

1 **Genome analysis and Hi-C assisted assembly of *Elaeagnus***
2 ***angustifolia* L., a deciduous tree belonging to *Elaeagnaceae***

3
4 Yunfei Mao¹, Qin Hu², Manman Zhang¹, Lu Yang¹, Lulu Zhang¹, Yunyun Wang¹,
5 Yijun Yin¹, Huiling Pang¹, Yeping Liu¹, Xiafei Su¹, Song Li³, XinXing Cui³,
6 Fengwang Ma⁴, Naibin Duan⁵, Donglin Zhang⁶, Yanli Hu¹, Zhiquan Mao¹, Xuesen
7 Chen¹, Xiang Shen^{1,*}.

8
9 ¹*College of Horticultural Science and Engineering/State Key Laboratory of Crop Biology, Shandong Agricultural*
10 *University, Tai'an, China.*

11 ²*College of Resources and Environment, Shandong Agricultural University, Tai'an, China.*

12 ³*Biomarker Technologies Corporation, Beijing, China.*

13 ⁴*College of Horticulture, Northwest Agriculture and Forestry University, Yangling, China.*

14 ⁵*Germplasm Resource Center of Shandong Province, Shandong Academy of Agricultural Sciences, Jinan, China.*

15 ⁶*Dept of Horticulture, University of Georgia, Athens, USA.*

16
17 **Abstract:** *Elaeagnus angustifolia* L. is a deciduous tree of the *Elaeagnaceae* family. It is widely
18 used in the study of abiotic stress tolerance in plants and for the improvement of
19 desertification-affected land due to its characteristics of drought resistance, salt tolerance, cold
20 resistance, wind resistance, and other environmental adaptation. Here, we report the complete
21 genome sequencing using the Pacific Biosciences (PacBio) platform and Hi-C assisted assembly
22 of *E. angustifolia*. A total of 44.27 Gb raw PacBio sequel reads were obtained after filtering out
23 low-quality data, with an average length of 8.64 Kb. And 39.56 Gb clean reads was obtained, with
24 a sequencing coverage of 75×, and Q30 ratio > 95.46%. The 510.71 Mb genomic sequence was
25 mapped to the chromosome, accounting for 96.94% of the total length of the sequence, and the
26 corresponding number of sequences was 269, accounting for 45.83% of the total number of
27 sequences. The genome sequence study of *E. angustifolia* can be a valuable source for the
28 comparative genome analysis of the *Elaeagnaceae* family members, and can help to understand
29 the evolutionary response mechanisms of the *Elaeagnaceae* to drought, salt, cold and wind
30 resistance, and thereby provide effective theoretical support for the improvement of
31 desertification-affected land.

32 .

33 **Keywords:** *Elaeagnus angustifolia* L.; PacBio sequencing; Hi-C assisted assembly; evolutionary
34 response mechanism; desertification-affected land

35

36 **Introduction**

37 *Elaeagnus angustifolia* L., also known as silver willow and cinnamon, is a deciduous tree
38 belonging to the *Elaeagnaceae* family (Fig. 1). It is native to central and western Africa and is
39 distributed in the United States, Canada, the Mediterranean coast, southern Russia, Iran, and India.
40 It shows a wide distribution area in China, where is is distributed in the Xinjiang, Gansu, Ningxia,
41 Inner Mongolia, and other provinces(Wang *et al.*, 2014). The fruit, branches, leaves, and flowers
42 of *E. angustifolia* can be used as medicine owing to multiple beneficial properties. The fruit is rich

*Correspondence (Tel +86 13705383303; e-mail shenx@sdau.edu.cn)

43 in sugars, flavonoids, and other substances that can regulate the blood circulation of the human
44 body and improve the immunity of the body; the branches, leaves, and flowers are beneficial for
45 anti-aging, and treatment of burns, bronchitis, dyspepsia, and neurasthenia(Min *et al.*,2006; Vitas
46 *et al.*, 2004; Wang *et al.*,2006). The flowers are also used for extracting aromatic oil, which is used
47 as a flavoring raw material in soap(Liu *et al.*, 2003).

48 At present, land desertification is a serious global phenomenon. Due to economic
49 development needs, the effects of various methods such as terraced fields and grazing control to
50 recover from land desertification are not significant in Spain, Greece, Turkey and other
51 countries(Salvati *et al.*, 2016). *E. angustifolia* shows the characteristics of drought resistance, salt
52 tolerance, cold resistance, wind resistance, easy reproduction, and strong adaptability(Huang *et al.*,
53 2005). The root rhizobium has important effects on nitrogen fixation and soil improvement, which
54 can reform saline-alkali land and improve desertification-affected land(Liu, 2015). In recent years,
55 *E. angustifolia* has been cultivated in Hebei, Heilongjiang, Henan, Shanxi, Shandong, and other
56 provinces in China(Guo *et al.*, 2008).

