

A peer-reviewed version of this preprint was published in PeerJ on 4 March 2020.

[View the peer-reviewed version](https://peerj.com/articles/8678) (peerj.com/articles/8678), which is the preferred citable publication unless you specifically need to cite this preprint.

Su Q, Liu L, Zhao M, Zhang C, Zhang D, Li Y, Li S. 2020. The complete chloroplast genomes of seventeen *Aegilops tauschii*: genome comparative analysis and phylogenetic inference. PeerJ 8:e8678
<https://doi.org/10.7717/peerj.8678>

The complete chloroplast genomes of seventeen *Aegilops tauschii*: Genome characteristic and comparative analysis

Qing Su ^{Equal first author, 1}, Luxian Liu ^{Equal first author, 1}, Mengyu Zhao ¹, Cancan Zhang ¹, Dale Zhang ¹, Youyong Li ^{Corresp., 2}, Suoping Li ^{Corresp. 1}

¹ Key Laboratory of Plant Stress Biology, School of Life Science, Henan University, Kaifeng, China

² School of Life Science and Technology, Henan Institute of Science and Technology, Xinxiang, China

Corresponding Authors: Youyong Li, Suoping Li
Email address: liyouyong@163.com, lisuoping@henu.edu.cn

As the diploid progenitor of common wheat, *Aegilops tauschii* Cosson (DD, $2n = 2x = 14$) is regarded to be a potential genetic resource for improving common wheat, which is naturally distributed in central Eurasia, spreading from northern Syria and Turkey to western China. In this work, the chloroplast genomes of seventeen *Ae. tauschii* accessions showed 135 551~ 136 009 bp in length and contained the typical quadripartite structure of angiosperms. Meanwhile, a total of 127 functional genes, including 78 protein-coding genes, 4 rRNAs, 26 tRNAs, and 19 duplicated genes were identified. Overall genomic structure including gene number, gene order were well conserved with identical IR/SC boundary regions, but few variations predominantly were detected in non-coding regions (intergenic spacer regions). IR expansion and contraction with identical structure among 17 *Aegilops tauschii* accessions were not influence chloroplast genomes in length. Four cpDNA markers including *rpl32-trnL-UAG*, *ccsA-ndhD*, *rbcL-psaI* and *rps18-rpl20* showed high nucleotide polymorphisms which may be used to study on inter- and intra-specific genetic structure and diversity of *Ae. tauschii*. The *ndhF* gene in AY46 accession appeared the highest ω value, which might be involved in the adaptation to high altitude ecological environment during the evolution of AY46 accession. The phylogenetic relationships constructed by the complete genome sequences strongly support that *Ae. tauschii* in the Yellow River region might be directly originated from Central Asia rather than Xinjiang. The specific spreading route of *Ae. tauschii* revealed in this work, reflects the frequent cultural exchange through the silk road from one point of view. We confirmed that *Ae. tauschii* derived from monophyletic speciation rather than hybrid speciation at the chloroplast genome level.

1 **The complete chloroplast genomes of seventeen *Aegilops tauschii*: Genome characteristic**
2 **and comparative analysis**

3 Qing Su¹⁺, Luxian Liu¹⁺, Mengyu Zhao¹, Cancan Zhang¹, Dale Zhang¹, Youyong Li^{2*}, Suoping Li^{1*}

4 ¹Key Laboratory of Plant Stress Biology, School of Life Science, Henan University, Kaifeng 475000, China

5 ²School of Life Science and Technology, Henan Institute of Science and Technology, Xinxiang 453003, China

6 ⁺The two authors contributed equally to this work

7 ^{*} Corresponding author

8 Email list for all authors:

9 Qing Su: suqingangel@163.com; Luxian Liu: liushuangcx2007@126.com; Mengyu Zhao:
10 835591597@qq.com; Cancan Zhang: 894275874@qq.com; Dale Zhang: zhangdale@henu.edu.cn; Youyong
11 Li: liyoyong@163.com; Suoping Li: lisuoping@henu.edu.cn.

12

13

14

15

16

17

18

19

20

21

22

23

24

25
26
27
28
29
30
31
32
33
34

35 **Abstract**

36 As the D genome progenitor of bread wheat, *Aegilops tauschii* Cosson (DD, $2n = 2x = 14$) is considered to be
37 a potential genetic resource for improving bread wheat, which is naturally distributed in central Eurasia,
38 ranging from northern Syria and Turkey to western China. In this work, the chloroplast genomes of seventeen
39 *Ae. tauschii* accessions showed 135 551~ 136 009 bp in length and contained the typical quadripartite structure
40 of angiosperms. Meanwhile, a total of 127 functional genes, including 78 protein-coding genes, 4 rRNAs, 26
41 tRNAs, and 19 duplicated genes were identified. Overall genomic structure containing gene order, gene
42 number were well conservative with identical IR/SC boundary regions, but few variations predominantly were
43 detected in non-coding regions (intergenic spacer regions). IR expansion and contraction with identical
44 structure among 17 *Aegilops tauschii* accessions were not influence chloroplast genomes in length. Four
45 cpDNA markers including *rpl32-trnL-UAG*, *ccsA-ndhD*, *rbcL-psaI* and *rps18-rpl20* showed high nucleotide
46 polymorphisms, which may be used to study on interspecific and intraspecific genetic structure and diversity
47 of *Ae. tauschii*. The *ndhF* gene in AY46 accession appeared the highest ω value, which might be involved in
48 the adaptation to high altitude ecological environment during the evolution of AY46 accession. The
49 phylogenetic relationships constructed by the complete genome sequences well sustained that *Ae. tauschii* in

50 the Yellow River region might be directly originated from Central Asia rather than Xinjiang. The specific
51 spreading route of *Ae. tauschii* revealed in this work, reflects the frequent cultural exchange through the silk
52 road from one point of view. We confirmed that *Ae. tauschii* derived from monophyletic speciation rather than
53 hybrid speciation at the chloroplast genome level.

54

55 **Keywords:** inverted repeats region, next-generation sequencing, phylogenetic tree, genetic
56 differentiation, selective pressure

57 **Introduction**

58 *Aegilops tauschii* Cosson (DD, $2n = 2x = 14$) is the donor of common wheat D genome contained abundant
59 genetic variation, especially strong tillering ability and high plant tolerance including disease resistance,
60 drought resistance and abiotic stress resistance (Singh et al., 2012). Naturally, it is widely distributed in central
61 Eurasia, spreading from northern Syria and Turkey to western China (Yili area of Xinjiang). In addition, as a
62 kind of farmland weed accompanying common wheat, *Ae. tauschii* is also found in the middle reaches of the
63 Yellow River (including Henan and Shannxi provinces, China) (Wei et al., 2008). In genetic studies, *Ae.*
64 *tauschii* is preferable to be further divided into into two sublineages based on nuclear genome sequences,
65 recognized as L1 and L2, broadly affiliating with *Ae. tauschii* ssp. *strangulata* and *Ae. tauschii* ssp. *tauschii*,
66 respectively (Mizuno et al., 2010; Dvorak et al., 1998). Previous studies have shown that L2 lineage is
67 involved in the origin of common wheat, which is limited in a narrow area within the whole species
68 distribution (Wang et al., 2013; Dvorak et al., 2012). Relatively, L1 lineage could adapt to more diversified
69 ecological environment conditions (Dudnikov, 2012). It is proposed that the genetic differentiation types of *Ae.*
70 *tauschii* (mainly L1 lineage) are more enrich than that of the common wheat D genome with the long genetic
71 distance between L1 and L2 (Lubbers et al., 1991; Wang et al., 2013; Dvorak et al., 2012). Thus, as many wild
72 crop progenitors, *Ae. tauschii* is considered to be a valuable gene germplasm for genetic improvement research
73 of common wheat (Kilian et al., 2011).

