapping ratio of both genomic and RNA-  
27  
28 Seq reads suggest a high quality of the *A. eriantha* ‘White’ assembly.  
29  
30  
31  
32

33 We then identified synteny between the *A. eriantha* ‘White’ assembly and the assembly of  
34  
35 *A. chinensis* ‘red5’ using MUMMER [29] (version 4.0.0beta2). In general, the two assemblies  
36  
37 showed a high macro-collinearity, with only a few inconsistencies (Fig. 3b). Detailed check of the  
38  
39 inconsistent regions using mate-pair read alignments supported the correct assemblies in the *A.*  
40  
41 *eriantha* ‘White’ genome, and therefore the inconsistencies could be due to errors in the ‘red5’  
42  
43 assembly or structure variations between ‘White’ and ‘red5’ (Fig. S1).  
44  
45  
46  
47  
48  
49

### 50 ***Repeat annotation***

51  
52 Repeats were annotated following a protocol described in Campbell et al [30]. The customized  
53  
54 repeat library was built to include both known and novel repeat families. We first searched the  
55  
56 assembly for miniature inverted transposable elements (MITEs) using MITE-Hunter [31] with  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 default parameters. The long terminal repeat (LTR) retrotransposons were then identified from the  
5  
6 *A. eriantha* ‘White’ genome using LTRharvest and LTRdigest wrapped in the GenomeTools  
7  
8 package [32]. The LTR identification pipeline was run iteratively to collect both recent (sequence  
9  
10 similarity  $\geq 99\%$ ) and old (sequence similarity  $\geq 85\%$ ) LTR retrotransposons. Candidates from each  
11  
12 run were filtered based on the elements typically encoded by LTR retrotransposons. The default  
13  
14 parameters (-minlenltr 100 -maxlenltr 6000 -mindistltr 1500 -maxdistltr 25000 -mintsd 5 -maxtsd  
15  
16 5 -motif tgca) were used in LTR calling according to Campbell *et al.* [30]. An initial repeat  
17  
18 masking of *A. eriantha* ‘White’ genome was performed with the repeat library derived by  
19  
20 combining the identified MITEs and LTR transposons. The repeat masked genome was fed to  
21  
22 RepeatModeler (<http://www.repeatmasker.org/RepeatModeler/>) to identify novel repeat families.  
23  
24 Finally, all identified repeat sequences were combined and searched against a plant protein  
25  
26 database where transposons encoding proteins were excluded. Elements with significant similarity  
27  
28 to plant genes were removed. The final repeat library contained 1,670 families, and 526 of them  
29  
30 were potentially novel repeat families. We used this species-specific repeat library to mask the *A.*  
31  
32 *eriantha* ‘white’ genome. Approximately 43.3% of the *A. eriantha* ‘White’ genome was masked,  
33  
34 and the largest family of repeats was LTR transposons (Table S3). Repeat content identified in *A.*  
35  
36 *eriantha* ‘White’ was much higher than that in *A. chinensis* (e.g. 36% in Hongyang [15]), and this  
37  
38 difference could be largely due to the improvement of the repeat region assembly with PacBio  
39  
40 long reads. In addition, variations between the two kiwifruit species could also contribute to this  
41  
42 difference.  
43  
44  
45  
46  
47  
48  
49  
50  
51