57 Although the nrDNA ITS sequence data of *Elaeagnaceae* are abundant in the GenBank at
58 present(He, 2012), studies on genome sequencing of *Elaeagnaceae* have not yet been reported,
59 and the genome is an important basis for analyzing the evolution of *Elaeagnaceae*. At present,
60 Pacific BioSciences(PacBio) technology, a third-generation sequencing technology, and Hi-C
61 assisted assembly technology have become increasingly reliable and the genome sequencing has
62 been completed for *Saccharum spontaneum* L.(Zhang *et al.*, 2018) and *Ammopiptanthus*
63 *nanus*(Gao *et al.*, 2018).

64 In this study, we applied PacBio technology and Hi-C assisted assembly technology to
65 sequence the genome of *E. angustifolia*, which is a valuable source for comparative genomic
66 analysis of the *Elaeagnaceae* family members. Genome sequencing can help understand the
67 response mechanism of the *Elaeagnaceae* to drought, salt, cold and wind resistance, and provide
68 an effective theoretical basis for planting *E. angustifolia* to recover from global land
69 desertification.



70
71 Figure 1. *Elaeagnus angustifolia*

72

73 **Materials and Methods**

74 **Sample collection**

75 Samples from an *Elaeagnus angustifolia* L. tree (imported from Xinjiang province, NCBI

76 Taxonomic ID, 36777) were collected from the south campus of Shandong Agricultural University
77 for genomic DNA sequencing, and Hi-C assisted assembly.

78

79 **Genomic DNA sequencing and Hi-C assisted assembly**

80 After collection, tissues were immediately immersed in liquid nitrogen and stored until DNA
81 extraction. DNA was extracted using the Cetyltrimethyl Ammonium Bromide (CTAB) method.
82 The quality of the extracted genomic DNA was checked using 1% agarose gel electrophoresis, and
83 the concentration was quantified using a Qubit fluorimeter (Invitro-gen, Carlsbad, CA, USA).
84 After checking the quantity and quality of the DNA sample, the library was constructed as shown
85 in Supplementary Figure S1 in the order from left to right as shown in Supplementary Figure S2.

86

87 **Results and discussion**

88 **Genomic results and statistics**

89 We constructed two 270-bp libraries using genomic DNA of *E. angustifolia* samples. A total of
90 60.15 Gb of high-quality data was sequenced and filtered on Illumina Hiseq sequencing platform
91 (San Diego, CA, USA), and the total sequencing depth was about 131×, which met the sequencing
92 requirement of more than 50× (Supplementary Table S1). A total of 5,125,675 subreads were
93 obtained by filtering low-quality data, and a total of 44.27 Gb raw PacBio sequel reads were
94 obtained, with an average length of 8.64 kb (Supplementary Table S2). The subread N50 was
95 12,635 bp, and the average length was 8,636 bp (Supplementary Table S3). Subreads were
96 corrected and assembled by Canu(Koren *et al.*, 2017), and the estimated genome size was found to
97 be 781.09 Mb and Contig N50 was 486.92 Kb (Supplementary Table S4).

98 A kmer map of $k = 19$ was constructed using the two 270-bp library data (Supplementary
99 Figure S3), which was used to evaluate genome size, repeat sequence ratio, and heterozygosity.
100 The highest peak in the kmer distribution curve was found at the k-mer depth of 111. The
101 sequences with kmer depth more than twice of the corresponding depth of the main peak, i.e. kmer
102 sequences with a depth greater than 223, were repetitive sequences. The sequence with kmer depth
103 appearing at half of the depth corresponding to the main peak, i.e. the kmer sequence with depth
104 appearing around 55 was a heterozygous sequence. The total number of kmer obtained from
105 sequencing data was 52,917,129,364. After removing those with depth abnormality, a total of
106 51,064,317,165 kmer sequences were used for the estimation of genome length, whose calculated
107 length was about 456.24 Mbp. Based on distribution of kmer, the genome of this species was
108 found to be a complex genome with high heterozygosity, with the content of repeat sequences
109 estimated to be about 39.24%, and the degree of heterozygosity estimated to be about 1.47%.

