THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Evolutionary origin of a tetraploid allium species in the Qinghai-Tibet Plateau

Citation for published version:
Li, M, Zheng, Z, Liu, J, Yang, Y, Ren, G, Ru, D, Zhang, S, Du, X, Ma, T, Milne, R & Liu, J 2021, 'Evolutionary origin of a tetraploid allium species in the Qinghai-Tibet Plateau', *Molecular Ecology*. https://doi.org/10.1111/mec.16168

Digital Object Identifier (DOI):
10.1111/mec.16168

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Molecular Ecology

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 24. Sep. 2022

Evolutionary origin of a tetraploid *Allium* species in the Qinghai-Tibet Plateau

Minjie Li[1†], Zeyu Zheng[1†], Juncheng Liu[2], Yongzhi Yang[1], Guangpeng Ren[1], Dafu Ru[1], Shangzhe Zhang[1], Xin Du[1], Tao Ma[2], Richard Milne[3,4], Jianquan Liu[1,2*]

[1]State Key Laboratory of Grassland Agro-ecosystem, Institute of Innovation Ecology & School of Life Sciences, Lanzhou University, Lanzhou 730000, People's Republic of China

[2]Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610065, People's Republic of China

[3]Institute of Molecular Plant Sciences, School of Biological Sciences, The University of Edinburgh, Daniel Rutherford Building, The King's Buildings, Mayfield Road, Edinburgh, EH9 3JH, UK

[4]Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh, EH3 5LR, UK

[*]Corresponding author: Jianquan Liu; **Email:** liujq@nwipb.cas.cn

[†] These authors contributed equally to this work.

22 **Abstract**

23 Extinct taxa may be detectable if they were ancestors to extant hybrid species, which
24 retain their genetic signature. In this study, we combined phylogenomics, population
25 genetics and fluorescence *in situ* hybridization (GISH and FISH) analyses to trace the
26 origin of the alpine tetraploid *Allium tetraploideum* (2n = 4x = 32), one of the five
27 known members in the subgenus *Cyathophora*. We found that *A. tetraploideum* was an
28 obvious allotetrapoploid derived from ancestors including at least two closely related
29 diploid species, *A. farreri* and *A. cyathophorum*, from which it differs by multiple
30 ecological and genomic attributes. However, these two species cannot account for the
31 full genome of *A. tetraploideum*, indicating that at least one extinct diploid is also
32 involved in its ancestry. Furthermore, *A. tetraploideum* appears to have arisen via
33 homoploid hybrid speciation (HHS) from two extinct allotetraploid parents, which
34 derived in turn from the aforementioned diploids. Other modes of origin were possible,
35 but all were even more complex and involved additional extinct ancestors. Our study
36 together highlights how some polyploid species might have very complex origins,
37 involving both HHS and polyploid speciation and also extinct ancestors.

38 **Keywords:** allotetraploid, extinction, homoploid hybrid speciation, *in situ*
39 hybridization, phylogenomics, population genetics

40

41

42

43 **Introduction**

44 Speciation, hybridization and extinction are major factors in plant evolution (Abbott,
45 Barton, & Good, 2016; Abbott et al., 2013; Anderson & Stebbins, 1954; Mittelbach et
46 al., 2007), but only the timing and nature of the first two can be detected, often with
47 precision, using molecular data (e.g., Abbott et al., 2016; Taylor & Larson, 2019;
48 Wang et al., 2020). Conversely, extinction has proved almost undetectable, except
49 using fossils (Quental & Marshall, 2010). Birth-death models, which estimate
50 extinction rates across a phylogeny, appear to be unreliable (Louca & Pennell, 2020).
51 However, hybrid speciation presents an opportunity to detect extinct or 'ghost' species,
52 which might leave their genomes, or part thereof, in their hybrid species offspring. This
53 provides a possible pathway for detecting extinct taxa and placing them onto molecular
54 phylogenies. However, examples of this remain very rare, e.g., one extinct contributor
55 to a hybrid species was inferred in *Paeonia* (Ferguson & Sang, 2001). The best
56 opportunities for detecting extinct lineages might be provided by allopolyploid species
57 where two diploid species are involved in the formation of an allopolyploid species
58 with one diploid extant and the other extinct (Luo et al., 2017).

59 While hybrid species make the detection of extinct progenitors possible, the
60 extinction of any one ancestral species will however challenge the identification of the
61 true lineage of any hybrid species, because missing reference homolog(s), might result
62 in incorrect identification of the potential progenitors (e.g., Jakob & Blattner, 2010;
63 Luo et al., 2017; Roelofs et al., 1997). Nonetheless, understanding the role of extinction

64 is critical to gaining proper understanding of the hybrid speciation and reticulate
65 evolution, because the newly formed hybrid species may displace their parent species
66 (reviewed in Soltis et al., 2016), and potentially could cause their extinction due to
67 fitness advantages of the neo-hybrid. To show that the progenitors of a hybrid species
68 are extinct, they must be shown to be the sister to and not direct ancestors of the extant
69 non-hybrid species. For example, using genomic *in situ* hybridization (GISH), the
70 subgenome of such an extinct species was successfully distinguished and identified in
71 polyploid species *Oxyria sinensis*, via chromosomal or genomic differentiation with
72 other extant ancestors (Luo et al., 2017). In this way, patterns of reticulate evolution
73 involving extinct species may be elucidated, especially where multiple hybrid
74 speciation events are inferred. Such cases might involve both homoploid and
75 allopolyploid hybrid speciation, and may require precise analysis to work out which
76 type of event happened and when. Allopolyploidization leads to genome divergence
77 within a species, allowing new adaptations to arise via subfunctionalisation or
78 neofunctionalization (Adams & Wendel, 2005), as seen in crops (Doebley, Gaut, &
79 Smith, 2006; Mayer et al., 2014). Homoploid hybrid speciation (HHS) leads to
80 transgressive traits and genomic (or karyotypic) restructuring (Lamichhaney et al.,
81 2018; Li et al., 2016; Rieseberg et al., 1995) though this has only rarely been
82 demonstrated (Abbott et al., 2016; Schumer, Rosenthal, & Andolfatto, 2018; Wang et
83 al. 2021; Yakimowski & Rieseberg, 2014). Thus far, only one documented case of
84 HHS involve the parents that themselves might be polyploids: *Paeonia officinalis.* This

85 species is believed to have originated through HHS between two extant
86 allotetraploids, one of which itself has an extinct ancestor (Ferguson & Sang, 2001).
87 Hence much remains to be learned about homoploid hybrid species with complex
88 ancestry, especially where one or more ancestor species might be extinct.

89 The current study examines the ancestry and origin of an allotetraploid *Allium*
90 *tetraploideum*. It occurs at altitudes ranging from 3100 and 3700 m in the eastern
91 Qinghai-Tibet Plateau (QTP) (Figure 1A and 1B**)** (Li, Liu, Guo, Xiao, & He, 2019), a
92 region with numerous endemic species (Liang et al., 2018). Two closely related
93 diploid species (2n = 2x = 16), *A. cyathophorum* and *A. farreri* occur to the southwest
94 and northeast respectively, with little or no range overlap (Figure 1B), while two other
95 diploids, *A. mairei* (this also has tetraploid lineages; Li et al., 2016) and *A. spicatum*,
96 occur further south and west, respectively (Figure 1A). These five species constitute a
97 well-supported clade, known as *Allium* subg. *Cyathophora*, in all phylogenetic analyses
98 (Huang et al., 2014; Li et al., 2016). A previous study suggested that both *A.*
99 *cyathophorum* and *A. farreri* may have been involved in the origin of *A. tetraploideum*
100 (Li et al., 2016), but here we report a comprehensive analysis of its origin, implicating
101 in addition at least one extinct diploid and two other tetraploid extinct progenitors, as
102 recent ancestors. Furthermore, we determine that, like *P. officinalis* (Ferguson & Sang,
103 2001), *A. tetraploideum* originated by HHS from allopolyploid parent species, but in
104 this case as part of a complex history probably involving allopolyploidization prior to
105 HHS, plus at least three extinction events.

106

107 **Materials and Methods**

108 **Sampling plants, genetic extraction and sequencing**

109 We sampled one individual each of the five species of subg. *Cyathophora*, plus one of
110 *A. przewalskianum* as the outgroup, respectively for whole chloroplast genome
111 sequencing and RAN-Seq. We sampled 20 *A. tetraploideum*, 18 *A. cyathophorum*, 22
112 *A. farreri*, 5 *A. spicatum*, 1 *A. mairei* and 1 *A. przewalskianum* individuals for
113 population genetics (Figure S1; Table S1) using *Sanger* sequencing in Tsingke
114 (Beijing, China). We further sampled 32 *A. tetraploideum*, 27 *A. cyathophorum*, 12 *A.*
115 *farreri*, 3 *A. spicatum* and 3 *A. mairei* individuals for specific-locus amplified
116 fragment sequencing (SLAF-seq) (Figure S1; Table S1), in order to obtain the
117 simplified genome-scale data for population genomics analyses. Full genome
118 sequencing was impractical because of the exceptionally large C-values of all *Allium*
119 species (https://cvalues.science.kew.org /search/angiosperm) and the giant assembled
120 genome size (16.24 gigabases) of one congener (*A. sativum*) (Sun et al., 2020). Fresh
121 leaves for total genomic DNA (gDNA) extraction were immediately dried in silica gel
122 in the field, and fresh leaves intended for RNA-seq were placed in a 5 ml EP tube and
123 stored in a liquid nitrogen container.

124 Total RNAs were isolated from the frozen leaves for each sample using the
125 RNAprep pure Plant Kit (Tiangen Biotech, Beijing, China) following manufacturer's
126 instructions. RNA quantity was examined using Qubit2.0 (Life Technologies,

127 Carlsbad, CA, USA) and RNA quality was assessed using an Agilent 2100
128 Bioanalyzer. Sequencing libraries were prepared and then sequenced on Illumina
129 HiSeq X-ten platform (Biomarker, Beijing, China) with standard procedures of
130 RNA-seq methodology. Total gDNA was isolated from the silica-dried leaves for each
131 sample using Takara plant DNA isolation reagent following the
132 manufacturer-recommended protocol (Beijing, China). For chloroplast genomes,
133 sequencing libraries were prepared and sequenced on Illumina HiSeq 2000 (Novogene,
134 Beijing, China). For SLAF-seq, gDNA was digested by HaeIII (New England Biolabs)
135 and then sequenced on Illumina HiSeq 2500 system (Biomarker, Beijing, China)
136 following the manufacturer's protocol.