### 52 53 54 55 ***Prediction and functional annotation of protein-coding genes*** 56 57 58 59 60 61 62 63 64 65

1  
2  
3  
4 Protein-coding genes were predicted from the repeat-masked *A. eriantha* ‘White’ genome with the  
5  
6 MAKER-P program [30] (version 2.31.10), which integrates evidence from protein homology,  
7  
8 transcripts and *ab initio* predictions. The homology-based evidence was derived by aligning  
9  
10 proteomes from 20 plant species to the ‘White’ genome assembly with exonerate (v2.26.1;  
11  
12 <https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate>). SNAP [33], AUGUSTUS  
13  
14 [34] (version 3.3), and GeneMark-ES [35] (version 4.35) were used for *ab initio* gene predictions.  
15  
16  
17 RNA-Seq data generated in this study were assembled *de novo* with Trinity and the assembled  
18  
19 contigs were aligned to the ‘White’ genome assembly to provide transcript evidence. Predictions  
20  
21 supported by the three different sources of evidence were finally integrated by MAKER-P, which  
22  
23 resulted in a total of 52,514 primitive gene models. We then filtered and polished these gene  
24  
25 models by two steps. First, we combined our RNA-Seq data with others collected from a previous  
26  
27 study [36], and mapped the reads to the ‘White’ genome using the STAR program [22], and a total  
28  
29 of 266 million read pairs were mapped. Based on the mapping, raw count for each predicted gene  
30  
31 model was derived and then normalized to CPM (counts per million mapped read pairs). Gene  
32  
33 models with ultra-low expression (CPM < 0.1) were less likely to be real genes. Furthermore, we  
34  
35 found that these lowly expressed genes had relatively high annotation edit distance (AED) score,  
36  
37 an indication of low-confidence as defined by the MAKER-P program. Therefore, for gene models  
38  
39 with CPM < 0.1, we only kept those containing both pfam domains and homologous sequences in  
40  
41 the NCBI nr protein database. After this filtering process 42,613 gene models were kept. Second,  
42  
43 the predicted protein-coding genes of kiwifruit *A. chinensis* ‘red5’ have been manually curated  
44  
45 [16], and therefore these gene models should have relatively higher accuracy and could be used to  
46  
47 modify *A. eriantha* ‘White’ gene models whose predictions were not consistently supported by the  
48  
49 different types of evidence. To this end, we performed another two *ab initio* predictions using  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 BRAKER2 [37] and GeMoMa [38] (version 1.5.2) with ‘red5’ proteome as the sole evidence.  
5  
6 These two predictions were compared with the gene models predicted by MAKER-P.  
7  
8 Consequently, a total of 237 gene models not predicted by MAKER-P were added and another 415  
9  
10 gene models which had better predictions by BRAKER2 or GeMoMa were used to replace the  
11  
12 corresponding gene models predicted by MAKER-P. Finally, we obtained a total of 42,850  
13  
14 protein-coding genes in the *A. eriantha* ‘White’ genome, with a mean coding sequence (CDS) size  
15  
16 of 1,004 bp and containing an average of five exons.  
17  
18

19  
20  
21 The predicted genes were functionally annotated by blasting their protein sequences against  
22  
23 TAIR, Swiss-Prot and TrEMBL databases with an E-value cutoff of 1e-5. Functional descriptions  
24  
25 of the protein hits were assembled with the AHRD program  
26  
27 (<https://github.com/groupschoof/AHRD>) and transferred to *A. eriantha* genes. Protein domains  
28  
29 were identified using InterProScan [39] (version 5.29-68.0) by searching the protein sequences  
30  
31 against domain databases including PANTHER, Pfam, SMART, and PROSITE. The Gene  
32  
33 Ontology (GO) terms were assigned to the *A. eriantha* ‘White’ predicted genes using the Blast2GO  
34  
35 program [40] with entries from NCBI protein database and InterProScan. Collectively, 91.2%  
36  
37 (N=39,075) of the predicted genes contain at least one annotation from the above databases (Table  
38  
39 S4).  
40  
41  
42  
43  
44  
45  
46  
47

#### 48 ***Evolutionary and comparative genomic analysis***

49

50 To infer the divergence time between *A. eriantha* and *A. chinensis*, we identified gene orthology  
51  
52 between the two species using MCSanX [41] and calculated synonymous substitution rate (Ks)  
53  
54 between each orthologous pair. Three additional species, cultivated tomato (*Solanum*  
55  
56 *lycopersicum*), wild tomato (*S. penellii*) and potato (*S. tuberosum*), were also included in the  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 analysis. The Ks distribution (Fig. 4a) suggested that the divergence between the two kiwifruit  
5  
6 species was earlier than that between the two tomato species. We dated the divergence by assuming  
7  
8 a strict molecular clock [42], and the time when *A. eriantha* and *A. chinensis* diverged from the  
9  
10 common ancestor was estimated to be ~3.3 million years ago (Mya), compared to ~1.9 Mya  
11  
12 between *S. lycopersicum* and *S. penellii* and ~6.0 Mya between *S. lycopersicum* and *S. tuberosum*.  
13  
14 Gene family evolution was analyzed by comparing genomes of *A. eriantha*, *A. chinensis*, *S.*  
15  
16 *lycopersicum*, *S. tuberosum*, *Vitis vinifera*, *Arabidopsis thaliana* and *Oryza sativa*. A total of  
17  
18 17,593 orthogroups were defined by OrthoFinder [43] (version 2.2.6), among which 1,246 were  
19  
20 single-copy gene families (Fig. 4b). The single-copy family genes were aligned and concatenated  
21  
22 to build a species phylogenetic tree using IQ-TREE [44] (version 1.5.5) with a best-fitting model  
23  
24 (Fig. 4c). Gene family expansion/contraction along the branches of the phylogenetic tree was  
25  
26 analyzed by CAFÉ [45] (version 4.1). Finally, a total of 1,740 and 1,345 gene families were found  
27  
28 apparently expanded and contracted, respectively, in *A. eriantha* (Fig. 4c).  
29  
30  
31  
32  
33  
34  
35  
36  
37