110 Due to the relatively low conservation of repeat sequences among species, it is necessary to
111 construct a specific repeat sequence database for the prediction of repeat sequences for specific
112 species. With the help of LTR FINDER v1.05(Xu *et al.*, 2007), MITE Hunter(Han *et al.*, 2010),
113 RepeatScout v1.0.5(Price *et al.*, 2005), and piler-df v2.4(Edgar *et al.*, 2005), the repeat sequence
114 database of *E. angustifolia* genome was constructed based on the structure prediction and the
115 principle of de novo prediction. The database was classified by PASTEClassifier(Wicker *et al.*,
116 2007), and then merged with the database of Repbase(Jurka *et al.*, 2005) as the final repetitive
117 sequence database, and then repeated sequences were identified based on the constructed repeat

118 sequence database using RepeatMasker v4.0.6(Tarailo-Graovac *et al.*, 2009) software. The
 119 prediction yielded a repeat of about 263.44 Mb, accounting for 50.01%. The detailed prediction
 120 results are shown in Table 1.

121 Table 1 Repeating sequence statistics

Type	Number	Length (bp)	Percentage (%)
ClassI/DIRS	57,476	39,462,537	7.49
ClassI/LINE	17,420	6,130,877	1.16
ClassI/LTR	1,192	1,341,892	0.25
ClassI/LTR/Copia	170,211	112,045,341	21.27
ClassI/LTR/Gypsy	89,775	74,832,142	14.2
ClassI/PLE/LARD	87,646	29,294,594	5.56
ClassI/SINE	3,134	580,471	0.11
ClassI/TRIM	4,191	2,037,167	0.39
ClassI/Unknown	277	111,257	0.02
ClassII/Crypton	10	712	0
ClassII/Helitron	9,255	2,368,390	0.45
ClassII/MITE	8,168	1,544,498	0.29
ClassII/Maverick	1,511	278,308	0.05
ClassII/TIR	26,737	12,037,464	2.29
ClassII/Unknown	7,008	1,957,120	0.37
PotentialHostGene	4,766	1,419,371	0.27
SSR	41,290	8,338,047	1.58
Unknown	96,908	27,036,916	5.13
Total without overlap	626,975	263,437,176	50.01

122

123 TopHat(Trapnell *et al.*, 2009) was used to compare the raw transcriptome data with the
 124 genome of *E. angustifolia*, and the number of bases in the Exon, Intron, and Intergenic regions
 125 were separately counted to evaluate the results of the gene prediction (Supplementary Table S5).
 126 The prediction of the genetic structure of *E. angustifolia* mainly used de novo prediction,
 127 homologous species prediction, and Unigene prediction, and then integrated the prediction results
 128 using EVM v1.1.1(Haas *et al.*, 2008) software. Genscan(Burge *et al.*, 1997), Augustus v2.4(Stanke
 129 *et al.*, 2003), GlimmerHMM v3.0.4(Majoros *et al.*, 2004), GeneID v1.4(Blanco *et al.*, 2007),
 130 SNAP (version 2006-07-28) (Korf, 2004) were used for head-to-head prediction. GeMoMa
 131 v1.3.1(Keilwagen *et al.*, 2016) was used for de novo prediction. His v2.0.4(Pertea *et al.*, 2016) and
 132 Stringtie v1.2.3(Pertea *et al.*, 2016) were used for assembly based on reference transcript, and
 133 TransDecoder v2.0(Haas *et al.*, 2016)and gene marks-t v5.1(Tang *et al.*, 2015) was used for gene
 134 prediction. PASA v2.0.2(Campbell *et al.*, 2006) was used to predict the Unigene sequences

135 without reference assembly based on transcriptome data. Finally, EVM v1.1.1(Haas *et al.*, 2008)
 136 was used to integrate the prediction results obtained by the above three methods, and 31,730 genes
 137 were obtained after modification with PASA v2.0.2. The specific predicted information is shown
 138 in Table 2 and Supplementary Table S6. The number of genes supported by the three prediction
 139 methods was integrated, as shown in Supplementary Figure S4. As shown, the number of genes
 140 supported by homologous prediction and transcriptome prediction resulted in 30,771 genes,
 141 accounting for 96.98%, indicating the high prediction quality. At the same time, according to the
 142 gene function annotation, 96.89% of the genes could be annotated into NR and other databases,
 143 which further indicated that the gene prediction was reliable.