74 Iran is widely regarded as the center of the origin and genetic diversity for *Ae. tauschii* (Dvorak et al.,

75 1998). After a long period of dispersal and adaptation, the spatial distribution of *Ae. tauschii* shows distinctly
76 difference. L2 lineage is mainly restricted from Transcaucasia (Armenia and Azerbaijan) to eastern Caspian
77 Iran, whereas L1 lineage widely spreads across the whole species range, including the middle reaches of the
78 Yellow River and the Yili area of Xinjiang in China (Kihara et al., 1965; Jaaska 1980; Matsuoka et al., 2015;
79 Wei et al., 2008). Owing to its prominent advantages such as moderate nucleotide replacement rate, significant
80 variations in molecular evolution speed between non-coding and coding regions, moderate genome in size and
81 the desirable collinear properties among different species (Kimura et al., 1984), chloroplast sequence has been
82 considered to be an effective strategy in exploring intra- and interspecific evolutionary relationships as well as
83 comparative genomic studies (Matsuoka et al., 2005; Yamane and Kawahara 2005; Tabidze et al., 2014;
84 George et al., 2015; Liu et al., 2017). In angiosperms, the size of chloroplast (cp) genome and its gene
85 arrangement are highly conserved with a circular chromosome ranging from 120 to 160 kb, including a small
86 single-copy region (SSC), a large single-copy region (LSC) and a couple of inverted repeats region (IRs)
87 (Palmer, 1991; Yang et al., 2010). Particularly, the phylogenetic analysis of chloroplast sequences could
88 provide specific identification on maternal lineages because the chloroplasts are primarily non-recombining
89 and uniparentally inherited (Sang, 2002). Up to now, genetic variation and phylogenetic analysis in common
90 wheat and its relatives have been researched based on the complete chloroplast genomes (Middleton et al.,
91 2014; Gornicki et al., 2014; Gogniashvili et al., 2016). To shed light on the genetic variation of *Ae. tauschii*
92 and the origin of Chinese landraces, here we report a sequence analysis of seventeen complete chloroplast
93 genomes of *Ae. tauschii* from western Turkey to eastern China. The results not only supplements new
94 molecular phylogenetic information for taxonomy optimization of *Ae. tauschii*, but also provides promising
95 germplasm resources for the genetic improvement of bread wheat.

96 **Materials and methods**

97 **Plant materials**

98 *Ae. tauschii* accessions included in this study, containing origin country and collection region, were listed in
99 Table 1. The accessions marked as ‘XJ’, ‘T’, and ‘S’ represent those from Xinjiang, Henan, and Shannxi,
100 respectively. The other accessions named ‘AY’ and ‘AS’ were respectively provided by the US National Plant
101 Germplasm Center and Institute of Genetics and Developmental Biology, Chinese Academy of Sciences. The
102 accessions locate in a rather wide range from Turkey, Georgia, Iran, Turkmenistan, Kazakhstan, Tajikistan,

103 Afghanistan, Pakistan, India to Xinjiang, Shannxi and Henan provinces in China (Figure 1), including 15
104 accessions for L1 lineages and 2 accessions for L2 lineages based on single nucleotide polymorphisms (Wang
105 et al., 2013).

106 **Next-generation sequencing of chloroplast genomes and annotation**

107 Total genomic DNA of 17 *Ae. tauschii* accessions were extracted from fresh leaves of seedling on one-week
108 old by plant genomic DNA kit (TIANGEN Biotech (Beijing) Co., Ltd), respectively. Approximately 5 ~ 10 µg
109 of DNA was trimmed to generate fragments. Then, the quality of DNA sequences was examined by Agilent
110 Bioanalyzer 2100 (Agilent Technologies). A paired-end sequencing library was constructed using ~400 bp
111 fragment by Genomic DNA Sample Prep Kit (Illumina) in accordance with the manufacturer's protocol. Then,
112 the genome sequencing was carried out on Majorbio Bio-Pharm Technology Co., Ltd. (Shanghai, China) using
113 the HiSeq X ten sequencing platform (Illumina Inc., San Diego, CA) with 150 bp read length. Low quality
114 reads with phred score < 30 and 0.001 probability error were sheared utilizing Trimmomatic v0.36
115 (<http://www.usadellab.org/cms/index.php?page=trimmomatic>), and the remaining high quality fragments were
116 assembled into contigs employing the SOAPdenovo v2.21 (<http://soap.genomics.org.cn/soapdenovo.html>). The
117 assembled contigs were further aligned to the reference genomes of AL8/78 (GenBank No. KJ 614412) using
118 BLASTN (<http://www.ncbi.nlm.nih.gov>). Finally, the sequence gaps of genome were filled by GapCloser
119 v1.12 (<http://soap.genomics.org.cn/soapdenovo.html>).

120 The complete chloroplast genome of *Ae. tauschii* was annotated using Dual Organellar GenoMe
121 Annotator program (DOGMA, Wyman et al., 2004), which were further verified artificially. The annotation of
122 the tRNA genes was verified using tRNAscan-SE. The circular chloroplast genome of *Ae. tauschii* was
123 constructed using the online software OGDRAW (Lohse et al., 2013).

124 **Comparative chloroplast genomic analysis**

125 In order to study intraspecific variations, LSC, SSC and expansion and interspecific variations of inverted
126 repeats region (IR) were conducted by Geneious 9.0.5 software (Biomatters, Auckland, New Zealand).
127 Utilizing AL8/78 as a reference genome, the chloroplast genomes of T093 and AY81 were aligned by
128 mVISTA to visualize sequence variations between L1 and L2 (Gogniashvili et al., 2016).

129 **Analysis of sequences divergence and molecular markers**

130 The divergence analysis of chloroplast genome was performed based on homologous regions of 17 *Ae. tauschii*
131 accessions employing MAFFT alignment (Kato et al., 2002). Polymorphic sequences within complete
132 chloroplast genomes were selected based on nucleotide diversity (Pi) analysis utilizing DnaSP v5.0 software
133 (Librado and Rozas, 2009).

134 **Synonymous (K_S) and Non-synonymous (K_A) substitution rates analysis**

135 To compute synonymous and non-synonymous substitution rates, the uniform encoding exons from individual
136 cp genomes were extracted and aligned separately with Chinese spring TA3008 (GenBank No. KJ614396) as
137 the reference genome utilizing Geneious v9.0.5 software (Biomatters, Auckland, New Zealand). The relative
138 evolutionary rates were calculated based on synonymous and non-synonymous substitutions as well as their
139 ratios ($\omega = K_A/K_S$).