## 38 **Conclusion**

39  
40 Here, we report a high-quality reference genome of kiwifruit *A. eriantha* cv. White. The assembly  
41  
42 from single-molecular sequencing combined with Hi-C scaffolding yielded a much more  
43  
44 continuous and complete genome than the two previously published kiwifruit genomes. This  
45  
46 genome will provide a valuable source for exploration of genetic basis of unique traits in kiwifruit  
47  
48 and also facilitate the studying of sexual determination loci in the dioecious plants.  
49  
50  
51  
52  
53  
54

## 55 **Availability of supporting data**

56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 This Whole Genome Shotgun project has been deposited at DBJ/ENA/GenBank under the  
5  
6 accession QOVS00000000. The version described in this paper is version QOVS01000000. Raw  
7  
8 sequencing reads have been deposited in the Sequence Read Archive (SRA) database under the  
9  
10 accession number SRP155011. The *Actinidia eriantha* ‘White’ genome sequence and the  
11  
12 annotation are also available at Kiwifruit Information Resource (<http://bdg.hfut.edu.cn/kir/>).  
13  
14  
15  
16  
17  
18

### 19 **Competing interests**

20  
21 The authors have no competing interests to declare.  
22  
23  
24  
25

### 26 **Abbreviation**

27  
28 CTAB: Cetyl trimethylammonium bromide;  
29

30  
31 NCBI: National Center for Biotechnology Information;  
32

33  
34 RNA-Seq: RNA sequencing;  
35

36  
37 SMRT: Single Molecule Real-Time;

38  
39 MITE: miniature inverted transposable element;

40  
41 LTR: long terminal repeat;  
42

43  
44 CPM: counts per million mapped read pairs;

45  
46 Mya: million years ago  
47  
48  
49

### 50 **Acknowledgement**

51  
52  
53 This work was supported by grants from the National Natural Science Foundation of China  
54  
55 (31471157 and 31700266), National Foundation for Germplasm Repository of Special  
56  
57 Horticultural Crops in Central Mountain Areas of China (NJF2017-69), National Science Fund for  
58  
59  
60  
61  
62  
63  
64  
65



1  
2  
3  
4 Distinguished Young Scholars (30825030), Key Project from the Government of Sichuan Province  
5  
6 (2013NZ0014, 2016NZ0105), Key Project from the Government of Anhui Province  
7  
8 (2012AKKG0739;1808085MC57), and the US National Science Foundation (IOS-1339287 and  
9  
10 IOS-1539831).  
11  
12

### 13 14 15 16 **Author contribution**

17  
18 W.T., X.S. and J.Y. contributed equally to this work. W.T., J.Y., X.T., Y.Y., X.N., M.M., D.Z.,  
19  
20 S.H., W.S., C.F. and M.L. collected plant samples, extracted DNA/RNA, and performed  
21  
22 transcriptome sequencing and gene expression analyses; W.T., X.S., J.Y., X.T., C.J., Z.F. and Y.L.  
23  
24 performed DNA sequencing, genome assembly, gene annotation, evolution and comparative  
25  
26 genomic analyses, and website construction; X.S., W.T., Z.F. and Y.L. wrote and revised the  
27  
28 manuscript; Y.L. and Z.F. conceived strategies, designed experiments and managed projects. All  
29  
30 authors read and approved the manuscript.  
31  
32  
33  
34  
35  
36  
37

### 38 **Figure legends**

39  
40  
41 **Figure 1.** Tree and fruits of *A. eriantha* cv. White.  
42  
43  
44

45  
46 **Figure 2.** Chromatin interaction map of *A. eriantha* derived from Hi-C data. Each group represents  
47  
48 an individual chromosome.  
49  
50

51  
52  
53 **Figure 3.** Genome of *A. eriantha* and synteny between the two kiwifruit species. (a) Genome  
54  
55 landscape of *A. eriantha* cv. White. Track A: gene density, Track B: repeat density, Track C: GC  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 content; all were calculated in a 500-kb window; (b) Genome synteny between *A. eriantha* cv.  
5 White and *A. chinensis* cv red5.  
6  
7  
8  
9