137 **Niche distinctness and ecological fitness of *A. tetraploideum***

138 *A. cyathophorum* and *A. farreri* are two closely related diploids of *A. tetraploideum*.
139 In order to examine niche differentiation between these three species, we recorded a
140 total of 47 georeferenced coordinates from sites for these species, and projected them
141 (Table S2) in ArcGIS Desktop 10.5.1 to define their background areas by adding one
142 decimal degree to the maximum and minimum known geographic coordinates, to
143 represent the "potential available areas" where species are likely to disperse (Barve et
144 al., 2011). Then a set of randomly selected points in this background area were used to
145 extract altitude and 19 current bioclimatic variables (Bio1–19) (Table S3) at the spatial
146 resolution of 2.5 arc-min from the World-Clim database (Hijmans, Cameron, Parra,
147 Jones & Jarvis, 2005). The R *ecospat* package

148 (https://cran.r-project.org/web/packages/ecospat/) was performed to quantify the
149 multivariate climatic overlap by estimating Schoener's D metric (Schoener, 1970) and
150 to execute niche equivalency/similarity tests (Warren, Glor & Turelli, 2008) with 100
151 iterations at a given resolution (R = 100). Density profile estimations were performed
152 for each bioclimatic variable using the sm.density.compare function of the R *sm*
153 package (https://cran.r-project.org/web/packages/sm/). The significance of the
154 difference was tested using 10,000 bootstrap replicates. All ecological variables
155 corresponding to the total 47 georeferenced coordinates were retrieved from the
156 World-Clim database at 30 arc-second resolution. Of these, six variables (Bio1, Bio2,
157 Bio3, Bio4, Bio12 and Bio15) exhibiting pairwise Pearson correlation coefficients $r <$
158 0.7 were selected to conduct principal component analysis (PCA). The principal
159 component (PC) loading scores for each species along axes one (PC1) and two (PC2)
160 (Table S4) were extracted using the R *factoextra* package
161 (http://www.sthda.com/english/rpkgs/factoextra). Based on these six variables, PC1
162 and PC2 scores, analyses of similarities (ANOSIMs) and permutational multivariate
163 analyses of variance (ADONIS2) were performed in the R *vegan* package
164 (https://cran.r-project.org/web/packages/vegan/) to statistically test whether there are
165 significant niche differences for each species pairs, with significance measured by 999
166 random permutations.

167 Stomata size and seed mass were compared between *A. tetraploideum* and the two
168 diploid relatives *A. cyathophorum* and *A. farreri*. Stomata size was calculated by

169 length×width (mm^2) at 20× objective of an electron microscope based on
170 measurements of 535, 508 and 470 stomata for those three species, each of 10
171 individuals from different populations. The 1000-seed masses were measured using an
172 electronic scale. For a germination experiment, seeds were first moist-cold stratified
173 (about –4°C) to break dormancy for 24 hours before being left in covered plastic petri
174 dishes (6 cm wide × 6 cm long) on the double layers of moistened filter paper. Seed
175 germination was carried out on a regulated heating block with three replicates of 50
176 seeds of each species at each of these temperatures: 5–10, 6, 8, 10, 15, 20, 25°C. Seeds
177 were determined to have germinated if they had macroscopically detectable buds; at
178 each temperature, this was checked every 24 hours. Seedling length was measured
179 every 48 hours after the 10th day of seed germination. Seed germination percentage
180 (GP) was estimated as (number of germinating seed)/(number of viable seeds), and
181 germination rate (GR) was measured as (G1+G2+…+Gn)/(G1T1+G2T2+ +…+GnTn)
182 ×100. Gn was the number of germinating seeds on the nth day, Tn was the number
183 corresponding days of germination. Pearson correlations between temperature and each
184 of GR, GP or seedling length were respectively tested using Mantel tests, by running
185 999 permutations for significance. ANOSIM and ADONIS2 tests were used to examine
186 whether GP, GR and seedling length were similar between *A. tetraploideum* and its two
187 diploid relatives.

188 **Phylogenetic analyses of *A. tetraploideum* within *Allium* based on chloroplast**
189 **genomes and transcriptomes**

190 To further confirm the position of *A. tetraploideum* within the phylogeny of *Allium*,
191 we examined a total of 47 chloroplast genomes from 43 *Allium* species, plus 46
192 transcriptomes from 34 *Allium* species, to reconstruct the phylogenies of *Allium* with
193 *Narcissus* species used as the outgroup (see Tables S5 and S6 for GenBank accession
194 numbers). We used Fast-Plast v.1.2.6 (https://doi.org/10.5281/zenodo.973887) to
195 rapidly finish *de novo* assembly of whole chloroplast genomes. Manual checking was
196 implemented in Geneious version 10 (Kearse et al., 2012). The genome annotation was
197 carried out using Plann version 1.1.2 (Huang & Cronk, 2015) by mapping to the
198 reference genome of *Allium cepa* (GenBank accession KM088015). To promote the
199 quality of multiple alignment of chloroplast genomes, we used
200 'get_annotated_regions_from_gb.py' to isolate individual fasta files of the annotated
201 genes and inter-genes from the joint file of 48 chloroplast genomes, and then applied
202 MAFFT (Katoh & Standley, 2013) with the 'L-INS-i' algorithm to align each fasta file.
203 Finally, 'concatenate_fasta.py' was used to concatenate the aligned fasta files together.
204 The Maximum-likelihood (ML) tree of the aligned chloroplast genomes was
205 constructed in RAxML (Stamatakis, 2014) using '-f a' parameter by running 100
206 bootstrap replicates to find the best-scoring ML tree.

207 For transcriptome analyses, the 150 bp long raw paired-end short reads were
208 filtered by removing adapter reads, N reads and low-quality reads using
209 FASTX-Toolkit (available at http://hannonlab.cshl.edu/fastx_toolkit/). The clean reads
210 of RNA-seq were mapped to the CDs file of *A. sativum*

211 (https://doi.org/10.6084/m9.figshare.12570947.v1) using Burrows-Wheeler Aligner
212 (BWA, Li & Durbin, 2009) with the 'mem' algorithm. PCR duplicate alignments
213 were masked using Picard MarkDuplicates (http://broadinstitute.github.io/picard).
214 The alignment bam files were sorted and indexed using SAMtools (Li & Durbin,
215 2009). The program ANGSD (Korneliussen, Albrechtsen, & Nielsen, 2014) was
216 employed to identify collinear blocks using all the bam files. A home-made Perl script
217 was used to extract CDs and those with the percentage of missing data over 50% in
218 each sequence were discarded. We used RAxML to reconstruct ML trees for each CD
219 sequence and the concatenated supermatrix of all CDs using the same parameters as
220 above. ASTRAL-III (Zhang, Sayyari & Mirarab, 2018) was performed to infer
221 species tree using CD gene trees with bootstrap support (BP) ≥70%.

222 **Phylotranscript analyses of *A. tetraploideum* within subg. *Cyathophora***

223 Our phylogenetic analyses of *Allium* indicated that *A. tetraploideum* belongs to subg.
224 *Cyathophora* (see results), and we thus performed phylogenetic analyses of subg.
225 *Cyathophora* with *A. przewalskianum* as the outgroup, to further clarify the
226 evolutionary history of *A. tetraploideum*. We first extracted the concatenated matrices
227 of single-copy genes (SCGs), CDs, non-CDs, codon1, codon2, codon1+2, codon3 and
228 protein sequences from the six MAFFT-alignment chloroplast genomes, as well as the
229 whole chloroplast genome matrix, to reconstruct ML trees using RAxML as above. We
230 also counted the shared single nucleotide polymorphisms (SNPs) between *A.*
231 *tetraploideum* and the other diploids in subg. *Cyathophora*, and those between *A.*

232 *cyathophorum* and the other three diploid species. Given the tetraploid state (2n = 4x =

233 32) of *A. tetraploideum* (Li et al., 2017), we expected the phylotranscriptomics of subg.

234 *Cyathophora* (Rabier, Ta, & Ané, 2014) to provide evidence for the genomic

235 heterogeneity of *A. tetraploideum*. The clean paired-end short reads of RNA-seq were

236 assembled into contigs using Trinity V 2.1.5 with default parameters (available at

237 https://github.com/trinityrnaseq/trinityrnaseq/wiki). CD-HIT (Fu, Niu, Zhu, Wu, & Li,

238 2012) was applied to retain non-redundant sequences with a threshold value of 0.95.

239 TRANSDECODER-v5.0.2 (https://github.com/TransDecoder/) was used to predict the

240 CDs and protein sequences in the open read frame (ORF) by setting the length of amino

241 acids to be longer than 100. OrthoMCL (Li, Stoeckert, & Roos, 2003) was employed to

242 identify orthologous sequence clusters of the five species of subg. *Cyathophora* and the

243 outgroup *A. przewalskianum* from the TRANSDECODER-predicted protein

244 sequences. For phylotranscriptomic analyses, we first generated orthologous clusters

245 that contained one sequence each from each of the six species (i.e., a 1:1:1:1:1:1 ratio)

246 using a custom Perl script. Then the homologous clusters that had between 1 and 4 copy

247 sequences of *A. tetraploideum* plus one copy sequence each from each single diploid

248 species (i.e., 1:1, 2:1, 3:1 or 4:1) were extracted using the above Perl script with

249 modifications. The selected homologous clusters were aligned using MAFFT with the

250 'L-INS-i' algorithm. Any homologous clusters that have one or more sequences with

251 length <300 bps or >50% missing characters were removed. The ML trees for

252 homologous clusters were reconstructed using RAxML as above. Genomic networks

253 were parsimoniously inferred in PhyloNet (Than, Ruths, & Nakhleh, 2008) using
254 single-copy homologous tree topologies with BP ≥70%, under the assumption of one or
255 two past hybrid events. We converted the gene trees that had BP ≥70 into ultra-metric
256 trees using the R *Phybase* package (https://github.com/lliu1871/phybase), and then
257 used Densitree (Bouckaert, 2010) to plot a superimposed tree to describe the resulting
258 phylogenetic topologies. Tree topologies in each homologous dataset were statistically
259 visualized using 'twisst.py'
260 (https://github.com/simonhmartin/twisst/blob/master/twisst.py).