140 **Phylogenetic analysis**

141 The phylogenetic relationships of *Ae. tauschii* were conducted using the complete chloroplast genome
142 sequences of seventeen accessions. In addition, in order to verified the origin of *Ae. tauschii* within Triticeae,
143 ninety-nine accession were used for phylogenetic analysis, including seventeen assembled from in this study,
144 representatives of fifty-six species of Aegilops genus, twenty-four of Triticum genus with two species of
145 *Hordeum vulgare* as outgroups.

146 These chloroplast genome sequences were aligned using Geneious 9.0.5 software (Biomatters, Auckland,
147 New Zealand). Gaps were adjusted manually or removed. The alignment length with all gap positions removed
148 were determined to be 135 984 bp (17 accessions) and 131 116 bp (99 accessions). The two phylogenetic trees
149 was constructed based on the maximum-likelihood (ML) method utilizing MEGA7.0 software (Kumar et al.,
150 2016). The bootstrap values were evaluated with 1000 replications. The best-fit nucleotide substitution model
151 (GTR+I+G) was selected by jModelTest 2.1.4 software (Posada, 2008) with default parameters.

152 **Results**

153 **The characteristic and comparison of *Ae. tauschii* chloroplast genomes**

154 In this work, seventeen *Ae. tauschii* chloroplast genomes were similar in size and gene content. They are
155 generally 135 551~ 136 009 bp long and contain four blocks: two IR regions separated by a SSC region (12
156 773 ~ 12 826 bp) and a LSC region (79 723 ~ 80 140 bp) (Figure2 and Table 2). The cp genome sequences
157 were submitted into GenBank with the accession number in Table 1. The intergenic regions of these sequences
158 range within 75 635 ~ 77 699 bp in length, accounting for the total genomes of 53.2 ~ 57.2 %. The remaining
159 gene regions are composed of coding regions, intron regions, tRNA genes, and rRNA genes. Particularly, GC
160 contents are relatively stable in 17 chloroplast genomes, ranging from 38.31 % to 38.35 % (Table 2).

161 The compositions of seventeen *Ae. tauschii* chloroplast genomes are highly conservative which all

162 contain a total of 122 functional genes, containing 82 protein-coding genes, 8 rRNAs, 32 tRNAs, with 16
163 duplicated genes in each one accession of *Ae. tauschii* (Table 2). Among the 106 unique genes, in which 56
164 fragments are related to self-replication and 46 genes are associated with photosynthesis. In addition, the
165 functions of 4 other genes are also annotated: maturase (*matK*), envelope membrane protein (*cemA*), C-type
166 cytochrome synthesis (*ccsA*) as well as protease (*clpP*) (Table S1).

167 The representative genomes of three *Ae. tauschii* accessions were relatively conservative, and any
168 translocations or inversions compared to any of the genomes were not identified. IR regions had lower
169 sequence variation than that in the LSC and SSC regions (Figure S1). Further, the exact IR/SSC and IR/LSC
170 boundary location and their neighbor genes among the 17 *Ae. tauschii* chloroplast genomes were strictly
171 identical in length (21 548 bp) and border structure (Figure S2).

172

173 **Sequence divergences and molecular marker development of *Ae. tauschii* chloroplast genomes**

174 The sequence divergences of coding genes, intron regions and intergenic regions, and were further
175 analyzed to elucidate the variation characteristics among 17 *Ae. tauschii* chloroplast genomes. Totally 56
176 variation loci were detected, in which 38 loci were located in non-coding regions (34 intergenic regions and 4
177 intron regions) and the other 18 loci were found in coding genes. As a result, relatively high value of
178 nucleotide variability (P_i) was determined for non-coding regions, ranging from 0.00008 to 0.00635 with an
179 average of 0.00133, which was approximately 3 times than that in the coding regions (averaged 0.000432)
180 (Figure 3). Specifically, four of these variable loci, *rpl32-trnL-UAG* (0.00478), *ccsA-ndhD* (0.00483), *rbcL-*
181 *psaI* (0.00492), and *rps18-rpl20* (0.00635) locating in intergenic regions displayed more higher nucleotide
182 polymorphisms (Figure 3), of which the former two loci were in the SSC region and the latter two loci were
183 located in the LSC region.

184

185 **Synonymous and non-synonymous substitution rates analysis**

186 To measure selection pressures of cp genes, the K_s , K_A and ω values of 78 protein-coding genes were
187 calculated and compared among 17 *Ae. tauschii* chloroplast genomes with Chinese Spring TA3008 as a
188 reference genome. Only 9 protein-coding genes exhibited ω values, with K_A and K_s values ranging from 0.004
189 \sim 0.006 and 0.0027 \sim 0.0069 (Table S2), respectively, which was neglected for the remaining genes due to K_s
190 or $K_A = 0$. Synonymous and non-synonymous substitution rates among intraspecific accessions of *Ae. tauschii*
191 were conservative (Figure 4). Specially, *ndhF* in AY46 displayed the largest evolutionary rate ($\omega = 1.4605$). In
192 addition, we found that only *atpB* gene are under positive selection in *Ae. tauschii* based on the K_A/K_s values.

193

194 **Phylogenetic tree**

195 The two datasets of the seventeen complete cp genome sequences assembled in this study and ninety-nine the
196 chloroplast genome sequence were used to construct the phylogenetic relationships. As shown in Figure 5,
197 seventeen *Ae. tauschii* are obviously clustered into three groups. Group I are sister to the remaining accessions,
198 group II and group III, which form one clade. Interestingly, group I is found derived from the L2 lineage,
199 while groups II and III are all originated from the L1 lineage, in accordance with the observation using 7185
200 SNP markers (Wang et al., 2013). This indicates that complete chloroplast genome sequences were equal
201 effective to nuclear sequences for the phylogenetic construction of *Ae. tauschii*. In group III, SC1 and T093
202 from the Yellow River region display a close genetic relationship, clustering into a small branch with AY22
203 from Pakistan. In addition, *Ae. tauschii* in the Yellow River region shows a larger distance with landraces of
204 Xinjiang compared with AY22. The above results imply that the chloroplast genome of *Ae. tauschii* derived
205 from the Yellow River region (Henan and Shaanxi) has closer relationship with that from Central Asia, which
206 exhibits relatively larger genetic differentiation with that from Xinjiang.

207 Ninety-nine accessions were used to constructed phylogenomic tree of Poaceae faminly. The two genetic
208 clusters were retrieved, one was Triticum cluster, another cluster was refered to the Aegilops species (Figure 6).
209 *Ae. speltoides* are gathered with a polytomy together with almost all *Triticum* species comprised the *Triticum*
210 clulster. *Triticum urartu* and *Triticum monococum* combined with the remaining Aegilops species formed the
211 Aegilops cluster. All D-genome species including *Ae. tauschii*, *Ae. ventricosa* and *Ae. cylindrica* clustered into

212 one clade.