10  
11 **Figure 4.** Evolutionary and comparative genomic analyses. (a) Distribution of synonymous  
12 substitution rate (Ks) between *A. eriantha* and *A. chinensis*, *S. lycopersicum* and *S. penellii*, and *S.*  
13 *lycopersicum* and *S. tuberosum*; (b) Orthogroups shared by selected species; (c) Species  
14 phylogenetic tree and gene family evolution. Numbers on the branch indicate counts of gene family  
15 that under either expansion (red) or contraction (green).  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25

26 **Figure S1.** An example of genome assembly inconsistency between *A. eriantha* cv. White and *A.*  
27 *chinensis* cv red5. (a) A chromosomal segment assembled into the Chr23 in ‘red5’, is syntenic to  
28 the region located at the terminus of Chr19 in ‘White’ (b) Snapshots of Illumina mate-pair reads  
29 mapped to the junctions of the break point as well as nearby regions supporting the assembly of  
30 ‘White’.  
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

## Reference

1. Ferguson AR, Ferguson LR. Are kiwifruit really good for you? *Acta Hort* 2013;**610**:131-138
2. Richardson DP, Ansell J, Drummond LN. The nutritional and health attributes of kiwifruit: a review. *Eur J Nutr.* 2018;1-18.
3. Li JQ, Li XW, Soejarto DD. Actinidiaceae. In: Wu ZY, Raven PH, Hong DY, eds. *Flora of China*. Beijing: Science Press & St. Louis: Missouri Plant Garden Press; 2007. **12**:pp334-362.
4. Ferguson AR, Huang H. Genetic resources of kiwifruit: domestication and breeding. *Hortic Rev.* 2007;**33**:1-121.
5. Testolin R. Kiwifruit (*Actinidia* spp.) in Italy: The history of the industry, international scientific cooperation and recent advances in genetics and breeding. *ISHS Acta Horticulturae* 2015;47-61.
6. Jo YS, Cho HS, Park MY, Bang GP. Selection of a sweet *Actinidia eriantha* 'bidan'. *ISHS Acta Horticulturae* 2017; 253-258.
7. Wu Y, Xie M, Zhang Q et al. Characteristics of 'White': a new easy-peel cultivar of *Actinidia eriantha*. *N Z J Crop Hortic Sci* 2009;**37**(4):369-373.
8. Atkinson RG, Sharma NN, Hallett IC et al. *Actinidia eriantha*: a parental species for breeding kiwifruit with novel peelability and health attributes. *N Z J For Sci* 2009;**39**:207-216.
9. Guo R, Landis JB, Moore MJ et al. Development and application of transcriptome-derived microsatellites in *Actinidia eriantha* (Actinidiaceae). *Front Plant Sci* 2017;**8**:1383.
10. Prakash R, Hallett IC, Wong SF et al. Cell separation in kiwifruit without development of a specialised detachment zone. *BMC Plant Biol.* 2017;**17**(1):86.
11. Shi ZJ, Zhang HQ, Hui Q et al. The resistance evaluation of different kiwifruit varieties to canker. *Acta Agriculturae Zhejiangensis* 2014;**26**(3):752-759
12. Zhang D, Gao C, Li R et al. TEOA, a triterpenoid from *Actinidia eriantha*, induces autophagy in SW620 cells via endoplasmic reticulum stress and ROS-dependent mitophagy. *Arch Pharm Res* 2017;**40**(5):579-591.
13. Wang T, Ran Y, Atkinson RG et al. Transformation of *Actinidia eriantha*: a potential species for functional genomics studies in Actinidia. *Plant Cell Rep.* 2006;**25**(5):425.
14. Wu JG, Ma L, Lin SH et al. Anticancer and anti-angiogenic activities of extract from *Actinidia eriantha* Benth root. *J Ethnopharmacol* 2017;**203**:1-10.
15. Huang S, Ding J, Deng D et al. Draft genome of the kiwifruit *Actinidia chinensis*. *Nat Commun* 2013;**4**:2640.
16. Pilkington SM, Crowhurst R, Hilario E et al. A manually annotated *Actinidia chinensis* var. *chinensis* (kiwifruit) genome highlights the challenges associated with draft genomes and gene prediction in plants. *BMC Genomics* 2018;**19**(1):257.
17. Hopping ME. Flow cytometric analysis of *Actinidia* species. *N Z J Bot* 1994;**32**(1):85-93.
18. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**:2114-20.
19. Leggett RM, Clavijo BJ, Clissold L et al. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics* 2013;**30**(4):566-568.
20. Rao SS, Huntley MH, Durand NC et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014;**159**(7):1665-1680.
21. Haas BJ, Papanicolaou A, Yassour M et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* 2013;**8**(8):1494.
22. Dobin A, Davis CA, Schlesinger F et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**(1):15-21.
23. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 2015;**31**(2):166-169.
24. Koren S, Walenz BP, Berlin K et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* 2017;**27**(5):722-736.
25. Chakraborty M, Baldwin-Brown JG, Long AD et al. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res* 2016;**44**(19):e147.