261 **Estimates of synonymous divergence**

262 Methods based on the fraction of synonymous substitutions per synonymous site (Ks)
263 between a given pair of orthologs or paralogs can be employed to estimate the time
264 since divergence or duplication (e.g., Guan et al., 2019; Lynch & Conery, 2000). This
265 methodology is now widely applied to infer interspecific relationships and
266 whole-genome duplications (WGDs) at genome or transcriptome level (Yang et al.,
267 2019). We used INPARANOID Version 4.1. (Sonnhammer & Östlund, 2015) to
268 identify the interspecific homologous sequences, and then one-to-one orthologs
269 between the diploid species pairs and two-to-one homologs between *A. tetraploideum*
270 and *A. cyathophorum* or *A. farreri* were extracted using a custom Perl script. A
271 maximum-likelihood method was implemented in the YN00 program of the PAML
272 package Version 4.1 (Yang, 2007) to calculate Ks between homologous sequence pairs
273 aligned by MAFFT. A WGDdetector pipeline

274 (https://github.com/yongzhiyang2012/wgddetector) employed an integrated analysis to
275 finally output the Ks values of pairwise paralogs of *A. tetraploideum* from
276 protein-guided DNA alignments. The resulting Ks estimates from interspecific
277 orthologs were plotted as a histogram (Ks plot) with age distributions <1.0, while those
278 gene pairs with Ks ≥ 1.0 were filtered out due to sparse data. We used the Gaussian
279 finite mixture model by the expectation-maximization (EM) algorithm in the R *mclust*
280 package (Scrucca et al., 2016), to fit univariate Gaussian components of the resulting
281 Ks density distribution of *A. tetraploideum* with random start points and merging best
282 results. According to the best Bayesian information criterion (BIC) values, three
283 optimal components in the Gaussian mixture model were identified with Ks > 0.02
284 (Table S7). We thus plotted the Ks distributions with age distributions from 0.02 to 0.2
285 when *A. tetraploideum* was analyzed because the range of divergence ages among the
286 diploid relatives was <0.2 (Figure S2). We implemented the Gaussian mixture
287 distributions obtained from the EM algorithm to infer significant peaks in each
288 observed Ks distribution using the R *mixtools* package (Benaglia, Chauveau, Hunter, &
289 Young, 2009). A parametric bootstrap (B = 100) of the likelihood ratio statistic was
290 performed using the boot.comp function in the R *mixtools* package to test the most
291 likely number of Gaussian components that fit each Ks distribution. Sizer tests
292 (Chaudhuri & Marron, 1999) for visual assessment significance of features,
293 distinguishing between increases, decreases, or values not significantly different from
294 zero, were used to further corroborate the Gaussian mixture model components across a

295 distribution at various (log transformed) bandwidths, to distinguish true data features

296 from noise. For datasets excluding or including *A. tetraploideum*, Ks values of 1.0 or

297 ≤0.2, respectively, and Ks bandwidths of 0.01–0.1, and 0.01–0.2, respectively, were

298 used to identify significant (alpha = 0.05) features in each observed Ks distribution.

299 Those Gaussian plots which significantly shifted from increase to decrease in Sizer

300 were identified as true peaks. A same Ks peak identified simultaneously by *mclust* and

301 *mixtools* were defined to range between peaks determined by those two packages.

302 Assuming a generation time of two years for *Allium* species, and a mean mutation rate

303 of 1.25×10^{-8} (1.0×10^{-8}–1.5×10^{-8}) per site per year for plants (Ossowski et al., 2010),

304 the ages corresponding to the Ks plots were calculated with the formula: Ks/generation

305 time/mutation rate.

306 **Population genetic analyses based on multiple loci**

307 We selected 10 genes from the orthologous groups of subg. *Cyathophora* species for

308 Sanger sequencing at the population level. The corresponding primer pairs (Table S8)

309 were designed in their conserved sequences using Primer Premier 6. Polymerase Chain

310 Reaction (PCR) amplifications were conducted in 30 μL volume following protocols

311 summarized in Table S8. Amplified products of the diploids were purified and directly

312 sequenced using homologous primers. However, those of *A. tetraploideum* were

313 purified using a Tiangen Universal DNA Purification Kit (Biotech, Beijing, China),

314 and then ligated into pMD19-T Vector with a pMD19-T Vector Cloning Kit (TaKaRa,

315 Dalian, China), and finally 10 positive clones for each gene per individual were

316 sequenced using general primer pairs. SeqMan (DNAstar, Burlan, 2000) was used to
317 edit DNA sequences and to obtain a consensus sequence. Each gene sequence was
318 aligned using MEGA 6.0 (https://www.megasoftware.net/) and checked manually.
319 Clone sequences that have only one variant compared to other clones from the same
320 individual were removed. DNAsp version 6.0 (http://www.ub.edu/dnasp) was used to
321 generate a haplotype file for each nuclear gene. The newly obtained sequences were
322 deposited in GenBank with accession numbers MT479228–MT481754. PhyML
323 (Guindon et al., 2010) was applied to gene sequence clones of *A. tetraploideum*, to
324 rapidly generate clusters of closely related genes using a likelihood-based method.
325 Each resulting tree topology was visualized using FigTree v1.4.4
326 (http://tree.bio.ed.ac.uk/software/figtree).

327 **Population genomics analyses using SLAF-seq reads**

328 For SLAF-seq, fragments that were between 464 and 494 bp were defined as
329 SLAF-tags. We used LAST (https://gitlab.com/mcfrith/last) to cluster the homologous
330 SLAF reads into a fabricated reference genome with a threshold value of 0.95. The
331 SLAF reads were then mapped to this fabricated reference genome using BWA to
332 generate polymorphic SLAFs. GATK (Depristo et al., 2011) and SAMtools were
333 together applied to produce accurate SNPs. A total of 168,045 SNPs were selected to
334 construct the ML tree of subg. *Cyathophora* in RAxML using the same parameters as
335 above; and totally 222,781 SNPs were selected for population genetic analyses of *A.*
336 *cyathophoraum*, *A. farreri* and *A. tetraploideum*, after filtering the SNP data with minor

337 allele frequency ≥0.05 and missing data ≤0.5 at each site. The genetic differentiation
338 (F_{ST}) scores within each species and between each of the three species pairs were
339 calculated using VCFtools (Danecek et al., 2011). Principal component analysis (PCA)
340 was performed using the SmartPCA program from EIGENSOFT v6.0.1
341 (https://github.com/argriffing/eigensoft). The population structure was completed in
342 ADMIXTURE v1.2.3 (Alexander & Lange, 2011) by partitioning all involved
343 individuals into different genetic clusters (K) according to maximum-likelihood
344 probabilities, with the number of K varying from 1 to 6. The cross-validation errors
345 were calculated in admixture by running 100 bootstrap replicates and the optimal K
346 value was fixed by a lower cross-validation error.

347 **Coalescent test of the hybrid origin of *A. tetraploideum* based on SLAF-seq site**
348 **frequency spectra**

349 We employed FASTSIMCOAL v2.6 (Excoffier, Dupanloup, Huerta-Sánchez, Sousa,
350 & Foll, 2013) to test the hybrid origin of *A. tetraploideum* with coalescent modelling to
351 fit demographic models using folded site frequency spectra (SFS). We constructed the
352 joint SFS for the three species using EasySFS
353 (https://github.com/isaacovercast/easySFS). We simulated the likely origin history of
354 *A. tetraploideum* by testing 15 different evolutionary models (Figure S3), of which
355 eight (model1–8) indicated dichotomous or radiative tree topologies with or without
356 gene flow after isolation, three models (model9–11) represented allotetraploid
357 speciation via a single hybridization event with or without migration/size change after

358 divergence, and four models (model12–15) described allotetraploid formation via HHS
359 between two polyploid parents that were in turn derived from *A. cyathophorum*, *A.*
360 *farreri* and an earlier diverging ghost lineage during the origin of *A. tetraploideum*. We
361 assumed a mutation rate of 1.25×10^{-8} per site per year and a generation time of 2 years.
362 A total of 329 parameters were further estimated and used to perform comparison tests
363 between the observed and simulated data. All parameters were estimated using global
364 maximum-likelihood with 100 independent runs for each model by setting '-n 100000
365 –N 100000 -M 1e-5 -w 1e-5 -l 10 -L 40 --msfs'. Each run started from random starting
366 parameters extracted from uniform and log-uniform distributions, and then the results
367 were modelled as FREQ. Without the formation for population size of each species, a
368 log-uniform distribution with large interval of 10^2–10^6 was used for effective
369 population sizes for each prior. The interval of the recent population change was set to
370 1×10^4–1×10^6. All time intervals for the species split were set based on molecular dating
371 analysis of the subg. *Cyathophora* (Figure S4). The interval of the basal divergence
372 time was set to 1×10^5–1×10^7 (uniform). The large interval of 1×10^4–1×10^6 was used for
373 the time difference between the two hybridization events. Gene flow between each
374 population was set to 10^{-2}-10^2 haplotypes. The optimal demographical model was
375 identified by the maximum value of likelihood and the minimum value of Akaike
376 information criterion (AIC). The 95% Confidence intervals for the parameters from the
377 best model were constructed using a parametric bootstrapping approach with 100
378 independent runs per bootstrap.

379 ***In situ* hybridization analyses**

380 The seeds from at least 10 individuals of each species were collected in the field and
381 germinated in the laboratory for cytological analyses. The fresh root tips were
382 pre-treated with Nitrous Oxide for two hours, fixed in 3:1 ethanol–acetic acid for one
383 hour and stored in 70% ethanol based on the protocol of Mandáková & Lysak, (2016).
384 The root tips were further digested in 6% pectinase and cellulase in citrate buffer at
385 37℃ for 45 minutes. After digestion, these root tips were rinsed in distilled water and
386 squashed in 20 μL of 60 % acetic acid on the microscope slides. The most well-spread
387 metaphase chromosomes were selected for slide denaturation and hybridization.