144 BLAST v2.2.31(Birney *et al.*, 2004) with an E-value cutoff of 1E-5 was used to align the
 145 predicted gene sequences with functional databases such as NR(Griffiths-Jones *et al.*, 2005),
 146 KOG(Griffiths-Jones *et al.*, 2006), GO(Nawrocki *et al.*, 2013), KEGG(Lowe *et al.*, 1997), and
 147 TrEMBL(She *et al.*, 2009). Functional annotation analyses, namely the KEGG pathway annotation
 148 analysis, KOG functional annotation analysis, and GO functional annotation analysis were
 149 performed. A total of 30,743 of the predicted genes were annotated into databases such as the NR
 150 (Supplementary Table S7). By comparison with GenBlastA v1.0.4(She *et al.*, 2009), homologous
 151 gene sequences were found in the genome with the true locus screened. GeneWise v2.4.1(Birney
 152 *et al.*, 2004) was used to find immature termination codons and frame-shift mutations in the gene
 153 sequences, and pseudogenes were identified. A total of 2,173 pseudogenes were predicted
 154 (Supplementary Table S8).

155 Table 2 Gene prediction result statistics

Method	Software	Species	Gene number
<i>Ab initio</i>	Genscan	-	26,696
	Augustus	-	38,539
	GlimmerHMM	-	48,103
	GeneID	-	39,104
	SNAP	-	44,716
Homology-based		<i>Oryza sativa</i>	26,741
		<i>Ziziphus jujuba</i>	27,261
	GeMoMa	<i>Arabidopsis thaliana</i>	28,297
		<i>Prunus persica</i>	30,248
		<i>Pyrus bretschneideri</i>	29,355
RNAseq	PASA	-	63,071
	GeneMarkS-T	-	54,579
	TransDecoder	-	86,897
Integration	EVM	-	31,730

156

157 **Hi-C assisted assembly**

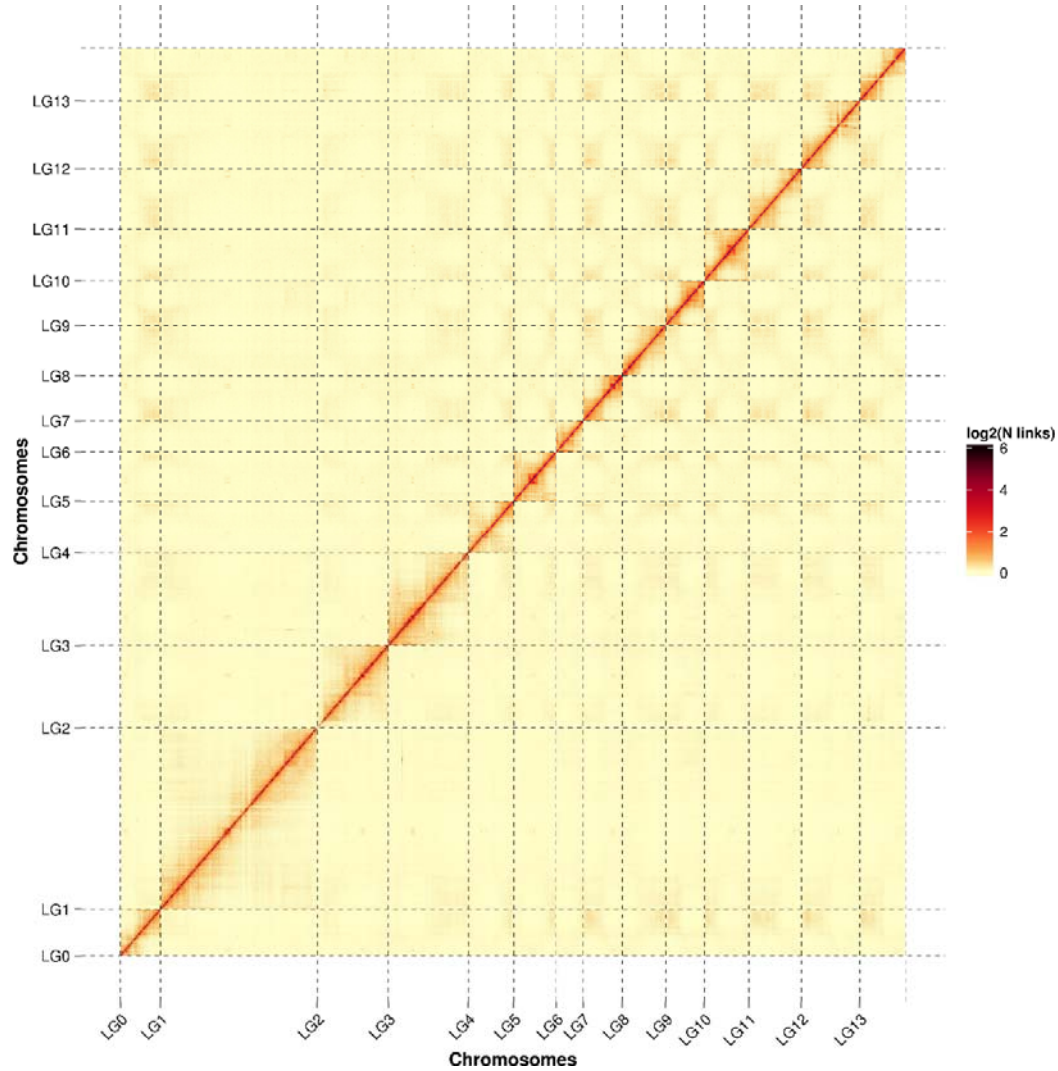
158 Based on Sequencing By Synthesis (SBS) technology, the Illumina high-throughput sequencing