213 **Discussion**

214 In this study, all the seventeen cp genomes displayed the typical structure of angiosperms, and identically
215 harbored 106 unigenes arrayed in the same order (Figure 2). The chloroplast genomes of seventeen *Ae.*
216 *tauschii* accessions were relatively conservative, and the IR region were more conservative than the LSC or
217 SSC regions (Figure S1). Intraspecific differences among cp genome of *Ae. tauschii* mainly contributed to
218 genome size, rather than expansion and contraction of IR regions. The seventeen cp genome present similar
219 size, with ranging from 135 551 bp to 136 009 bp (Table 2), of which AY22 exhibited the biggest cp genome,
220 and AY46 had the smallest one. IRs regions of all observed cp genomes had identical length (21548 bp). We
221 detected the differences in length is mainly attributed to to the variation in the non-coding regions, especially
222 in intergenic regions size (Table 2). Compared to the reference genome, the seventeen analyzed genome had
223 no any gene loss. Inconsistent with interspecific statement (Terakami et al., 2012), in this study, identical
224 LSC/IR and SSC/IR border regions among *Ae. tauschii* accessions is not suitable for a useful evolutionary tool
225 at intraspecific level. Thus, general cp genome structure including gene order, gene number were well
226 conservative with identical IR/SC boundary regions, but few variations predominantly were detected in non-
227 coding regions (intergenic spacer regions).

228 To identify the genetic divergence, the nucleotide variability (Pi) of coding genes, intron regions and
229 intergenic regions from seventeen *Ae. tauschii* individuals were performed using DnaSP. The results showed
230 IRs appeared the lower sequence divergence than the LSC and SSC regions, which also occurred in other
231 angiosperm and possibly on account of copy correction of IR regions by gene conversion (Khakhlova and
232 Bock, 2006). The four non-coding regions contained *rpl32-trnL-UAG*, *ccsA-ndhD*, *rbcL-psaI* and *rps18-rpl20*
233 showed high nucleotide polymorphisms. Two of them (*rpl32-trnL-UAG*, *ccsA-ndhD*) were in the SSC region,
234 whereas the other two loci (*rbcL-psaI* , *rps18-rpl20*) were in the LSC region (Figure 3). Previous studies
235 identified *Ae. tauschii* accessions using chloroplast non-coding sequences *trnF-ndhJ*, *trnC-rpoB*, *atpI-atpH*
236 and *ndhF-rpl32* (Yamane and Kawahara, 2005; Dudnikov, 2012). In this study, we developed additional four
237 regions (*rpl32-trnL-UAG*, *ccsA-ndhD*, *rbcL-psaI* and *rps18-rpl20*) with relatively high levels of intraspecific

238 variation, which can be used for population genetic analyses or serve as specific DNA barcodes of *Ae.*
239 *tauschii* (Zhang et al., 2011; Maier et al., 1995).

240 To pinpoint whether genes of intraspecific species underwent adaptive evolution in *Ae. tauschii*
241 plastomes, we carried out the identification of substitution rates for individual genes. Comparing with the
242 outgroup *Triticum aestivum*, the values (ω) showed few differences among gene individuals (Fig.4). Identical
243 K_A/K_S value of *atpB* among seventeen cp genomes are more than 1, indicating *atpB* are under positive
244 selection in *Ae. tauschii*. The remaining genes manifested purifying selection on the cp encoding genes of *Ae.*
245 *tauschii* due to less than one of K_A/K_S values. Specially, *ndhF* gene in AY46 accession appeared the highest ω
246 value. The *ndhF* gene encoding the F subunit of NADH dehydrogenase is involved in photosystem I and
247 electron transport in chloroplast thylakoid membranes (Lascano et al., 2003). Previous research indicated that
248 the evolutionary rates of *ndhF* were positively correlated with altitude (Li et al., 2016). In this study, we found
249 that the highest ω value was 1.4605 for *ndhF* gene, far surpassing one in AY46 accession compared to other 16
250 *Ae. tauschii* accessions. *ndhF* gene had been identified as the most differentiation one in plastome of vascular
251 plant (Kim et al., 2004). Moreover, AY46 accession showed the largest evolutionary pressure under 1296
252 meters in altitude, implying that *ndhF* gene may be involved in the adaptation to the high altitude environment
253 during the evolutionary process of AY46 accession.

254 A substantial sequence length is believed to be required in a robust phylogenetic analysis due to the
255 lower nucleotide substitution rates in chloroplast genomes than those in nuclear genomes (Wolfe et al., 1987;
256 Khakhlova and Bock, 2006). Therefore, the complete chloroplast genome sequence is beneficial for exploring
257 the phylogenetic relationship of angiosperms to acquire more comprehensive information (Kim et al., 2015).
258 As for the origin of *Ae. tauschii* in China, Yili area of Xinjiang is undoubtedly considered to be the
259 easternmost of the natural distribution for *Ae. tauschii* wild population (Matsuoka et al., 2015; Gogniashvili et
260 al., 2016; Wang et al., 2013), whereas the origin of landraces in the Yellow River region is still controversial.
261 The result in this work provides an evidence that *Ae. tauschii* in the Yellow River region might be directly
262 originated from Central Asia. Analogously, Iran is believed to be the origin of *Ae. tauschii* in the Yellow River
263 region according to the results of Wei et al., (2008). Based on the above recognition, the possible introduction
264 route for *Ae. tauschii* dispersing eastward to China is proposed. Specifically, *Ae. tauschii* accessions

265 distributed in Iran were eastward spread to Central Asia and Yili of Xiinjiang through human activities or
266 natural extension, followed by further eastward spreading to the Yellow River region from the former area
267 through the south line of silk road, which gradually evolved to *Ae. tauschii* accessions of the Yellow River
268 region.

269 In order to further illuminate D genomes origin of common wheat, ninety chloroplast genomes were
270 constructed, with the similar topological structure to Bernhardt (2017). Most researchers consider that *Ae.*
271 *tauschii*, the donator of the D genome of common wheat, derived from monophyletic speciation (Dvorak et al.,
272 1998, 2012; Wang et al., 2013; Matsuoka et al., 2015). However, another study estimate the evolution
273 relationship of the A, B and D genome lineages based on the genome sequences (2269 genes) of hexaploid
274 bread wheat subgenomes and five diploid relatives. The results suggest that the D genome are produced by
275 homoploid hybrid speciation of the A and B genomes (Marcussen et al., 2014). Li (2018) also pointed out that
276 a maternal origin of the D genome lineage might be the A-genome or some other relatively close lineage
277 through ancient hybridization based on phylogenetic analysis of chloroplast DNAs. In this study, *Ae. tauschii*
278 together with the other *Ae.* species cluster into a small branch and then sister to *Triticum urartu* clade,
279 excluding that D genome is derived from ancient hybridization and A genome acts as maternal parent. While,
280 there is an evidence for D genome is not derived from A crossing with B (B genome acting as maternal parent),
281 *Ae. tauschii* has closer genetic distance between *Triticum urartu* than to *Ae. speltoides*. We propose that these
282 results presented phylogeny estimates as a reference framework for future studies on *Triticeae* or *Ae. tauschii*.