- 1
- 2
- 3
- 4 26. Walker BJ, Abeel T, Shea T et al. Pilon: an integrated tool for comprehensive microbial variant
- 5 detection and genome assembly improvement. *PLoS One* 2014;**9**(11):e112963.
- 6
- 7 27. Servant N, Varoquaux N, Lajoie BR et al. HiC-Pro: an optimized and flexible pipeline for Hi-C
- 8 data processing. *Genome Biol* 2015;**16**(1):259.
- 9
- 10 28. Burton JN, Adey A, Patwardhan RP et al. Chromosome-scale scaffolding of de novo genome
- 11 assemblies based on chromatin interactions. *Nat Biotechnol* 2013;**31**(12):1119.
- 12
- 13 29. Kurtz S, Phillippy A, Delcher AL et al. Versatile and open software for comparing large genomes.
- 14 *Genome Biol* 2004;**5**(2):R12.
- 15
- 16 30. Campbell M, Law M, Holt C et al. MAKER-P: a tool-kit for the rapid creation, management, and
- 17 quality control of plant genome annotations. *Plant Physiol* 2013;**164**(2):513-524.
- 18
- 19 31. Han Y, Wessler SR. MITE-Hunter: a program for discovering miniature inverted-repeat
- 20 transposable elements from genomic sequences. *Nucleic Acids Res* 2010;**38**(22):e199.
- 21
- 22 32. Gremme G, Steinbiss S, Kurtz S. GenomeTools: a comprehensive software library for efficient
- 23 processing of structured genome annotations. *IEEE/ACM Trans Comput Biol Bioinform*
- 24 **2013**;**10**(3):645-656.
- 25
- 26 33. Korf I. Gene finding in novel genomes. *BMC Bioinformatics* 2004;**5**(1):59.
- 27
- 28 34. Stanke M, Keller O, Gunduz I et al. AUGUSTUS: ab initio prediction of alternative transcripts.
- 29 *Nucleic Acids Res.* 2006;**34**:W435-W439.
- 30
- 31 35. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO et al. Gene identification in novel eukaryotic
- 32 genomes by self-training algorithm. *Nucleic Acids Res* 2005;**33**(20):6494-6506.
- 33
- 34 36. Wang Z, Liu Y, Li D et al. Identification of circular RNAs in kiwifruit and their species-specific
- 35 response to bacterial canker pathogen invasion. *Front Plant Sci.* 2017;**8**:413.
- 36
- 37 37. Hoff KJ, Lange S, Lomsadze A et al. BRAKER1: Unsupervised RNA-Seq-Based Genome
- 38 Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* 2016;**32**(5):767-769.
- 39
- 40 38. Keilwagen J, Wenk M, Erickson JL et al. Using intron position conservation for homology-based
- 41 gene prediction. *Nucleic Acids Res* 2016;**44**(9):e89.
- 42
- 43 39. Zdobnov EM, Apweiler R. InterProScan—an integration platform for the signature-recognition
- 44 methods in InterPro. *Bioinformatics* 2001;**17**(9):847-848.
- 45
- 46 40. Conesa A, Götz S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int*
- 47 *J Plant Genomics* 2008;**2008**:619832.
- 48
- 49 41. Wang Y, Tang H, DeBarry JD et al. MCSanX: a toolkit for detection and evolutionary analysis of
- 50 gene synteny and collinearity. *Nucleic Acids Res* 2012;**40**(7):e49.
- 51
- 52 42. Ossowski S, Schneeberger K, Lucas-Lledó JI et al. The rate and molecular spectrum of spontaneous
- 53 mutations in *Arabidopsis thaliana*. *Science* 2010;**327**(5961):92-94.
- 54
- 55 43. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons
- 56 dramatically improves orthogroup inference accuracy. *Genome Biol* 2015;**16**(1):157.
- 57
- 58 44. Nguyen LT, Schmidt HA, von Haeseler A et al. IQ-TREE: a fast and effective stochastic algorithm
- 59 for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2014;**32**(1):268-74.
- 60
- 61 45. De Bie T, Cristianini N, Demuth JP et al. CAFE: a computational tool for the study of gene family
- 62 evolution. *Bioinformatics* 2006;**22**(10):1269-71.
- 63
- 64
- 65