159 platform was used to sequence the Hi-C library to produce a large number of high-quality reads.
160 Raw data for sequencing samples included two FASTQ files, including reads measured at both
161 ends of all Hi-C constructed library fragments (Supplementary Figure S5). We obtained 39.56 Gb
162 clean reads, with sequencing coverage of 75×, and Q30 ratio of > 95.46% (Supplementary Table
163 S9).

164 BWA(Li *et al.*, 2009) and SAMtools (version: 0.7.10-r789) were used to map the pair-end
165 data with the assembled genome sequence. The ratio of reads mapped to the assembled genome
166 was 90.68%, and the ratio of Unique Mapped Read Pairs was 61.13%, indicating that the Hi-C
167 data were good enough for subsequent analysis (Supplementary Table S10). We used
168 HiC-Pro(Servant *et al.*, 2015) to filter and evaluate the Hi-C data. The Invalid Interaction Pairs
169 ratio cannot exceed 80% if it is a usable Hi-C library(Belton *et al.*, 2012). Invalid Interaction Pairs
170 mainly include Self-circle Ligation, Dangling Ends type, Re-ligation type, and other discarded
171 types(Belton *et al.*, 2012; Hu *et al.*, 2013; Imakaev *et al.*, 2012; Lajoie *et al.*, 2015; Servant *et al.*,
172 2015). A total of 80.79 M pairs of reads on the genome were obtained in this experimental library.
173 Among them, 72.97 M pairs were valid Hi-C data, accounting for 90.32% of the data on the
174 genome, and the ratio of Invalid Interaction Pairs was 9.68% (Supplementary Table S11).

175 After Hi-C assembly, a total of 51.71 Mb of genomic sequence was mapped to the
176 chromosome, accounting for 96.94% of the total length of the sequence, and the corresponding
177 number of sequences was 269, accounting for 45.83% of the total number of sequences. Among
178 the sequences located on the chromosome, the sequence length that could determine the order and
179 direction was 473.91 Mb, accounting for 92.8% of the total length of the sequence located on the
180 chromosome, and the number of corresponding sequences was 104, accounting for 38.66% of the
181 total number of sequences located on the chromosome (Supplementary Table S12).

182 For Hi-C assembled into the genome of the chromosome, the length was cut into a bin of 100
183 Kb, and then the number of Hi-C Read Pairs was covered between any two bins as the intensity
184 signal of the interaction between the two Bins (Fig 2). A total of 14 chromosome groups could be
185 clearly distinguished; within each group, it could be seen that the intensity of the interaction at the
186 diagonal position was higher than that of the non-diagonal position, indicating that the interaction
187 strength between adjacent sequences (diagonal position) in the result of Hi-C chromosome
188 assembly was high, while that between non-adjacent sequences (non-diagonal position) was weak,
189 which was consistent with the principle of Hi-C assisted genome assembly and proved that the
190 genome assembly had a good effect.



191

192

193

Fig 2 Hi-C assembly chromosome interaction heat map

194 Conclusion

195 In this study, the genome of *Elaeagnus angustifolia* L. was obtained using PacBio technology, and
196 Hi-C assisted assembly technology. Thus, our findings are a valuable source for comparative
197 genomic analyses of the *Elaeagnaceae* and can help understand the response mechanism of the
198 *Elaeagnaceae* to drought, salt, cold and wind resistance, thereby providing an effective theoretical
199 basis for planting *E. angustifolia* to reverse global land desertification.