283 **Conclusion**

284 In this work, genomic structure in gene order, gene number were well conservative with identical IR/SC
285 boundary regions among assembled the seventeen *Ae. tauschii* chloroplast genomes determined using Illumina
286 next-generation DNA sequencing technology. Intraspecific differences dominantly were detected in non-
287 coding regions (intergenic spacer regions) among cp genome of *Ae. tauschii* mainly contributed to genome size.
288 Four cpDNA markers including *rpl32-trnL-UAG*, *ccsA-ndhD*, *rbcL-psaI* and *rps18-rpl20* can be used to study
289 on inter- and intraspecific genetic structure and diversity of *Ae. tauschii*. The phylogenetic relationships
290 performed by the complete genome sequences well sustained that *Ae. tauschii* in the Yellow River region

291 might be directly originated from Central Asia rather than Xinjiang. We confirmed that *Ae. Tauschii* derived
292 from monophyletic speciation rather than hybrid speciation.

293 **Acknowledgements**

294 This work was supported by National Natural Science Foundation of China (Grant Nos. 31401379 and
295 31571649) and Project of Major Science and Technology in Henan Province (Grant No. 161100110400).

296 **References**

297 Bernhardt N, Brassac J, Kilian B, and Blattner FR. 2017. Dated tribe-wide whole chloroplast genome
298 phylogeny indicates recurrent hybridizations within Triticeae. *BMC Evolutionary Biology* 17(1):141.

299 Dudnikov AJ. 2012. Chloroplast DNA non-coding sequences variation in *Aegilops tauschii* Coss.:
300 evolutionary history of the species. *Genetic Resources and Crop Evolution* 59:683-699.

301 Dvorak J, Deal KR, Luo MC, You FM, Borstel KV, and Dehghani H. 2012. The origin of spelt and free-
302 threshing hexaploid wheat. *Journal of Heredity* 103:426-441.

303 Dvorak J, Luo MC, Yang ZL, and Zhang HB. 1998. The structure of the *Ae. tauschii* genepool and the
304 evolution of hexaploid wheat. *Theoretical and Applied Genetics* 97:657-670.

305 George B, Bhatt BS, Awasthi M, George B, and Singh AK. 2015. Comparative analysis of microsatellites in
306 chloroplast genomes of lower and higher plants. *Current Genetics* 61:665-677.

307 Gogniashvili M, Jinjikhadze T, Maisaia I, Akhalkatsi M, Kotorashvili A, and Kotaria N. 2016. Complete
308 chloroplast genomes of *Aegilops tauschii* Coss. and *Aegilops cylindrica* host sheds light on plasmon D
309 evolution. *Current Genetics* 62:791-798.

310 Gornicki P, Zhu H, Wang J, Challa GS, Zhang Z, Gill BS, and Li W. 2014. The chloroplast view of the
311 evolution of polyploid wheat. *New Phytologist* 204:704-714.

312 Jaaska V. 1980. Electrophoretic survey of seedling esterases in wheats in relation to their phylogeny.
313 *Theoretical and Applied Genetics* 56:273-284.

314 Katoh K, Misawa K, Kuma KI, and Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence
315 alignment based on fast fourier transform. *Nucleic acids research* 30:3059-3066.

316 Khakhlova O, and Bock R. 2006. Elimination of deleterious mutations in plastid genomes by gene conversion.
317 *Plant Journal* 46:85-94.

318 Kihara H, Yamashita H, and Tanaka M. 1965. Morphologic, physiological, genetical, and cytological studies
319 in *Ae.* and *Triticum* collected in Pakistan, Afghanistan, Iran; Kyoto University Press: Kyoto, 4-41.

- 320 Kilian B, Mammen K, Millet E, Sharma R, Graner A, Salamini F, Hammer K, and Ozkan H. 2011. Wild crop
321 relatives: genomic and breeding resources: Cereals. *Springer* 1-76.
- 322 Kim KJ, and Lee HL. 2004. Complete chloroplast genome sequences from Korean ginseng (*Panaxschinseng*
323 *Nees*) and comparative analysis of sequence evolution among 17 vascular plants. *DNA Research* 2004
324 11:247-261.
- 325 Kim K, Lee SC, Lee J, Yu Y, Yang K, and Choi BS. 2015. Complete chloroplast and ribosomal sequences for
326 30 accessions elucidate evolution of *Oryza* AA genome species. *Scientific Reports* 5:15655.
- 327 Kimura M. 1984. The neutral theory of molecular evolution. *Current Opinion in Genetics and Development*
328 170:224.
- 329 Kumar S, Stecher G, and Tamura K. 2016. MEGA7: Molecular evolutionary genetics analysis version 7.0 for
330 bigger datasets. *Molecular Biology and Evolution* 33:1870-1874.
- 331 Lascano HR, Casano LM, Martin M, and Sabater B. 2003. The activity of the chloroplast Ndh complex is
332 regulated by phosphorylation of the NDH-F subunit. *Plant Physiology* 13: 2256-2262.
- 333 Li CP, Sun XH, Conover JL., Zhang ZB, Wang JB, Wang XF, Deng X, Wang HY, Liu B, Wendel JF and
334 Gong L. 2018. Cytonuclear Coevolution following Homoploid Hybrid Speciation in *Ae. tauschii*[J].
335 *Molecular biology and evolution*. Doi:10.1093/molbev/msy215.
- 336 Li JJ, Liu HX, Mao SZ, Zhao B, and Huang SX. 2016. Adaptive evolution of the *ndhF* gene in the genus
337 *Rheum* (Polygonaceae). *Guihaia* 36:101-106.
- 338 Librado P, and Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data.
339 *Bioinformatics* 25:1451-1452.
- 340 Liu LX, Li R, Worth JR, Li X, Li P, Cameron KM, and Fu CX. 2017. The complete chloroplast genome of
341 Chinese Bayberry (*Morella rubra*, Myricaceae): implications for understanding the evolution of Fagales.
342 *Frontiers in Plant Science* 8:968.
- 343 Lohse M, Drechsel O, Kahlau S, and Bock R. 2013. Organellar Genome DRAW: a suite of tools for generating
344 physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic acids*
345 *research* 41:W575.
- 346 Lubbers EL, Gill KS, Cox TS, and Gill BS. 1991. Variation of molecular markers among geographically
347 diverse accessions of *Triticum tauschii*. *Genome* 34:354-361.
- 348 Maier RM, Neckermann K, Igloi GL, and Kossel H. 1995. Complete sequence of the maize chloroplast
349 genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing.
350 *International Journal of Biological Macromolecules* 251: 614-628.
- 351 Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M, Jakobsen K S, Wulff BBH, Steuernagel B, Msyer
352 KFX, and Olsen O-A. 2014. Ancient hybridizations among the ancestral genomes of bread wheat. *Science*

353 345(6194):1250092.

354 Matsuoka Y, Mori N, and Kawahara T. 2005. Genealogical use of chloroplast DNA variation for intraspecific
355 studies of *Ae. tauschii* Coss. *Theoretical and Applied Genetics* 111:265-271.

356 Matsuoka Y, Takumi S, and Kawahara T. 2015. Intraspecific lineage divergence and its association with
357 reproductive trait change during species range expansion in central Eurasian wild wheat *Aegilops tauschii*
358 Coss. (Poaceae). *BMC Evolutionary Biology* 15:1-10.