**Figure 1**

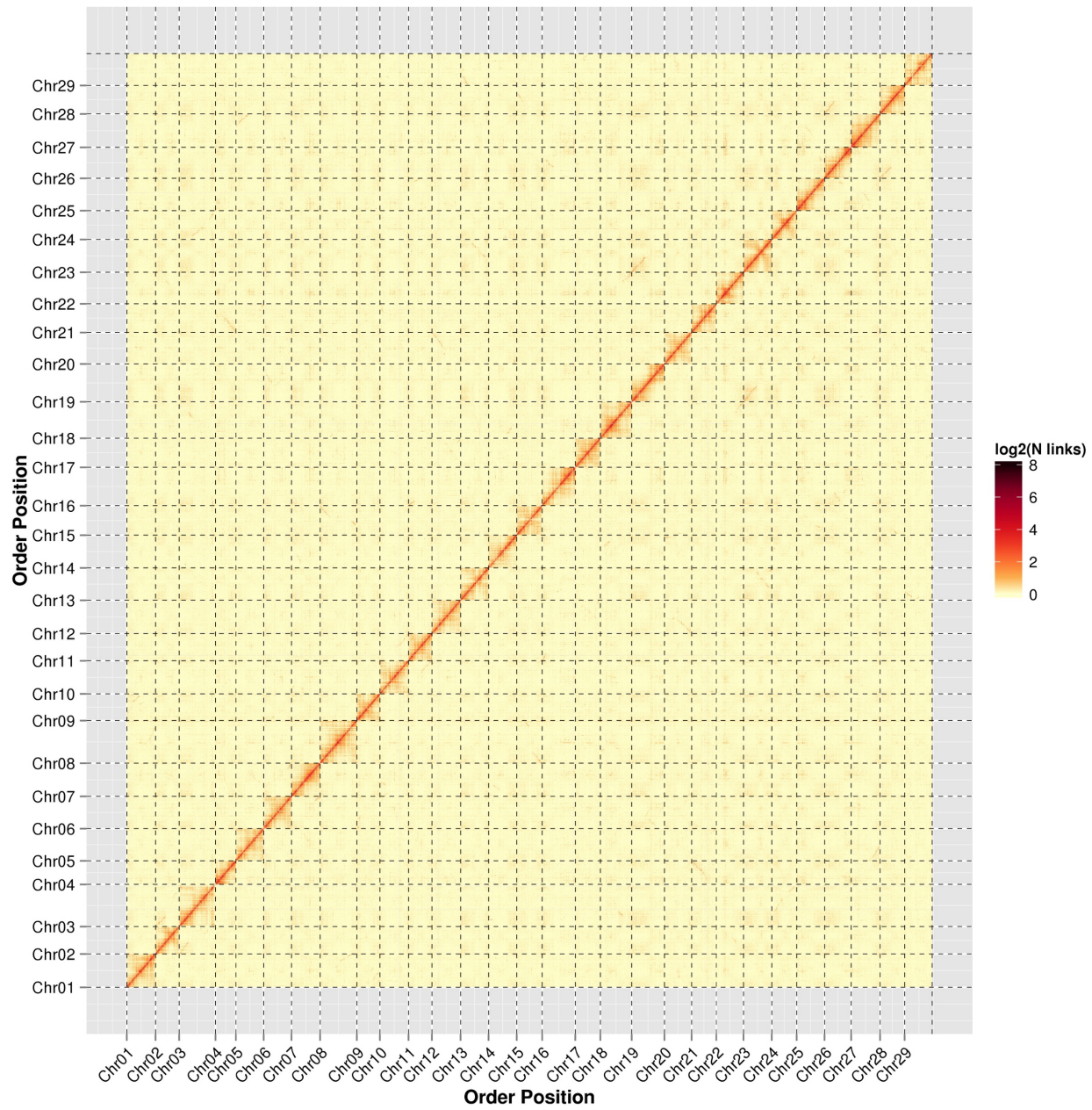
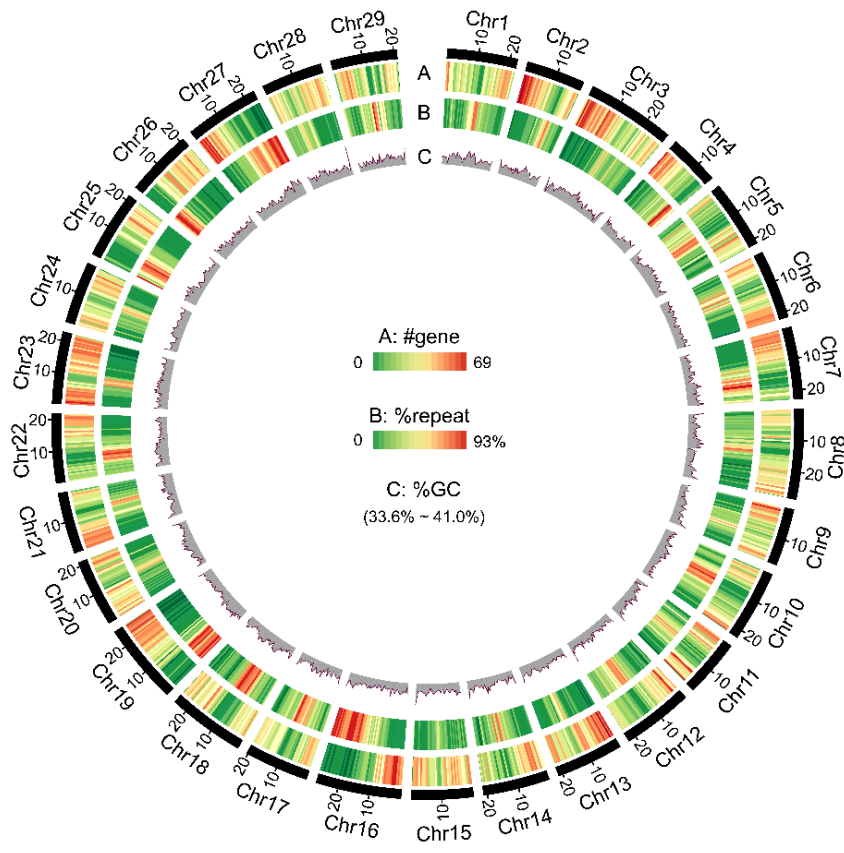