200

201 Conflict of interest

202 The authors have no conflict of interest to declare.

203

204 Acknowledgements

205 We would like to thank Editage (www.editage.com) for providing linguistic assistance during the
206 preparation of this manuscript. The research was supported by the National Science and
207 Technology Support Program, China(NO1: 2014BAD16B02), the Shandong Key Research and

208 Development Program, China(NO2: 2018GNC113019), the Fruit innovation team in Shandong
209 Province, China(NO3: SDAIT-06-07), and the fund of National Modern Agro-industry
210 Technology Research System of China(NO4: CARS-28).

211

212 **Author contributions**

213 Yunfei Mao, Xinxing Cui, Yanli Hu and Xiang Shen planned and designed the research. Yunfei
214 Mao, Qin Hu, Manman Zhang, Lu Yang, Lulu Zhang, Yunyun Wang, Yijun Yin, Huiling Pang,
215 Yeping Liu, Xiafei Su and Song Li performed experiments, conducted fieldwork, analysed data etc.
216 Yunfei Mao, Fengwang Ma, Naibin Duan, Donglin Zhang, Yanli Hu, Zhiquan Mao, Xuesen Chen
217 and Xiang Shen wrote the manuscript. Every author contributed equally.

218

219 **Reference**

- 220 Belton, J.M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y., Dekker, J. (2012) Hi-C: a
221 comprehensive technique to capture the conformation of genomes. *Methods*. **58**, 268-276.
- 222 Birney, E., Clamp, M., Durbin, R. (2004) GeneWise and genomewise. *Genome research*. **14**,
223 988-995.
- 224 Birney, E., Clamp, M., Durbin, R. (2004) GeneWise and genomewise. *Genome research*. **14**,
225 988-995.
- 226 Blanco, E., Parra, G., Guigó, R. (2007) Using geneid to identify genes. *Current protocols in*
227 *bioinformatics*. <https://doi.org/10.1002/0471250953.bi0403s18>.
- 228 Burge, C., Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA.
229 *Journal of molecular biology*. **268**, 78-94.
- 230 Campbell, M.A., Haas, B.J., Hamilton, J.P., Mount, S.M., Buell, C.R. (2006) Comprehensive
231 analysis of alternative splicing in rice and comparative analyses with Arabidopsis. *BMC*
232 *genomics*. **7**, 327.
- 233 Edgar, R.C., Myers, E.W. (2005) PILER: identification and classification of genomic repeats.
234 *Bioinformatics*. **21**, i152-i158.
- 235 Gao, F., Wang, X., Li, X.M., Xu, M.Y., Li, H.Y., Abla, M., Sun, H.G., Wei, S.J., Feng, J.C., Zhou,
236 Y.J. (2018) Long-read sequencing and de novo genome assembly of *Ammopiptanthus nanus*,
237 a desert shrub. *GigaScience*. **7**, 1-5.
- 238 Griffiths-Jones, S., Grocock, R.J., Van Dongen, S., Bateman, A., Enright, A.J. (2006) miRBase:
239 microRNA sequences, targets and gene nomenclature. *Nucleic acids research*. **34**,
240 D140-D144.
- 241 Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., Bateman, A. Rfam:
242 annotating non-coding RNAs in complete genomes. *Nucleic acids research*. **33**, D121-D124.
- 243 Guo, L.J., Wang, Y.T. (2008) Conservation Research and Prospects of *Elaeagnus* Germplasm
244 Resources and Utilization Values. *Chinese Wild Plant Resources*. **27**, 32-34.
- 245 Haas, B.J., Papanicolaou, A. (2016) TransDecoder (Find Coding Regions Within Transcripts).
246 <http://transdecoder.github.io>.
- 247 Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R.,
248 Wortman, J.R. (2008) Automated eukaryotic gene structure annotation using
249 EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biol*. **9**, R7.
- 250 Han, Y., Wessler, S.R. (2010) MITE-Hunter: a program for discovering miniature inverted-repeat
251 transposable elements from genomic sequences. *Nucleic acids research*. gkq862.