359 Middleton CP, Senerchia N, Stein N, Akhunov ED, Keller B, Wicker T, and Benjamin K. 2014. Sequencing of
360 chloroplast genomes from wheat, barley, rye and their relatives provides a detailed insight into the
361 evolution of the Triticeae Tribe. *PLoS ONE* 9:e85761.

362 Mizuno N, Yamasaki M, Matsuoka Y, Kawahara T, and Takumi S. 2010. Population structure of wild wheat
363 D-genome progenitor *Aegilops tauschii* Coss.: implications for intraspecific lineage diversification and
364 evolution of common wheat. *Molecular Ecology* 19:999-1013.

365 Palmer JD. 1991. Plastid chromosomes: structure and evolution. *The molecular biology of plastids* 7: 5-53.

366 Posada D. 2008. jModelTest: phylogenetic model averaging. *Molecular biology and evolution* 25:1253-1256.

367 Sang T. 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. *Critical Reviews in*
368 *Biochemistry and Molecular Biology* 37:121-147.

369 Singh S, Chahal GS, Singh PK, and Gill BS. 2012. Discovery of desirable genes in the germplasm pool of *Ae.*
370 *tauschii* Coss. *Indian Journal of Genetics and Plant Breeding* 72:271-277.

371 Tabidze V, Baramidze G, Pipia I, Gogniashvili M, Ujmajuridze L, Beridze T, Hernandez AG, and Schaal B.
372 2014. The complete chloroplast DNA sequence of eleven grape cultivars. Simultaneous resequencing
373 methodology. *Journal International Des Sciences De La Vigne Et Du Vin* 48: 99-109.

374 Terakami S, Matsumura Y, Kurita K, Kanamori H, Katayose Y, Yamamoto T, and Katayama H. 2012.
375 Complete sequence of the chloroplast genome from pear (*Pyrus pyrifolia*): genome structure and
376 comparative analysis. *Tree Genetics and Genomes* 8(4):841-854.

377 Wang J, Luo MC, Chen Z, You FM, Wei Y, and Zheng Y. 2013. *Aegilops tauschii* single nucleotide
378 polymorphisms shed light on the origins of wheat D-genome genetic diversity and pinpoint the
379 geographic origin of hexaploid wheat. *New Phytologist* 198:925-937.

380 Wei HT, Li J, Peng ZS, Lu BR, Zhao ZJ, and Yang WY. 2008. Relationships of *Aegilops tauschii* revealed by
381 DNA fingerprints: The evidence for agriculture exchange between China and the West. *Progress in*
382 *Natural Science-materials International* 18:1525-1531.

383 Wolfe KH, Li WH, and Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant
384 mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences*
385 84:9054-9058.

- 386 Wyman SK, Jansen RK, and Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA.
387 *Bioinformatics* 20:3252-3255.
- 388 Yamane K, and Kawahara T. 2005. Intra- and interspecific phylogenetic relationships among diploid *Triticum-*
389 *Aegilops* species (Poaceae) based on base-pair substitutions, indels, and microsatellites in chloroplast non-
390 coding sequences. *American Journal of Botany* 92:1887-1898.
- 391 Yang M, Zhang X, Liu G, Yin YX, Chen KF, Yun QZ, Zhao DJ, and Yu J. 2010. The complete chloroplast
392 genome sequence of date palm (*Phoenix dactylifera* L.). *PLoS ONE* 5:e12762.
- 393 Zhang YJ, Ma PF, and Li DZ. 2011. High-throughput sequencing of six bamboo chloroplast genomes:
394 Phylogenetic implications for temperate woody bamboos (Poaceae: Bambusoideae). *PLoS ONE* 6:e20596.

Figure 1

Geographical distribution of 17 *Aegilops tauschii* accessions from western Turkey to eastern China

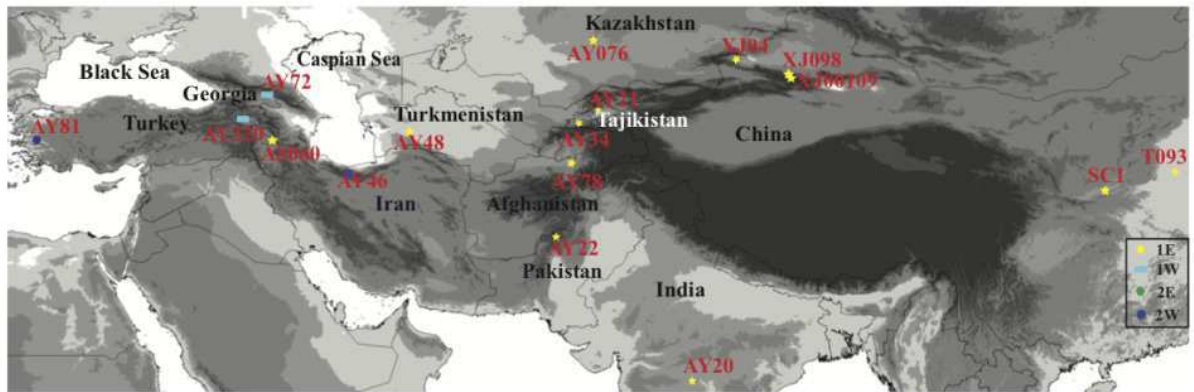
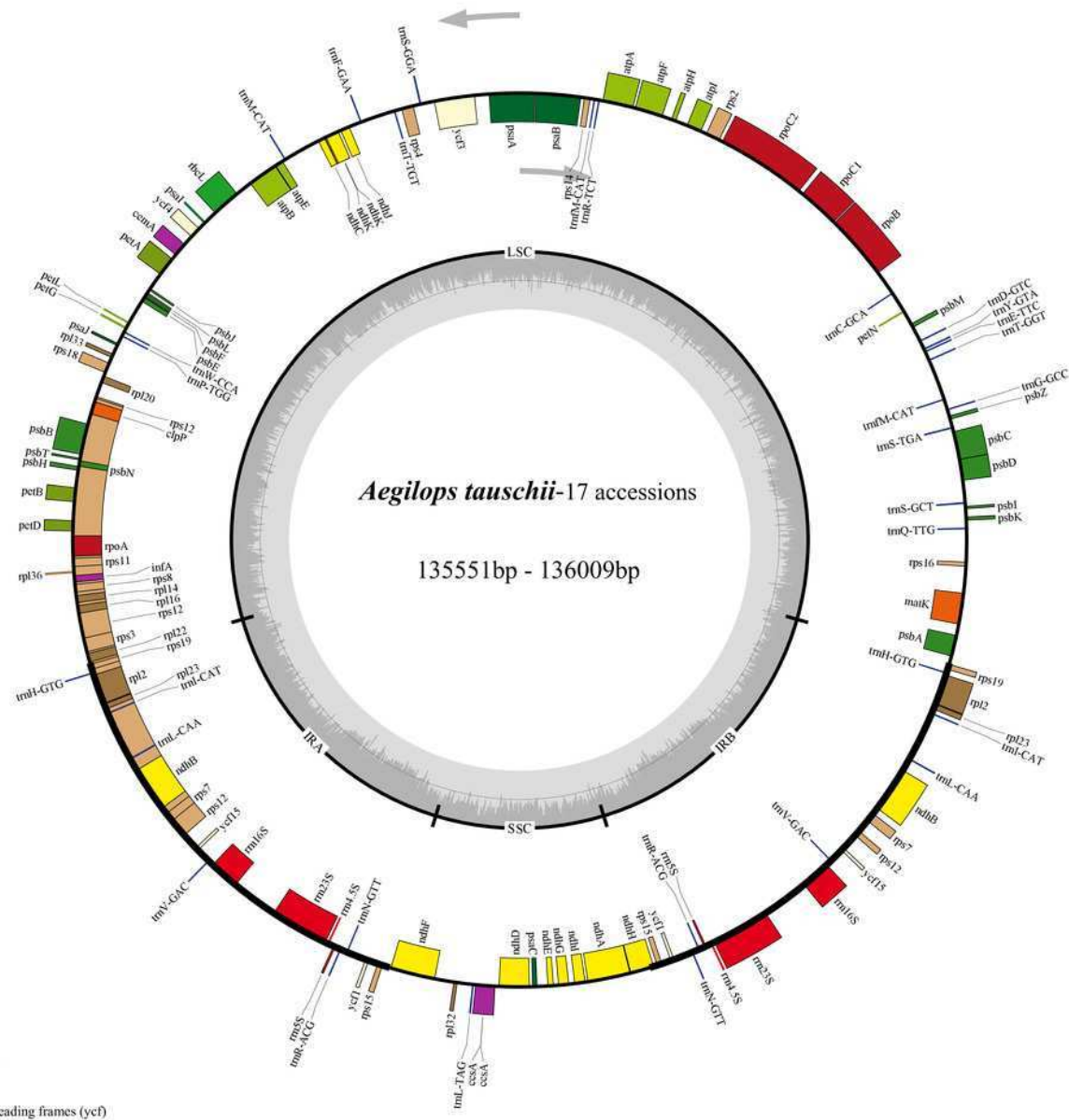