388 The high-quality gDNAs of *A. mairei*, *A. spicatum*, *A. cyathophorum* and *A.*
389 *farreri* were isolated as described above. The hybridization probes were prepared using
390 nick translation to obtain DNA fragments ranging from 100 to 500 bp. The 25 μL
391 mixture of ≥500 mg gDNA, 0.6 μL DNase I dilution buffer, 2.5 μL 10× nick
392 translation buffer, 1.2 μL DNA polymerase 1, 2.5 μL A,C,G dNTPs, 2.5 μL dUTPs
393 and 16.4 μL ddH_2O was incubated in a PCR amplifier at 15℃ for 95 minutes. Finally,
394 gDNA probes of *A. mairei*, *A. spicatum*, *A. cyathophorum* and *A. farreri* were
395 respectively labelled by Alexa Fluor 546-14-dUTP, 594-5dUTP, 647-aha-dUTP and
396 488-5dUTP (Thermo Fisher). The hybridization mixture was prepared in 200 μL per
397 slide with 100 μL 100 % formamide, 40 μL 50 % dextran sulphate and 20 μL labelled
398 gDNAs in 20 μL 20× saline–sodium citrate (SSC) and 20 μL ddH_2O, which was then
399 denatured on a hotplate at 90°C for 10 min. Chromosome denaturation was conducted

400 with 100 μL 70% formamide in 100 μL 2× SSC at 80−85°C for 1.5 minutes. Genomic
401 *in situ* hybridization (GISH) was conducted in a humid chamber at 42°C and left
402 overnight. Posthybridization washes of the GISH preparations were done for 30
403 minutes at 42°C in 5× SSC and twice at 37°C for 30 minutes respectively in 2× SSC,
404 1× SSC and 0.1 ×SSC, and transferred to 1× blocking solution (containing 20 μL 1×
405 TNA buffer and 1 μL rhodamine anti-DIG-sheep in 80 μL ddH$_2$O) at 4°C overnight
406 to counterstain chromosomes. The prepared GISH slides were visualized and digitally
407 photographed using a PERKIN ELMER & OLYMPUS ULTRAVIEW VOX & LX81
408 confocal microscope. At least three different cell figures were photographed for each
409 prepared slide. Photographs for each of the four fluorochromes were individually
410 acquired by using appropriate excitation and emission filters (U-MNIBA3,
411 U-MWIGA3, U-MWU2, U-MWBV2 and U-MWIY2). Due to combining different
412 gDNA-probes in the chromosomes of *A. tetraplodieum*, the four monochromatic
413 images were pseudo-coloured, merged, and processed using Adobe Photoshop CS.
414 The 45S rDNA and 5S rDNA probes which were respectively marked by FAM
415 and TAMRA were purchased in Shanghai Sangon Biotech
416 (https://www.sangon.com/). The diluted rDNA-probes were dissolved in 5 μL 1×
417 Tris-EDTA buffer solution and 5 μL 2× SSC per slide. Hybridization was performed
418 in a humid chamber at 37°C for 2 hours. Posthybridization washing of the
419 rDNA-probe slides were done at 42 °C for 3 min in 2× SSC and in ddH$_2$O, and
420 air-dried. Afterwards, the preparations were counterstained using DAPI solution (1

421 mg/mL). Hybridization signals from FISH slides were analyzed using an
422 OLYMPUS microscope BX43+DP2.
423
424 **Results**
425 **Ecological distinctness and fitness of *A. tetraploideum***
426 The alpine tetraploid *A. tetraploideum* has a distinctive distribution niche that is distinct
427 from the two closely related diploid species, *A. cyathophorum* and *A. farreri*. Similarity
428 tests showed that the niche overlaps of these three species were more dissimilar to each
429 other than would be expected by chance (all $p > 0.05$) (Figure S5). Each of these
430 species were significantly different from each of the two others along six major
431 bioclimatic variables (Table S9). The most occupied densities and parameter spans of
432 all three species exhibited significant ($P < 0.02$) differences in altitude and
433 temperature-related variables in univariate comparisons (Figure S6). The tetraploid *A.*
434 *tetraploideum* occupies a clearly different niche from the other two diploids, and shows
435 a striking preference for environments that are changeable according to the
436 temperature-related variables. In addition, *A. tetraploideum* has significantly greater
437 stomata size and seed weight (Figure 1C and 1D), and also seedling growing advantage
438 of germinated seeds (Figure S7; Tables S9 and S10). These comparisons suggested that
439 *A. tetraploideum* has distinct adaptions to high altitudes compared to its two diploid
440 relatives.

441 **Phylogenomics of *A. tetraploideum* based on chloroplast genome and**
442 **transcriptome data**

443 Our phylogenomic analysis of *Allium* strongly supported (BP = 100) that subg.
444 *Cyathophora* comprises the previously reported five species and that *A. tetraploideum*
445 has a close relationship with the morphologically similar *A. cyathophorum* (Figures
446 1E, 1F, and S8). The whole chloroplast genome-based phylogenetic tree of subg.
447 *Cyathophora* consistently indicated that *A. tetraploideum* is sister to *A. cyathophorum*,
448 and that this pair is sister to *A. spicatum* plus *A. farreri*, all four of which form a clade
449 sister to *A. mairei* (Figure 2A). This generally conforms to the previous population
450 genetic analyses of these species (Li et al., 2016) based on the chloroplast DNA
451 fragments (Figure S9). The counted SNPs showed that *A. tetraploideum* clearly shares
452 more SNPs with *A. cyathophorum* (1503; Figure 2A) than it does with the other three
453 diploid chloroplast genomes (454, 1066, 994; Figure 2A). Additionally, the number of
454 shared SNPs between the other three diploid species and *A. cyathophorum* (373, 986,
455 908; Figure 2A) are only slightly higher than between these and *A. tetraploideum*,
456 indicating that the chloroplast genomes of the other three diploid relatives are
457 approximately equidistant from those of *A. cyathophorum* and *A. tetraploideum*. These
458 attributes indicate that the tetraploid *A. tetraploideum* possesses a captured chloroplast
459 (genome) from the diploid *A. cyathophorum*.

460 To further identify whether other species contributed to the origin of *A.*
461 *tetraploideum*, we identified homologous genes from all subg. *Cyaphophora* species

462 plus the outgroup *A. przewalskianum* using OrthoMCL to reconstruct single copy gene
463 (SCG) trees. At first, 540 high-supported SCG trees with BP ≥70 for clades were
464 obtained (Figure 2B). Among these trees, we found that the phylogenetic relatedness
465 among *A. farreri* (A), *A. cyathophorum* (B), and *A. tetraploideum* (C) varied greatly,
466 with tree topologies A(B,C), B(A,C) and C(A, B) occurring respectively in 314, 195
467 and 31 gene trees (Table S11). Integrating these gene trees into a Densitree analysis,
468 we found that *A. tetraploideum* has medium strength connections to both *A.*
469 *cyathophorum* and *A. farreri*, and also much weaker connections to *A. spicatum* and *A*
470 *mairei* (Figure 2B). When excluding *A. przewalskianum*, a similar phylogenetic pattern
471 was observed for subg. *Cyathophora* among the 807 SCG trees having BP ≥70 (Table
472 S11). Combing both sets of SCG trees into a PhyloNet-based network (Table S12)
473 indicated that *A. farreri* contributed 0.46–0.47 of the *A. tetraploideum* genome, and *A.*
474 *cyathophorum* contributed 0.53–0.54 of the *A. tetraploideum* genome (Figure 2C).

475 Due to the hybrid and polyploidy origin of *A. tetraploideum*, the SCGs in its
476 diploid relatives would correspond to multiple copies in *A. tetraploideum*. We therefore
477 generated further homologous gene trees in which only SCGs members of subg.
478 *Cyathophora* were present, including between 2 and 4 copies of the same genes for *A.*
479 *tetraploideum*, to confirm the source of paralogous genes of *A. tetraploideum*. When all
480 five species of subg. *Cyathophora* were included, ~ 90% of all the trees presented two
481 tree topologies: ~ 53% had the (*farreri*,(*tetraploideum*,*cyathophorum*)) topology, and
482 ~37% had the (*cyathophorum*(*tetraploideum,farreri*,)) topology. Nevertheless, the

483 remaining ~10% had the topology (*tetraploideum*(*farreri*,*cyathophorum*)) which
484 indicated that there must be an additional genome source for *A. tetraploideum* (Figure
485 2D). The relatively low percentage for the latter tree topology likely resulted from the
486 misidentification of potential parental progenitors due to missing reference homolog(s)
487 (Figure S10), meaning that probably much more than 10% of the *A. tetraploideum*
488 genome comes from a source other than extant diploid species. When *A. mairei* or *A.*
489 *spicatum* were removed from the analysis, a similar result was obtained (Figure 2E).

490 **Ks-based genomic divergence and duplication**

491 The Ks-determined interspecific relationships between the four diploid species of subg.
492 *Cytahophora* were highly consistent with the phylotranscriptomic results (Figure 1F).
493 That is, the youngest Ks peak between *A. cyathophorum* and *A. farreri* (*c*. 0.042)
494 indicated they are the most recently diverged sister species; the intermediate Ks peaks
495 (*c*. 0.083–0.086) between *A. spicatum* and *A. cyathophorum* - *A. farreri*, indicated that
496 the next most recent divergence event was with *A. spicatum*; the largest Ks peaks (*c.*
497 0.106–0.120) between *A. mairei* and the other three, denoted the earliest divergence
498 event, involving *A. mairei* (Figure S2). The interspecific divergence ages estimated
499 from these Ks peaks (ranging between 4.80–1.68 million years ago (Mya); Table S13)
500 also coincide with those inferred using the constant mutation rate and sequence
501 variation of chloroplast DNA (ranging between 5.19–2.11 Mya; Figure S4).

502 We next examined the divergence between paralogs of *A. tetraploideum*, and that
503 between homologs of *A. tetraploideum* and each of *A. cyathophorum* and *A. farreri*,

504 based on those clusters that comprised two *A. tetraploideum* sequences and one from
505 the diploid (i.e., a ratio of 2:1). Our Gaussian mixture model identified three peaks for
506 Ks distribution of paralogs of *A. tetraploideum* (Figures S2 and S11; Table S14),
507 which were significant according to Sizer tests (Figure S2). Two of these significant
508 peaks (per Ks distribution) were also confirmed by the Sizer analyses for *A.*
509 *tetraploideum*:*A. cyathophorum* (2:1) and *A. tetraploideum*:*A. farreri* (2:1) homologs
510 (Figures S9 and S11). The second Ks peak (0.036–0.041) within *A. tetraploideum*
511 corresponded closely to the orthologous divergence between *A. cyathophorum* and *A.*
512 *farreri* (0.042) (Figure S2), and was fairly close to homologous divergence (1:2)
513 between *A. tetraploideum* and each of *A. cyathophorum* (0.043–0.049) and *A. farreri*
514 (0.048–0.049) (Figure 2F). This similarity likely reflects the divergence of these two
515 species' subgenome outside of, and then continuing both within and outside of, the
516 tetraploid *A. tetraploideum.* Another high peak occurred at 0.066–0.079 within *A.*
517 *tetraploideum*, corresponding probably to a peak of 0.086–0.088 between *A.*
518 *cyathophorum* and *A. tetraploideum*, and possibly also a 0.100–0.107 peak between *A.*
519 *farreri* and *A. tetraploideum* (Figure 2F); this might represent the earlier divergence of
520 an unknown ghost ancestor of *A. tetraploideum* from the common ancestor of *A.*
521 *cyathophorum* and *A. farreri* (Figure 2F). If these values do represent the same peak,
522 then from this data the predicted ghost lineage might have diverged after *A. mairei*
523 (peaks with *A. cyathophorum* and *A. farreri* at 0.106 and 0.112, respectively), but it
524 might be sister to *A. spicatum* (which has peaks with *A. cyathophorum* and *A. farreri*

525 at 0.083 and 0.086 respectively) (Figure S2), or possibly could have diverged after *A.*

526 *mairei* but before *A. spicatum*. The youngest significant Ks peak detected was

527 (0.025–0.026) in *A. tetraploideum*, and this might represent divergence of

528 homologous genes following polyploidization (Figure 2F). However, it also

529 corresponds closely to a peak of 0.024 in *A. tetraploideum* vs. *A. cyathophorum*,

530 whereas there was no comparable peak in *A. tetraploideum* vs. *A. farreri* (Figures S2

531 and S11); hence a more complex explanation is also possible, involving later gene

532 exchange with *A. cyathophorum*, resetting the clock for genomic divergence on

533 germplasm newly shared between these species. Hence the three peaks in *A.*

534 *tetraploideum* correspond, in chronological order, to the divergence of (i) the ghost

535 lineage, (ii) *A. cyathophorum* from *A. farreri*, and (iii) subgenomes following

536 polyploidization and/or possible introgression (Figure 2F).