- 252 He, Y.H. (2012) Bioinformatic Analysis of the Elaeagnaceae nrDNA ITS Sequences. *Northwest*
253 *Normal University, China*.
- 254 Hu, M., Deng, K., Qin, Z.H., Liu, J.S. (2013) Understanding spatial organizations of
255 chromosomes via statistical analysis of Hi-C data. *Quantitative Biology*. **1**, 156-174.
- 256 Huang, J.H., Maimaitijiang, Yang, C.H., Wang C.F. (2005) Present Situation and Prospect about
257 the Study of *Elaeagnus angustifolia* L.. *Chinese Wild Plant Resources*. **24**, 26-29, 33.
- 258 Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker,
259 J., Mirny, L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome
260 organization. *Nat Methods*. **9**, 999-1003.
- 261 Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J. (2005)
262 Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome*
263 *research*. **110**, 462-467.
- 264 Keilwagen, J., Wenk, M., Erickson, J.L., Schattat, M.H., Jan, G., Frank, H. (2016) Using intron
265 position conservation for homology-based gene prediction. *Nucleic acids research*. **44**,
266 e89-e89.
- 267 Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., Phillippy, A.M. (2017) Canu:
268 scalable and accurate long-read assembly via adaptivemer weighting and repeat separation.
269 *Genome Res*. **27**, 722-736.
- 270 Korf, I. (2004) Gene finding in novel genomes. *BMC bioinformatics*. **5**, 59.
- 271 Lajoie, B.R., Dekker, J., Kaplan, N. (2015) The Hitchhiker's guide to Hi-C analysis: practical
272 guidelines. *Methods*. **72**, 65-75.
- 273 Li, H., Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform.
274 *Bioinformatics*. **25**, 1754-1760.
- 275 Lin, M., Todoric, D., Mallory, M., Luo B.S., Trottier, E., Dan, H.H. (2006) Monoclonal antibodies
276 binding to the cell surface of *Listeria monocytogenes* serotype 4b. *Journal of Medical*
277 *Microbiology*. **55**, 291-299.
- 278 Liu, Y.W., Di, D.L., Wang, Q. (2003) Study on Chemical Components and Fingerprint of Volatile
279 Oil from Flowers of *Elaeagnus angustifolia* L. *Food Science*. **24**, 111-113.
- 280 Liu, Y.Z. (2015) Studies on Genetic Diversity of *Elaeagnus angustifolia* L.. *Hunan Agricultural*
281 *University, China*.
- 282 Lowe, T.M., Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA
283 genes in genomic sequence. *Nucleic acids research*. **25**, 0955-0964.
- 284 Majoros, W.H., Pertea, M., Salzberg, S.L. (2004) TigrScan and GlimmerHMM: two open source
285 ab initio eukaryotic gene-finders. *Bioinformatics*. **20**, 2878-2879.
- 286 Nawrocki, E.P., Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches.
287 *Bioinformatics*. **29**, 2933-2935.
- 288 Pertea, M., Kim, D., Pertea, G.M., Leek, J.T., Salzberg, S.L. (2016) Transcript-level expression
289 analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*. **11**,
290 1650.
- 291 Price, A.L., Jones, N.C. (2005) Pevzner PA: De novo identification of repeat families in large
292 genomes. *Bioinformatics*. **21**, i351-i358.
- 293 Salvati, L., Kosmas, C., Kairis, O., Karavitis, C., Acikalın, S., Belgacem, A., Solé-Benet, A.,
294 Chaker, M., Fassouli, V., Gokceoglu, C., Gungor, H., Hessel, R., Khatteli, H., Kounalaki, A.,
295 Laouina, A., Ocakoglu, F., Ouessar, M., Ritsema, C., Sghaier, M., Sonmez, H., Taamallah, H.,