Figure 2

Chloroplast genome map of 17 *Aegililops tauschii* accessions.

The genes lying outside and inside of circle are transcribed in the counterclockwise and clockwise directions, respectively. Different colored bars are used to represent different functional gene groups. The darker gray area in the inner circle displays GC content. The fine lines indicate the boundary of the inverted repeats (IR_A and IR_B) that split the genomes into large single copy (LSC) and small single copy (SSC) regions.



- photosystem I
- photosystem II
- cytochrome b/f complex
- ATP synthase
- NADH dehydrogenase
- RubisCO large subunit
- RNA polymerase
- ribosomal proteins (SSU)
- ribosomal proteins (LSU)
- clpP, matK
- other genes
- hypothetical chloroplast reading frames (ycf)
- ORFs
- transfer RNAs
- ribosomal RNAs
- origin of replication
- introns
- polycistronic transcripts

Figure 3

Comparative analysis of the nucleotide variability (Pi) values among 17 *Aegililops tauschii* accessions.

The homologous regions are oriented according to their locations in the chloroplast genomes.

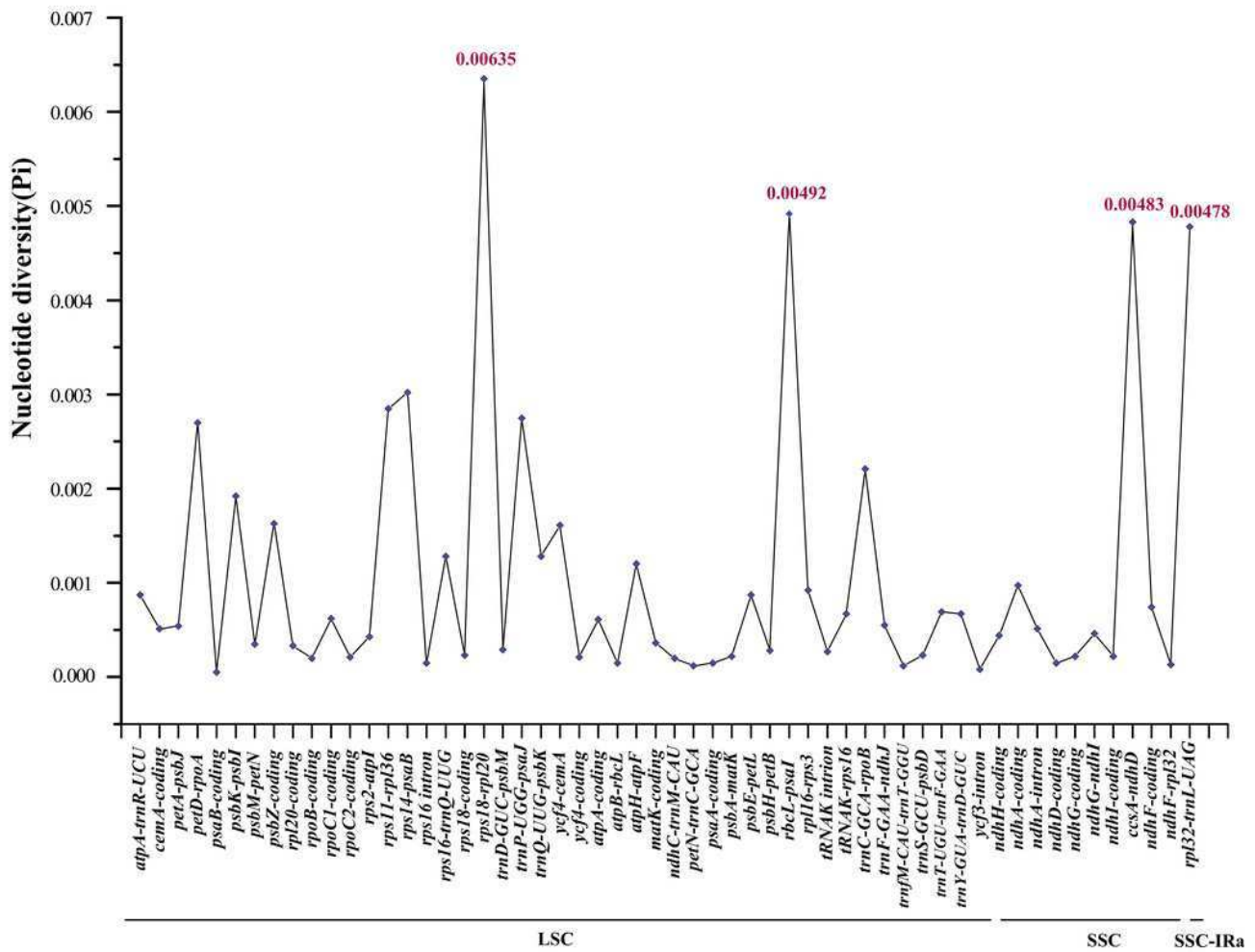


Figure 5

Phylogenetic tree constructed based on the complete chloroplast genomes of 17 *Aegililops tauschii* accessions by maximum likelihood (ML) method.