537 **Population genetic analyses based on multiple loci**

538 We cloned and sequenced nuclear sequences at ten gene loci to obtain haplotypes for

539 each gene. Our trees for these genes strongly indicated that haplotypes within *A.*

540 *tetraploideum* fall into three categories: those that clustered with *A. cyathophorum*

541 haplotypes (Te-Cy), those that clustered with *A. farreri* haplotypes (Te-Fa), and those

542 that clustered with neither, clustering with *A. spicatum* or not with any species (Te-Gh)

543 (Figure 3A). The Te-Cy type was present for every gene, with *A. tetraploideum*

544 haplotypes nested within *A. cyathophorum* in each case. The Te-Fa type was present for

545 six genes, with *A. tetraploideum* haplotypes being sister to the monophyletic lineage of

546 *A. farreri* in five cases (9266-2, 9360-14, 7214-27, 9341-10 and 9259-23) or nested
547 within *A. farreri* in one gene (10076-6). The Te-Gh was present for nine genes, the
548 exception being 7214-27 (Figure 3A). We found that 6 of the 10 gene trees were
549 topologically consistent with phylotranscriptomic analyses (Figures 1F and S8)
550 regarding the relationships of the four diploid species, in which the sequence of
551 divergence was *A. mairei*, then *A. spicatum* and finally *A. cyathophorum* from *A.*
552 *farreri*. However, *A. spicatum* was sister to *A. cyathophorum* or *A. farreri* in the four
553 remaining gene trees (Figure S12). Interestingly, when adding haplotype sequences of
554 *A. tetraploideum* to sequence file of the four diploids for phylogenetic analyses, the
555 position of *A. spicatum* varied greatly (Figure 3A), likely indicating a close genetic
556 relationship between *A. spicatum* and the ghost lineage.
557 **Population genomics analyses based on SLAF-SNPs**
558 Phylogenetic analysis of all SLAF-SNPs suggested that *A. mairei* was the first
559 diverging lineage within subg. *Cyathophora*, with all sampled individuals of the other
560 four species comprised a monophyletic clade. Within that clade, *A. tetraploideum* was
561 separated from the three diploid species, and among those *A. farreri* and *A.*
562 *cyathophorum* were more closely related to each other than to *A. spicatum* (Figure
563 3B). The *F*-statistics based on genome-wide SNPs exhibited a high genetic
564 differentiation (F_{ST} = 0.344) between *A. tetraploideum* and *A. farreri*. However, the
565 differentiation between *A. cyathophorum* and *A. tetraploideum* (F_{ST} = 0.238) was
566 similar that between *A. cyathophorum* and *A. farreri* (F_{ST}= 0.268) (Figure 3C). Both

567 PCA and Structure analyses suggested that *A. tetraploideum* was well differentiated
568 with the two diploid species, *A. cyathophorum* and *A. farreri* (Figure 3D and 3E).
569 We further examined the origin of *A. tetraploideum* based on SLAF-seq site
570 frequency spectra, comparing 15 models regarding the origin of *A. tetraploideum*
571 using coalescence simulation in a composite likelihood framework (Figure S3). The
572 best-fitting evolutionary model (Table S15) predicted that three hybridization events
573 had been involved in the origin of *A. tetraploideum* plus instances of bidirectional
574 gene flow (Table S15). Under this model, one assumed extinct diploid lineage (a
575 'ghost') first diverged at ~3.54 Mya (95% confidence interval: 4.04–3.04 Mya), and
576 the split between *A. cyathophorum* and *A. farreri* occurred around 2.87 Mya (95%
577 confidence interval: 3.52–2.21 Mya) (Figures 3F and S13). These ages coincide with
578 both Ks-based ages (Table S13) and molecular dating analysis (Figure S4). At ~2.01
579 Mya (95% confidence interval: 2.92–1.10 Mya), the ghost genome (GG) hybridized
580 independently with each of *A. cyathophorum* (CC) or *A. farreri* (FF), leading to the
581 two parental allotetraploids (GGCC and GGFF) of *A. tetraploidum*. The final
582 formation of *A. tetraploideum* involved a third, homoploid hybrid speciation event at
583 at 0.83 Mya (95% confidence interval: 1.55–0.12 Mya), between the two extinct
584 allotetraploids (GGCC and GGFF).

585 **GISH and FISH experiments on *A. tetraploideum* and related species**

586 Our phylogenomics and population genetics/genomics analyses (Figures 2 and 3)
587 strongly indicated that *A. tetraploideum* is a tetraploid hybrid species probably

588 generated from *A. cyathophorum*, *A. farreri*, and one other extinct/unknown diploid
589 lineage, apparently via two short-lived allotetraploid parents that are also extinct. To
590 further confirm this likely origin, we carried out GISH and FISH analyses of *A.*
591 *tetraploideum* with gDNAs of the other four diploid species and 45s and 5s rDNAs.
592 We found that 22 out of 32 chromosomes of *A. tetraploideum* were labelled by the
593 gDNA-probes of the four diploid species (Figure 4A), whereas the remaining 10
594 chromosomes were not (23–32 blue chromosomes; Figure 4A), and showed no
595 hybridization signs from any extant species. When separately visualizing each of the
596 four monochromatic images, we observed that gDNAs of *A. cyathophorum* and *A.*
597 *farreri* respectively hybridized to 16 (red; Figure 4B) and 14 (green; Figure 4C)
598 chromosomes of *A. tetraploideum*. However, when these two images were merged, ten
599 chromosomes were indicated to hybridize to gDNA-probes of both *A. cyathophorum*
600 and *A. farreri* (yellow; Figure 4F). When further merging the remaining two
601 monochromatic images (Figure 4D and 4E), six of those ten chromosomes (rose-pink;
602 Figure 4A) were labelled for all four diploid species, indicating that they hybridised
603 with all of them. Therefore, twelve chromosomes were stained by one species only,
604 specifically six of *A. cyathophorum*, four of *A. farreri* and two of *A. spicatum* (Figure
605 4A).

606 To test whether the six chromosomes that were labelled for all four species did so
607 because they derived from the species' common ancestor, reciprocal GISH analyses
608 between each diploid species pair were carried out. When gDNA of each species was

609 hybridized to its chromosomes, a uniform fluorescence signal was consistently found
610 across the whole chromosomal complement (Figure S14). However, when we
611 hybridized gDNA of one species onto the chromosomes of the other congeners with the
612 blocking DNAs, no hybridization signal was observed. We thus conducted reciprocal
613 GISH experiments using gDNA of one diploid species to another diploid without the
614 blocking DNAs. The observed presence or absence of fluorescence on chromosomes
615 were positively related to interspecific genomic affinities recovered by phylogenomics
616 or Ks-based divergence (Figures 2 and S2); for example, gDNA of *A. cyathophorum*
617 hybridised with eight, six and two from *A. farreri*, *A. spicatum* and *A. mairei*
618 respectively; while gDNA of *A. farreri* hybridized with four and two chromosomes
619 from *A. spicatum* and *A. mairei*, respectively; and gDNA of *A. spicatum* hybridized
620 with two chromosomes of *A. mairei* (Figure S14).

621 Our FISH results further showed two pairs of chromosomes with 45S and 5S
622 rDNA loci for diploid *A. cyathophorum* and *A. farreri,* respectively. Only one pair of
623 45S rDNA (Figure 4G and 4H) loci were found for each diploid species. However, *A.*
624 *farreri* has one pair of telomere 5S rDNA loci (Figure 4H), while two pairs of loci
625 (Figure 4G) were found for *A. cyathophorum* in a pair of homologous chromosomes.
626 In addition, only one pair rather than the expected two pairs of 45S rDNA loci were
627 found for *A. tetraploideum* (Figure 4I) although both *A. farreri* and *A. cyathophorum*
628 were assumed to be involved in the origin of this tetraploid (Figure 3F). In addition,

629 two pairs of telomere 5S rDNA loci were detected in *A. tetraploideum* (Figure 4I) and
630 both were highly analogous to those of *A. farreri* (Figure 4J).
631
632 **Discussion**
633 Resolving the ancestry of a current polyploid species is essential to understanding its
634 speciation history, and offers the potential to detect extinction events among its
635 ancestors. However, extinct ancestors also make this task extremely challenging. In this
636 study, we mainly aimed to identify the ancestors and infer the likely speciation pattern
637 of the alpine tetraploid *Allium tetraploideum* (2n = 4x = 32), which has distinct
638 adaptions to high altitude conditions compared to its closely related diploids, *A.*
639 *cyathophorum* and *A. farreri* (2n = 2x = 16). Our data indicated that both extant
640 diploids plus at least one extinct lineage might have contributed to the formation of *A.*
641 *tetraploideum*, and that this origin most likely involved both allopolyploid and
642 homoploid hybrid speciation although other even more complex scenarios could not
643 be excluded. In the interpretation that best fits the data, two further polyploid ancestors,
644 both now extinct, formed *A. tetraploideum* by homoploid hybrid speciation. Our
645 results highlight the complexity of the origin of some polyploid species, and suggest
646 how frequent lineage extinctions may be a part of this process.
647 **The origin of *A. tetraploideum* involved two extant diploid species plus at least one**
648 **extinct diploid lineage**