Figure 2.



a



b

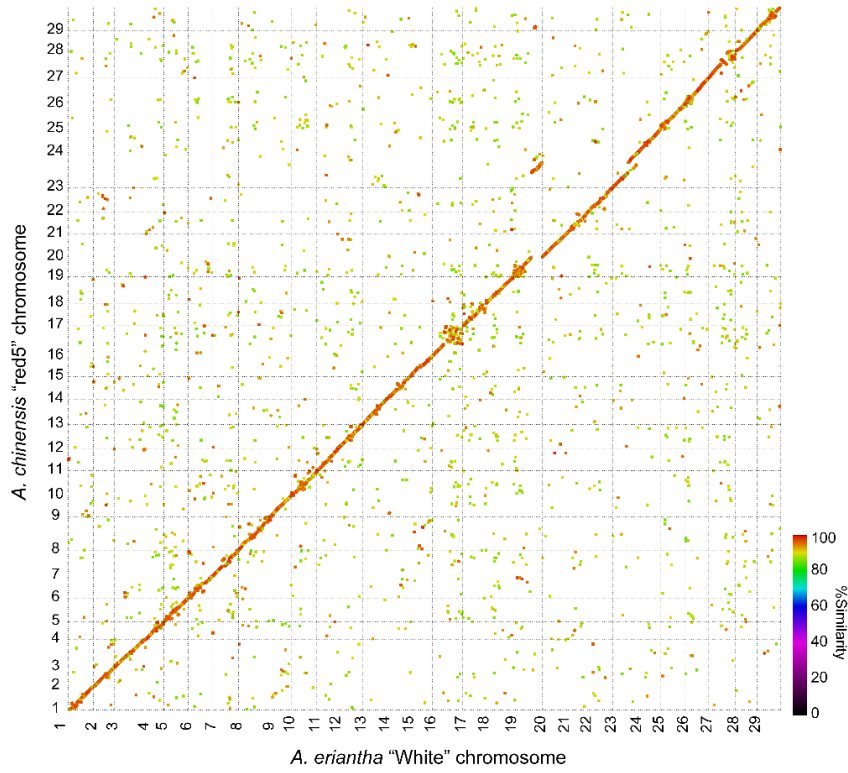


Figure 3

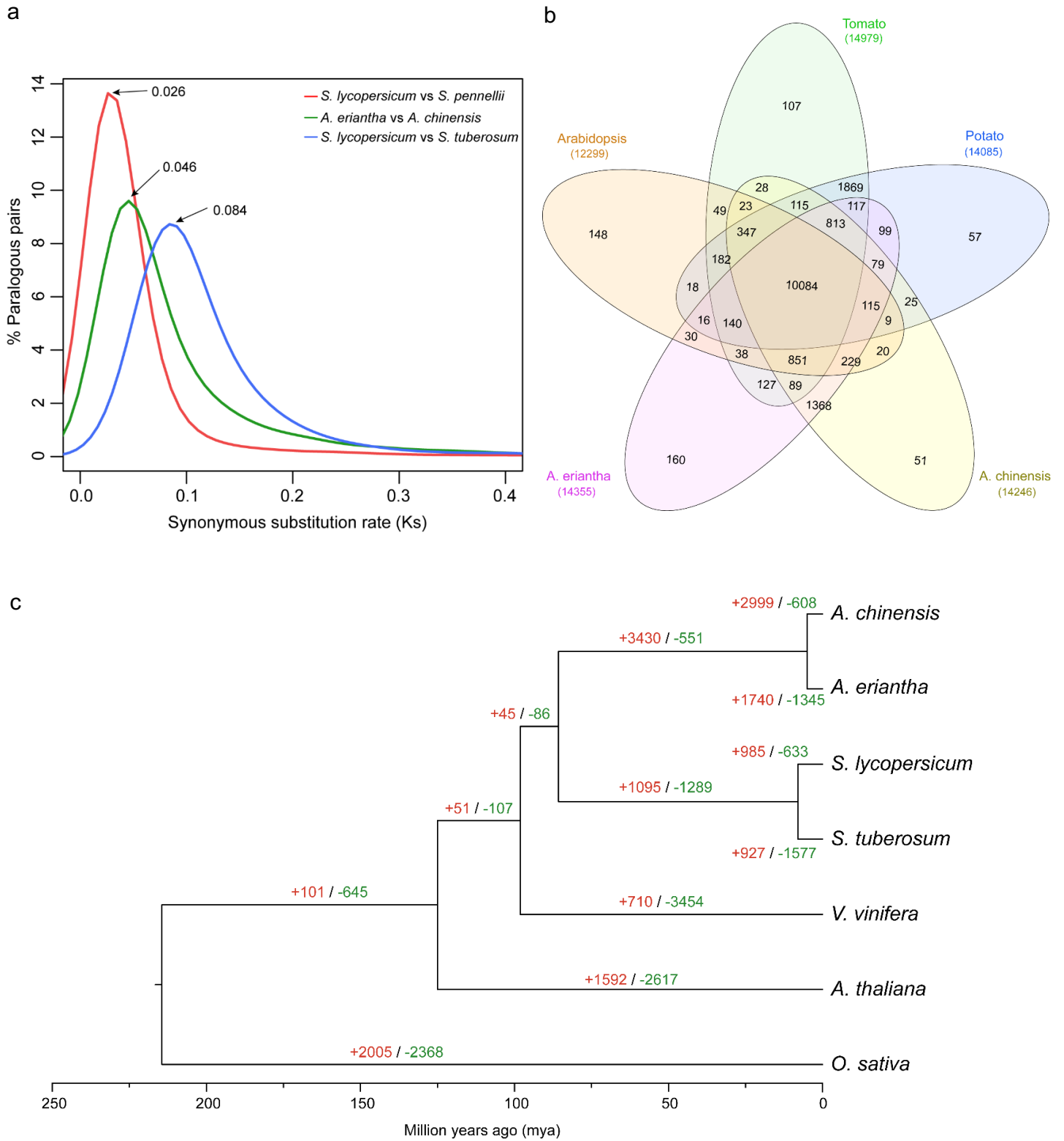

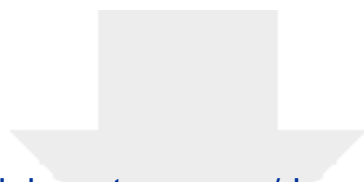


Figure 4





Click here to access/download  
**Supplementary Material**  
FigureS1.pdf



Click here to access/download  
**Supplementary Material**  
Supp\_Tables.xlsx