Bootstrap support values (<50%) were hided. The phylogenetic tree resulting from analysis of 135 984 bp in the alignment length of chloroplast genomes with all gap positions removed, including long stretches of the same nucleotide, short sequences appearing in opposite orientation and some sequences consisting short repeats.

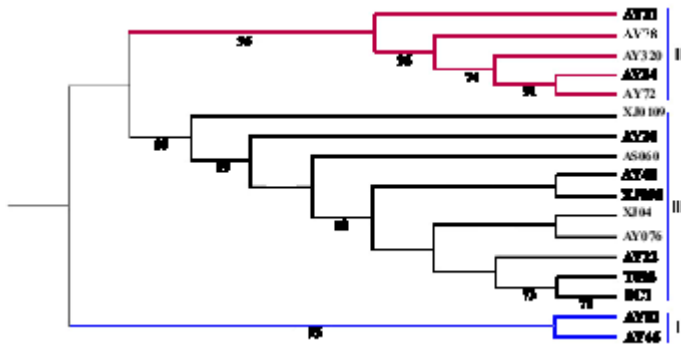
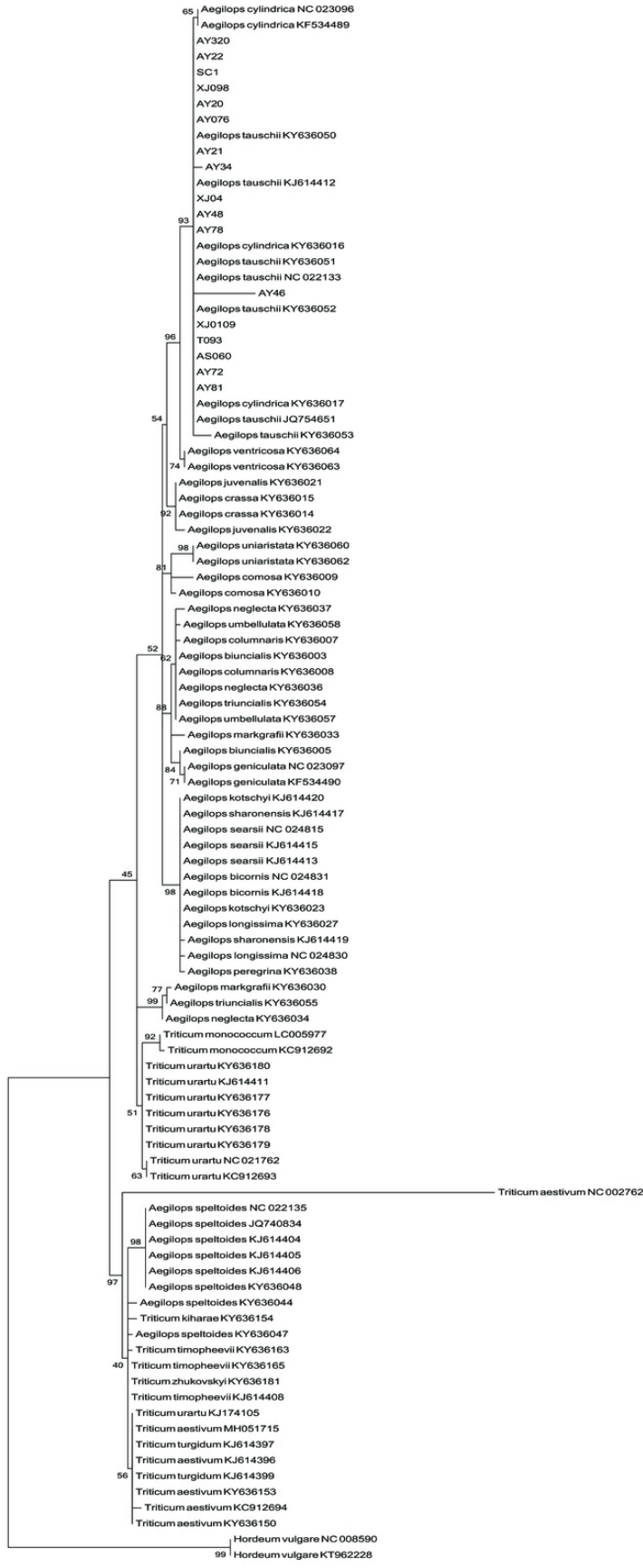


Figure 6

Phylogenetic tree constructed based on the complete chloroplast genomes of 99 Triticeae accessions by maximum likelihood (ML) method.

The phylogenetic tree resulting from analysis of 131 116 bp in the alignment length of chloroplast genomes with all gap positions removed, including long stretches of the same nucleotide, short sequences appearing in opposite orientation and some sequences consisting short repeats.



0.00050

Table 1 (on next page)

The origin country and collection regions of 17 *Aegilops tauschii* accessions

1 **Table 1**

Accessions	Source	Regions	GenBank accession number
SC1	Shannxi, China	N (34.158997), E (108.90699), Elevation: 428 meters	MN258085
AY81	Izmir, Turkey	Elevation: 30 meters	MN258083
AY34	Khujand, Tajikistan	N (39.771944), E (68.809444), Elevation: 433 meters	MN258078
AY22	Baluchistan, Pakistan	N (30.925), E (66.44638889), Elevation: 675 meters	MN223978
AY320	Kars, Turkey	N (40.13333333), E (43.06666667), Elevation: 1275 meters	MN258084
AY21	Khujand, Tajikistan	N (40.67388889), E (70.54694444), Elevation: 462 meters	MN223977
XJ04	Xinjiang, China	N (44.321239), E (80.77766), Elevation: 892 meters	MN258087
XJ0109	Xinjiang, China	N (43.386026), E (83.5977), Elevation: 1269 meters	MN258089
T093	Henan, China	N (35.728123), E (115.242698), Elevation: 52 meters	MN258086
XJ098	Xinjiang, China	N (43.386026), E (83.5977), Elevation: 1269 meters	MN258088
AY78	Konoz, Afghanistan	N (36.68333333), E (68.05), Elevation: 362 meters	MN258082
AS060	Iran	Unknown	MN223975
AY48	Turkmenistan, Balkan	N(38.48333333), E(56.3), Elevation: 730 meters	MN258080

AY076	Turkistan	N (45), E (70), Elevation: 210 meters	MN258090
AY20	India	N(20), E(77), Elevation: 509 meters	MN223976
AY72	Georgia	N (43),E (47) , Elevation: 90 meters	MN258081
AY46	Iran, Tehran	N(35.8), E(50.96666667), Elevation: 1296 meters	MN258079

2 The origin country and collection regions of 17 *Aegilops tauschii* accessions

3

4

5

Table 2 (on next page)

Comparative analysis of the chloroplast genomes among 17 *Aegilops tauschii* accessions

Number of tRNA genes	32(6)	32(6)	32(6)	32(6)	32(6)	32(6)	32(6)	32(6)	32(6)	32(6)	32(6)	32(6)	32(6)	32(6)	32(6)	32(6)	32(6)
Duplicated genes in IR	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16

3

4

5