649 Based on the maternally inherited chloroplast, *A. tetraploideum* is sister to the diploid
650 *A. cyathophorum*, while *A. farreri* is sister to *A. spicatum* (Figures 1E and 2A).
651 However, phylogenetic analyses of homologous genes comparing *A. tetraploideum*
652 with its four closest diploid relatives plus the outgroup, indicated reticulate
653 relationships, with *A. tetraploideum* equally linked to *A. cyathophorum* and *A. farreri*,
654 and less so to other species (Figure 2B and 2C). From this alone, the simplest
655 explanation would be that *A. tetraploideum* is the direct allopolyploid result of *A.*
656 *cyathophorum* × *A. farreri* with the evolutionary development of some new genomic
657 characters detected here. However, the involvement of a third, extinct, diploid ancestor,
658 is supported by the following lines of evidence. First, phylogenetic analyses of
659 homologs between the diploids and *A. tetraploideum* revealed topologies in which
660 some *A. tetraploideum* homologs formed a clade not close to either of the two
661 diploids (Figure 2D and 2E), suggesting the involvement of at least one extinct lineage
662 in the origin of *A. tetraploideum* (Figure S10). Second, under the single allopolyploid
663 hypothesis given above, only one genome divergence event, marked by a Ks peak,
664 would be expected, produced by divergence of the diploid ancestors (later
665 polyploidization would not provide a second peak because all chromosome sets are
666 already diverging). However, we identified three significant Ks peaks within the
667 paired homologs of *A. tetraploideum* (Figure 2F), of which one (0.036–0.041)
668 corresponded closely to the homologous divergence (0.042) between *A.*
669 *cyathophorum* and *A. farreri*, and another younger peak (0.025–0.026) might

670 represent polyploidization, but an older (0.066–0.079) peak could only represent
671 divergence from an earlier diverging, now extinct, ancestor lineage (Figure 2F).
672 Comparing peaks representing the divergence of *A. farreri* + *A. cyaphophorum* from
673 *A. tetraploideum* (~0.096), from *A. mairei* (~0.109) and from *A. spicatum* (~0.085)
674 (Table S14), suggests that the extinct lineage appears to have diverged after *A. mairei*
675 but before *A. spicatum*. Our coalescence-based simulations also support this, because
676 the split of *A. cyathophorum* + *A. farreri* from a ghost progenitor was estimated to
677 have occurred 3.54 Mya, equivalent to Ks = 0.089 (Figure 3F**)**. Third, phylogenetic
678 analyses on various nuclear loci, involving multiple individuals and ten cloned
679 sequences per individual, repeatedly showed admixture of some sequences from *A.*
680 *tetraploideum* with all those of *A. cyaphoporum*, whereas other *A. tetraploideum*
681 sequences formed one or two clades elsewhere in the tree, most often sister to *A.*
682 *farreri*, or occasionally closer to *A. spicatum* (Figure 3A). Fourth, our
683 population-scaled genomic data revealed a clearly different genetic complement for *A.*
684 *tetraploideum* when compared with *A. cyathophorum* and *A. farreri* (Figure 3B, 3C,
685 3D and 3E), and the coalescence simulations suggest that this complement might have
686 been contributed by an unknown ghost lineage (Figure 3F). Fifth, GISH analyses
687 hybridizing gDNAs of the four diploid relatives onto *A. tetraploideum* chromosomes,
688 indicated that its 32 chromosomes were divided into six that hybridized with all the
689 four diploids, 14 that hybridized with *A. cyathophorum* and/or *A. farreri*, two that
690 hybridized only with *A. spicatum*, and ten that hybridized with no extant diploids,

691 respectively (Figure 4A). Taken together, these lines of evidence seem to suggest that
692 *A. tetraploideum* is derived from *A. cyathophorum*, *A. farreri* and at least one extinct
693 lineage.

694 However, it should be noted that the fast evolution of repetitive sequences at 5S
695 and 45S rDNA loci and all chromosomes after the origin of the polyploid species may
696 distort such inference of origin when GISH and FISH results are employed. For
697 example, in our FISH results, we failed to detect the 5S rDNA loci of *A.*
698 *cyathophorum* in *A. tetraploideum* (Figure 4I) in contrast to GISH results and
699 phylogenomic analyses. However, we detected one more pair of 5S rDNA loci of *A.*
700 *farreri* in *A. tetraploideum* (Figure 4J). Despite this, our reciprocal GISH experiments
701 (Figure S14) between diploid species indicated eight chromosomes of the tetraploid
702 (six four-fluorochome, 1–6 pale-rose, plus four two-fluorochome, 7–10 yellow, Fig, 4
703 A) might have been derived from those with the ancestrally shared repeats of the
704 extant diploids. These results seem to indicate that the diploids *A. cyathophorum* and
705 *A. farreri* must have been involved in the origin of *A. tetraploideum* as suggested by
706 phylogenomic data. However, it remains unknown whether fast evolution of these
707 repetitive sequences or the addition of chromosomes from the extinct lineage was
708 responsible for the observed inconsistent GISH and FISH results in *A. tetraploideum.*
709 Such a distinction would require further investigations, especially based on the whole
710 genome sequences of all related species.
711

712 **Two allopolyploidizations, one homoploid hybrid speciation event, and three**
713 **extinctions?**

714 The possibility that *A. tetraploideum* is a tetraploid derived from three or more diploid
715 ancestors implies that its origin might have involved both polyploidization and
716 homoploid hybrid speciation (HHS) events. Our coalescence simulations indicated
717 that the simplest pathway for deriving *A. tetraploideum* from three ancestors would be
718 for an earlier diverging, extinct diploid to first formed allopolyploids independently
719 with each of *A. farreri* and *A. cyathophorum*, which is plausible because *Tragopogon*
720 *dubius* did this with *T. pratensis* and *T. porrifolius* to respectively form allopolyploids
721 *T. mirus* and *T. miscellus* (Lipman, Chester, Soltis, & Soltis, 2013). Those two
722 allotetraploids would then have crossed to produce *A. tetraploideum* by HHS (Figure
723 3F), as the formation of *T. mirus* × *T. miscellus* (Lipman, Chester, Soltis, & Soltis,
724 2013). In addition, this scenario would involve only two genomic divergence events:
725 first the extinct lineage from the others (Ks = 0.066–0.079), and then *A.*
726 *cyathophorum* from *A. farreri* (Ks = 0.036–0.041). The final hybrid speciation event,
727 at the homoploid level to generate *A. tetraploideum*, would create an additional peak
728 (Ks = 0.025–0.026; Figure 3F) because the genome of the ghost progenitor might
729 have diverged in each allotetraploid parents via subfunctionalisation or
730 neofunctionalization (Adams & Wendel, 2005; Doebley, Gaut, & Smith, 2006; Mayer
731 et al., 2014).

732 However, the youngest Ks peak in *A. tetraploideum* might instead reflect the
733 recent transfer of genetic material in either direction with *A. cyaphophorum*,
734 following which the newly transferred germplasm would begin to diverge anew from
735 the donor species. For enough material to be transferred to produce a Ks peak, this
736 event might be best viewed as a third round of hybrid speciation, meaning this
737 hypothesis still involves three extinct intermediaries. It is more complex than that
738 above, nevertheless it cannot be rejected. These are by no means the only possible
739 routes of origin for *A. tetraploideum*; other equally or more complex origins are still
740 possible (Figures S13, S15 and S16), involving >1 extinct diploid, >3 hybrid
741 speciation events and/or >1 autopolyploidization events (Figure S16). For example,
742 hypotheses involving two extinct lineages, one of them forming an allopolyploid with
743 *A. cyathophorum* and the other with *A. farreri* before the two products formed *A.*
744 *tetraploideum* by HHS, equally fits the data (Figure S15). Likewise, we cannot rule
745 out introgression between any extant or extinct lineages occurring during the process
746 as inferred by our analyses (Figure S13).
747 Difficulty in pairing between chromosomes of different species in hybrids can
748 cause post-zygotic breeding barriers and hence reproductive isolation (RI) (e.g., Ji,
749 Pertuzé, & Chetelat, 2004; Shibata & Hizume, 2002). HHS tends to occur where the
750 parent species are more closely related and this barrier, therefore, becomes less strong
751 allowing chromosomes to pair successfully with their interspecific counterparts (Lee,
752 Chang, & Chung, 2011). Conversely, allopolyploidization is more likely when RI is

753 greater, and chromosomes do not pair well, because following allopolyploidization,
754 chromosomes from each parent pair with their homolog from the other parent, with
755 likely diploid-like meiotic behavior and disomic inheritance (Bian et al., 2018; Lloyd &
756 Bomblies, 2016; Sybenga, 1996). Such a process has been demonstrated before
757 (Lipman et al., 2013; Lloyd & Bomblies, 2016) and under this scenario, *A.*
758 *tetraploideum* would have had strong RI from its alllotetraploid parents almost
759 immediately, because such disomic shifts may quickly occur after hybrid formation
760 This might have resulted in an incomplete haploid chromosomes originating from *A.*
761 *cyathophorum* and/or from *A. farreri*, under which homologous chromosome pairing
762 from any back-cross in either direction would have been broken.

763 All possible scenarios of the origin for *A. tetraploideum* involve multiple
764 instances of hybrid speciation, polyploidy, and extinction of intermediary taxa (Figure
765 S16), but our data can only inform us further about the first two; extinct lineages
766 leave behind no trace of why they disappeared. Likely reasons involve each ghost
767 lineage being outcompeted by its descendants, and/or failing to adapt to geological
768 changes and the strong Pleistocene climatic oscillations in the high-altitude QTP (Liu,
769 Sun, Ge, Gao, & Qiu, 2012). Conversely, the success of *A. tetraploideum* might reflect
770 its fitness advantage, aided by transgressive traits (Soltis et al., 2016), and epistatic
771 benefits from combining genomes (Adams & Wendel, 2005; Fawcett, Maere, & Van
772 De Peer, 2009; Van De Peer, Mizrachi, & Marchal, 2017). However, in lineages that
773 lack a fossil record, extinction events are often impossible to detect, especially as recent

774 research shows that birth-death models for phylogenies cannot distinguish between
775 scenarios with high, low or zero extinction rates (Louca & Pennell, 2020). Our results
776 seem to tentatively demonstrate that unravelling the origin of complex allopolyploids
777 provides an alternative means of demonstrating that lineages might have existed then
778 gone extinct. By combining multiple lines of evidence, our results showed that the
779 small subgenus *Cyathophora* seems to have once contained three additional extinct
780 lineages as well as five living ones (Figure 3F); hence ≥37.5% of its lineage diversity
781 might be extinct. In the other documented case of a species (e.g., *Paeonia officinalis*)
782 derived by HHS from allopolyploid parents, only one extinct ancestor (a diploid) could
783 be confidently inferred (Ferguson & Sang, 2001). Similar phylogenomic and
784 population genetic analyses presented here may shed light on complex origins of the
785 other genera with polyploids and/or individual species derived by allopolyploidization
786 events. For now, we can say that hybrid species, especially those with complex
787 histories, might provide another likely means besides fossils for generating possible
788 evidence for the existence of now extinct lineages.