- 296 Tezcan, L., de Vente, J., Kelly, C., Colantoni, A., Carlucci, M. (2016) Assessing the
297 effectiveness of sustainable land management policies for combating desertification: A data
298 mining approach. *Journal of Environmental Management*. **183**, 754-762.
- 299 Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E., Dekker, J.,
300 Barillot, E. (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing.
301 *Genome Biology*. **16**, 1-11.
- 302 She, R., Chu, J.S.C., Wang, K., Pei, J., Chen, N.S. (2009) GenBlastA: enabling BLAST to identify
303 homologous gene sequences. *Genome Research*. **19**, 143.
- 304 Stanke, M., Waack, S. (2003) Gene prediction with a hidden Markov model and a new intron
305 submodel. *Bioinformatics*. **19**, ii215-ii225.
- 306 Tang, S., Lomsadze, A., Borodovsky, M. (2015) Identification of protein coding regions in RNA
307 transcripts. *Nucleic Acids Research*. **43**, e78.
- 308 Tarailo-Graovac, M., Chen, N. (2009) Using RepeatMasker to identify repetitive elements in
309 genomic sequences. *Current Protocols in Bioinformatics*.
310 <https://doi.org/10.1002/0471250953.bi0410s25>.
- 311 Trapnell, C., Pachter, L., Salzberg, S.L. (2009) TopHat: discovering splice junctions with
312 RNA-Seq. *Bioinformatics*. **25**, 1105-1111.
- 313 Vitas, A.I., e Aguado, V. I.G.-J. (2004) Occurrence of *Listeria monocytogenes* in fresh and
314 processed foods in Navarra (Spain). *International Journal of Food Microbiology*. **90**, 349
315 -356.
- 316 Wang, B.S., Qu, H.Y., Ma, J., Sun, X.L., Wang D., Zheng Q.S., Xing D. (2014) Protective effects
317 of elaeagnus angustifolia leaf extract against myocardial ischemia/reperfusion injury in
318 isolated rat heart. *Journal of Chemistry*. **2014**, 1-6.
- 319 Wang, Y., Zhao, P., Wang, Y.L., Zhang Y. (2006) Nutritional composition of wild *Elaeagnus*
320 *angustifolia* fruits. *Journal of Gansu Agricultural University*. **41**, 130-132.
- 321 Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P.,
322 Morgante, M., Panaud, O. (2007) A unified classification system for eukaryotic transposable
323 elements. *Nature Reviews Genetics*. **8**, 973-982.
- 324 Xu, Z., Wang, H. (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR
325 retrotransposons. *Nucleic Acids Research*. **35**, W265-W268.
- 326 Zhang, J.S., Zhang X.T., Tang, H.B., Zhang Q., Hua, X.T., Ma, X.K., Zhu, F., Jones, T.,
327 X.G., Zhu, Bowers, J., Wai, C.M., Zheng, C.F., Shi, Y., Chen, S., Xu, X.M., Yue, J.J., Nelson,
328 D.R., Huang, L.X., Li, Z., Xu, H.M., Zhou, D., Wang, Y.J., Hu, W.C., Lin, J.S., Deng,
329 Y.J., Pandey, N., Mancini, M., Zerpa, D., Nguyen, J.K., Wang, L.M., Yu, L., Xin, Y.H., Ge,
330 L.F., Arro, J., Han, J.O., Chakrabarty, S., Pushko, M., Zhang, W.P., Ma, Y.H., Ma, P.P., Lv,
331 M.J., Chen, F.M., Zheng, G.Y. Xu, J.S., Yang, Z.H., Deng, F., Chen, X.Q., Liao, Z.Y., Zhang,
332 X.X., Lin, Z.C., Lin, H., Yan, H.S., Kuang, Z., Zhong, W.M., Liang, P.P., Wang, G.F., Yuan,
333 Y., Shi, J.X., Hou, J.X., Lin, J.X., Jin, J.J., Cao, P.J., Shen, Q.C., Jiang, Q., Zhou, P., Ma,
334 Y.Y., Zhang, X.D., Xu, R.R., Liu, J., Zhou, Y.M., Jia, H.F., Ma, Q., Qi, R., Zhang, Z.L., Fang,
335 J.P., Fang, H.K., Song, J.J., Wang, M.J., Dong, G.G., Wang, G., Chen, Z., Ma, T., Liu,
336 H., Dhungana, S.R., Huss, S.E., Yang, X.P., Sharma, A., Trujillo, J.H., Martinez,
337 M.C., Hudson, M., Riascos, J.J., Schuler, M., Chen, L.Q., Braun, D.M., Li, L., Yu,
338 Q.Y., Wang, J.P., Wang, K., Schatz, M.C., Heckerman, D., Sluys, M.-A.V., Souza,

339 G.M., Moore, P. H., Sankoff, D., Buren, R.V., Paterson, A.H., Nagai, C., Ming ^{□□}, R. (2018)
340 Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L.. *Nature*
341 *Genetics*. **50**, 1565–1573.

342

343 **Supporting information**

344 Additional Supporting information may be found in the online version of this article:

345 **Figure S1** DNA library components.

346 **Figure S2** Hi-C sequencing experiment process.

347 **Figure S3** Distribution of k-mers of length 19 from the Illumina Hiseq reads.

348 **Figure S4** The integrated gene is derived from the distribution map of three prediction methods.

349 **Figure S5** FASTQ file format.

350 **Table S1** Sample sequencing result statistics.

351 **Table S2** Length distribution of subreads of Pac-bio sequencing.

352 **Table S3** Filtering raw data of Pac-bio sequencing.

353 **Table S4** Genome assembly evaluation statistics.

354 **Table S5** Transcriptome comparison region statistics.

355 **Table S6** Gene information statistics.

356 **Table S7** Gene function annotation statistics.

357 **Table S8** Pseudogene Prediction Results.

358 **Table S9** Sequencing data volume statistics.

359 **Table S10** Clean data and genome alignment results statistics.

360 **Table S11** Hi-C sequencing data Validation.

361 **Table S12** Hi-C assembles data statistics.