789

790 **ACKNOWLEDGMENTS**

791 We thank Drs. Ying Wu (Northeast Normal University), Bo Liu and Quanwen Dou
792 (Northwest Institute of Plateau Biology, Chinese Academy of Science) for helping in
793 GISH experiments. Financial support was equally provided by Strategic Priority
794 Research Program of Chinese Academy of Sciences (XDB31010300) and further by

795 the National Natural Science Foundation of China (grant numbers 31901074 and
796 31590821). We thank the Supercomputing Center of Lanzhou University for
797 computation supporting.

798 AUTHOR CONTRIBUTIONS

799 J.L. designed the research. M.L., Z.Z., J.L., Y.Y., G.R., F.R., S.Z., X.D., and T.M.
800 carried out fieldwork, experiments and data analyses. J.L., M.L. and R.M. wrote and
801 revised the manuscript. M.L and Z.Z. contributed equally.

802 DATA ACCESSIBILITY STATEMENT

803 Raw RNA-Seq reads (under BioProject PRJNA598096), complete chloroplast
804 genomes (accession numbers: MN882559–MN882564), and sanger gene sequences
805 (accession numbers: MT479228–MT481754) have been deposited at NCBI, which is
806 publicly accessible. Raw RNA-Seq reads (accession number: CRA004506) and
807 SLAF-seq (accession number: CRA004214) have been archived in BIG Data Center
808 (https://ngdc.cncb.ac.cn/gsub/submit/gsa/list) under BioProject PRJCA005098.

809 CODE ACCESSIBILITY STATEMENT

810 The custom scripts have deposited in GitHub (https://github.com/
811 liminjie1234/Complex-evolutionary-origin-of-one-tetraploid-allium-in-the-Qinghai-T
812 ibet-Plateau).

813 Reference

814 Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J. E., Bierne, N., … Zinner,
815 D. (2013). Hybridization and speciation. *Journal of Evolutionary Biology*, *26*(2),
816 229–246. https://doi.org/10.1111/j.1420-9101.2012.02599.x
817 Abbott, R. J., Barton, N. H., & Good, J. M. (2016). Genomics of hybridization and its
818 evolutionary consequences. *Molecular Ecology*, *25*(11), 2325–2332.
819 https://doi.org/10.1111/mec.13685

820 Adams, K. L., & Wendel, J. F. (2005). Polyploidy and genome evolution in plants.
821 *Current Opinion in Plant Biology*, *8*(2), 135–141.
822 https://doi.org/10.1016/j.pbi.2005.01.001
823 Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm
824 for individual ancestry estimation. *BMC Bioinformatics*, *12*(1), 246.
825 https://doi.org/10.1186/1471-2105-12-246
826 Anderson, E. & Stebbins, G. L. (1954). Hybridization as an evolutionary stimulus.
827 *Evolution, 8*(4), 378–388.
828 Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson,
829 A. T., … Villalobos, F. (2011). The crucial role of the accessible area in ecological
830 niche modeling and species distribution modeling. *Ecological Modelling*,
831 *222*(2011), 1810–1819. https://doi.org/10.1016/j.ecolmodel.2011.02.011
832 Benaglia, T., Chauveau, D., Hunter, D. R., & Young, D. S. (2009). Mixtools: an R
833 package for analyzing finite mixture models. *Journal of Statistical Software*,
834 *32*(6), 1–29. https://doi.org/10.18637/jss.v032.i06
835 Bian, Y., Yang, C. W., Ou, X. F., Zhang, Z. B., Wang, B., Ma, W. W., … Liu, B.
836 (2018). Meiotic chromosome stability of a newly formed allohexaploid wheat is
837 facilitated by selection under abiotic stress as a spandrel. *New Phytologist*, *220*(1),
838 262–277. https://doi.org/10.1111/nph.15267
839 Bouckaert, R. R. (2010). DensiTree: making sense of sets of phylogenetic trees.
840 *Bioinformatics*, *26*(10), 1372–1373.
841 https://doi.org/10.1093/bioinformatics/btq110
842 Burland, T. G. (2000). *DNASTAR's Lasergene Sequence Analysis Software*. Totowa,
843 New Jersey: Humana Press, 71–91.
844 Chaudhuri, P., & Marron, J. S. (1999). SiZer for exploration of structures in curves.
845 *Journal of the American Statistical Association*, *94*(447), 807–823.
846 https://doi.org/10.1080/01621459.1999.10474186
847 Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., …
848 Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15),
849 2156–2158. https://doi.org/10.1093/bioinformatics/btr330
850 Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., …
851 Daly, M. J. (2011). A framework for variation discovery and genotyping using
852 next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–501.
853 https://doi.org/10.1038/ng.806
854 Doebley, J. F., Gaut, B. S., & Smith, B. D. (2006). The molecular genetics of crop
855 domestication. *Cell*, *127*(7), 1309–1321.
856 https://doi.org/10.1016/j.cell.2006.12.006
857 Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013).
858 Robust demographic inference from genomic and SNP data. *PLoS Genetics*,
859 *9*(10), e1003905. https://doi.org/10.1371/journal.pgen.1003905
860 Fawcett, J. A., Maere, S., & Van De Peer, Y. (2009). Plants with double genomes might
861 have had a better chance to survive the Cretaceous-Tertiary extinction event.

862 *Proceedings of the National Academy of Sciences of the United States of America*,
863 *106*(14), 5737–5742. https://doi.org/10.1073/pnas.0900906106
864 Ferguson, D., & Sang, T. (2001). Speciation through homoploid hybridization between
865 allotetraploids in peonies (*Paeonia*). *Proceedings of the National Academy of*
866 *Sciences of the United States of America*, *98*(7), 3915–3919.
867 https://doi.org/10.1073/pnas.061288698
868 Fredotović, Ž., Šamanić, I., Weiss-Schneeweiss, H., Kamenjarin, J., Jang, T. S., &
869 Puizina, J. (2014) Triparental origin of triploid onion, *Allium* × *cornutum*
870 (Clementi ex Visiani, 1842), as evidenced by molecular, phylogenetic and
871 cytogenetic analyses. *BMC Plant Bioogy*, *14*, 24.
872 https://doi.org/10.1186/1471-2229-14-24.
873 Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering
874 the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152.
875 https://doi.org/10.1093/bioinformatics/bts565
876 Guan, C. F., Liu, S. Y., Wang, M. K., Ji, H., Ruan, X. F., Wang, R. Z., & Yang, Y.
877 (2019). Comparative transcriptomic analysis reveals genetic divergence and
878 domestication genes in Diospyros. *BMC Plant Biology*, *19*(1), 227.
879 https://doi.org/10.1186/s12870-019-1839-2
880 Guggisberg, A., Mansion, G., Kelso, S., & Conti, E. (2006). Evolution of
881 biogeographic patterns, ploidy levels, and breeding systems in a diploid-polyploid
882 species complex of *Primula. New Phytologist*, *171*(3), 617–632.
883 https://doi.org/10.1111/j.1469-8137.2006.01722.x
884 Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O.
885 (2010). New algorithms and methods to estimate maximum-likelihood
886 phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, *59*(3),
887 307–321. https://doi.org/10.1093/sysbio/syq010
888 Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high
889 resolution interpolated climate surfaces for global land areas. *International*
890 *Journal of Climatology*, *25*(15), 1965–1978. https://doi.org/10.1002/joc.1276
891 Huang, D. I., & Cronk, Q. C. B. (2015). Plann: a command-line application for
892 annotating plastome sequences. *Applications in Plant Sciences*, *3*(8), 1500026.
893 https://doi.org/10.3732/apps.1500026
894 Huang, D. Q., Yang, J. T., Zhou, C. J., Zhou, S. D., & He, X. J. (2014). Phylogenetic
895 reappraisal of *Allium* subgenus *Cyathophor*a (Amaryllidaceae) and related taxa,
896 with a proposal of two new sections. *Journal of Plant Research*, *127*(2), 275–286.
897 https://doi.org/10.1007/s10265-013-0617-8
898 Jakob, S. S., & Blattner, F. R. (2010). Two extinct diploid progenitors were involved in
899 allopolyploid formation in the *Hordeum murinum* (Poaceae: Triticeae) taxon
900 complex. *Molecular Phylogenetics and Evolution*, *55*(2), 650–659.
901 https://doi.org/10.1016/j.ympev.2009.10.021
902 Ji, Y., Pertuzé, R., & Chetelat, R. T. (2004). Genome differentiation by GISH in
903 interspecific and intergeneric hybrids of tomato and related nightshades.

904 *Chromosome Research*, *12*(2), 107–116.
905 https://doi.org/10.1023/B:CHRO.0000013162.33200.61
906 Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software
907 version 7: improvements in performance and usability. *Molecular Biology and*
908 *Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010
909 Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., …
910 Drummond, A. (2012). Geneious basic: an integrated and extendable desktop
911 software platform for the organization and analysis of sequence data.
912 *Bioinformatics*, *28*(12), 1647–1649.
913 https://doi.org/10.1093/bioinformatics/bts199
914 Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: analysis of next
915 generation sequencing data. *BMC Bioinformatics*, *15*(1), 1–13.
916 https://doi.org/10.1186/s12859-014-0356-4
917 Lamichhaney, S., Han, F., Webster, M. T., Andersson, L., Grant, B. R., Grant, P.R.
918 (2018). Rapid hybrid speciation in Darwin's finches. *Science*, *359*(6372),
919 224–228. https://doi.org/10.1126/science.aao4593.
920 Lee, Y. I., Chang, F. C., & Chung, M. C. (2011). Chromosome pairing affinities in
921 interspecific hybrids reflect phylogenetic distances among lady’s slipper orchids
922 (Paphiopedilum). *Annals of Botany*, *108*(1), 113–121.
923 https://doi.org/10.1093/aob/mcr114
924 Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with
925 Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760.
926 https://doi.org/10.1093/bioinformatics/btp324
927 Li, L., Stoeckert, C. J. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog
928 groups for eukaryotic genomes. *Genome Research*, *13*(9), 2178–2189.
929 https://doi.org/10.1101/gr.1224503.candidates
930 Li, M. J., Guo, X. L., Li, J., Zhou, S. D., Liu, Q., & He, X. J. (2017). Cytotaxonomy of
931 *Allium* (Amaryllidaceae) subgenera *Cyathophora* and *Amerallium* sect.
932 *Bromatorrhiza*. *Phytotaxa*, *331*(2), 185–198.
933 https://doi.org/10.11646/phytotaxa.331.2.3
934 Li, M. J., Liu, J. Q., Guo, X. L., Xiao, Q. Y., & He, X. J. (2019). Taxonomic revision of
935 *Allium cyathophorum* (Amaryllidaceae). *Phytotaxa*, *415*(4), 240–246.
936 https://doi.org/10.11646/phytotaxa.415.4.9
937 Li, M. J., Tan, J. B., Xie, D. F., Huang, D. Q., Gao, Y. D., & He, X. J. (2016). Revisiting
938 the evolutionary events in *Allium* subgenus *Cyathophora* (Amaryllidaceae):
939 insights into the effect of the Hengduan Mountains Region (HMR) uplift and
940 Quaternary climatic fluctuations to the environmental changes in the
941 Qinghai-Tibet Plateau. *Molecular Phylogenetics and Evolution*, *94*(Part B),
942 802–813. https://doi.org/10.1016/j.ympev.2015.10.002
943 Li, M. J., Xie, D. F., Xie, C., Deng, Y. Q., Zhong, Y., Yu, Y., & He, X. J. (2019). A
944 phytogeographic divide along the 500 mm isohyet in the Qinghai-Tibet Plateau:

945 insights from the phylogeographic evidence of Chinese *Allium*s (Amaryllidaceae).
946 *Frontiers in Plant Science*, *10*, 149. https://doi.org/10.3389/fpls.2019.00149
947 Liang, Q., Xu, X., Mao, K., Wang, M., Wang, K., Xi, Z., & Liu, J. (2018). Shifts in
948 plant distributions in response to climate warming in a biodiversity hotspot, the
949 Hengduan Mountains. *Journal of Biogeography*, *45*(6), 1334–1344.
950 https://doi.org/10.1111/jbi.13229
951 Lipman, M. J., Chester, M., Soltis, P. S., & Soltis, D. E. (2013). Natural hybrids
952 between *Tragopogon mirus* and T. miscellus (Asteraceae): a new perspective on
953 karyotypic changes following hybridization at the polyploid level. *American*
954 *Journal of Botany*, *100*(10), 2016–2022. https://doi.org/10.3732/ajb.1300036
955 Liu, J. Q., Sun, Y. S., Ge, X. J., Gao, L. M., & Qiu, Y. X. (2012). Phylogeographic
956 studies of plants in China: avances in the past and directions in the future. *Journal*
957 *of Systematics and Evolution*, *50*(4), 267–275.
958 https://doi.org/10.1111/j.1759-6831.2012.00214.x
959 Lloyd, A., & Bomblies, K. (2016). Meiosis in autopolyploid and allopolyploid
960 Arabidopsis. *Current Opinion in Plant Biology*, *30*, 116–122.
961 https://doi.org/10.1016/j.pbi.2016.02.004
962 Louca, S., & Pennell, M. W. (2020). Extant time trees are consistent with a myriad of
963 diversification histories. *Nature*, *580*(7804), 502–505.
964 https://doi.org/10.1038/s41586-020-2176-1
965 Luo, X., Hu, Q., Zhou, P., Zhang, D., Wang, Q., Abbott, R. J., & Liu, J. (2017). Chasing
966 ghosts: allopolyploid origin of *Oxyria sinensis* (Polygonaceae) from its only
967 diploid congener and an unknown ancestor. *Molecular Ecology*, *26*(11),
968 3037–3049. https://doi.org/10.1111/mec.14097
969 Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of
970 duplicate genes. *Science*, *290*(5494), 1151–1155.
971 https://doi.org/10.1126/science.290.5494.1151
972 Mayer, K. F. X., Marcussen, T., Sandve, S. R., Heier, L., Pfeifer, M., Kugler, K. G., …
973 Olsen, O. A. (2014). A chromosome-based draft sequence of the hexaploid bread
974 wheat (*Triticum aestivum*) genome. *Science*, *345*(6194), 287–296.
975 https://doi.org/10.1126/science.1251788
976 Mittelbach, G. G., Schemske, D. W., Cornell, H. V., Allen, A. P., Brown, J. M., Bush,
977 M. B., … Turelli, M. (2007). Evolution and the latitudinal diversity gradient:
978 speciation, extinction and biogeography. *Ecology Letters*, *10*(4), 315–331.
979 https://doi.org/10.1111/j.1461-0248.2007.01020.x
980 Ossowski, S., Schneeberger, K., Lucas-Lledó, J. I., Warthmann, N., Clark, R. M.,
981 Shaw, R. G., … Lynch, M. (2010). The rate and molecular spectrum of
982 spontaneous mutations in *Arabidopsis thaliana*. *Science*, *327*(5961), 92–94.
983 https://doi.org/10.1126/science.1180677
984 Quental, T. B., & Marshall, C. R. (2010). Diversity dynamics: molecular phylogenies
985 need the fossil record. *Trends in Ecology and Evolution*, *25*(8), 434–441.
986 https://doi.org/10.1016/j.tree.2010.05.002

987 Rabier, C. E., Ta, T., & Ané, C. (2014). Detecting and locating whole genome
988 duplications on a phylogeny: a probabilistic approach. *Molecular Biology and*
989 *Evolution*, *31*(3), 750–762. https://doi.org/10.1093/molbev/mst263
990 Rieseberg, L. H., Van Fossen, C., & Desrochers, A. M. (1995). Hybrid speciation
991 accompanied by genomic reorganization in wild sunflowers. *Nature*, *375*(6529),
992 313–316. https://doi.org/10.1038/375313a0
993 Roelofs, D., Van Velzen, J., Kuperus, P., & Bachmann, K. (1997). Molecular evidence
994 for an extinct parent of the tetraploid species *Microseris acuminata* and *M.*
995 *campestris* (Asteraceae, Lactuceae). *Molecular Ecology*, *6*(7), 641–649.
996 https://doi.org/10.1046/j.1365-294X.1997.00225.x
997 Schoener, T. W. (1970). Nonsynchronous spatial overlap of lizards in patchy jabitats.
998 *Ecology*, *51*(3), 408–418. Https://doi.org/10.2307/1935376
999 Schumer, M., Rosenthal, G. G., & Andolfatto, P. (2018). What do we mean when we
1000 talk about hybrid speciation? *Heredity*, *120*(4), 379–382.
1001 https://doi.org/10.1038/s41437-017-0036-z
1002 Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). Mclust 5: clustering,
1003 classification and density estimation using Gaussian finite mixture models. *R*
1004 *Journal*, *8*(1), 289–317. https://doi.org/10.32614/rj-2016-021
1005 Shibata, F., & Hizume, M. (2002). The identification and analysis of the sequences that
1006 allow the detection of *Allium cepa* chromosomes by GISH in the allodiploid *A.*
1007 *wakegi*. *Chromosoma*, *111*(3), 184–191.
1008 https://doi.org/10.1007/s00412-002-0197-1
1009 Soltis, D. E., Visger, C. J., Blaine Marchant, D., & Soltis, P. S. (2016). Polyploidy:
1010 pitfalls and paths to a paradigm. *American Journal of Botany*, *103*(7), 1146–1166.
1011 https://doi.org/10.3732/ajb.1500501
1012 Sonnhammer, E. L. L., & Östlund, G. (2015). InParanoid 8: orthology analysis between
1013 273 proteomes, mostly eukaryotic. *Nucleic Acids Research*, *43*(D1), D234–D239.
1014 https://doi.org/10.1093/nar/gku1203
1015 Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and
1016 post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313.
1017 https://doi.org/10.1093/bioinformatics/btu033
1018 Sun, X., Zhu, S., Li, N., Cheng, Y., Zhao, J., Qiao, X., … Liu, T. (2020). A
1019 Chromosome-level genome assembly of garlic (*Allium sativum*) provides insights
1020 into genome evolution and allicin biosynthesis. *Molecular Plant*, *13*(9),
1021 1328–1339. https://doi.org/10.1016/j.molp.2020.07.019
1022 Sybenga, J. (1996). Chromosome pairing affinity and quadrivalent formation in
1023 polyploids: do segmental allopolyploids exist? *Genome*, *39*(6), 1176–1184.
1024 https://doi.org/10.1139/g96-148
1025 Taylor, S. A., & Larson, E. L. (2019). Insights from genomes into the evolutionary
1026 importance and prevalence of hybridization in nature. *Nature Ecology &*
1027 *Evolution*, *3*(2), 170–177. https://doi.org/10.1038/s41559-018-0777-y

1028 Than, C., Ruths, D., & Nakhleh, L. (2008). PhyloNet: a software package for analyzing
1029 and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, *9*,
1030 322. https://doi.org/10.1186/1471-2105-9-322
1031 Van De Peer, Y., Mizrachi, E., & Marchal, K. (2017). The evolutionary significance of
1032 polyploidy. *Nature Reviews Genetics*, *18*(7), 411–424.
1033 https://doi.org/10.1038/nrg.2017.26
1034 Wang, M., Zhang, L., Zhang, Z., Li, M., Wang, D., Zhang, X., … Ma, T. (2020).
1035 Phylogenomics of the genus Populus reveals extensive interspecific gene flow and
1036 balancing selection. *New Phytologist*, *225*(3), 1370–1382.
1037 https://doi.org/10.1111/nph.16215
1038 Wang, Z. F., Jiang, Y. Z., Bi, H., Lu, Z. Q., Ma, Y. Z., Yang, X. Y., ... Liu, J.Q. (2021).
1039 Hybrid speciation via inheritance of alternate alleles of parental isolating genes.
1040 *Molecular Plant*, *14*(2), 208–222. https://doi.org/10.1016/j.molp.2020.11.008
1041 Warren, D. L., Glor, R. E., & Turelli, M. (2008). Environmental niche equivalency
1042 versus conservatism: quantitative approaches to niche evolution. *Evolution*,
1043 *62*(11), 2868–2883. https://doi.org/10.1111/j.1558-5646.2008.00482.x
1044 Wu, L. L., Cui, X. K., Milne, R. I., Sun, Y. S., & Liu, J. Q. (2010). Multiple
1045 autopolyploidizations and range expansion of Allium przewalskianum Regel.
1046 (Alliaceae) in the Qinghai-Tibetan Plateau. *Molecular Ecology*, *19*(8),
1047 1691–1704. https://doi.org/10.1111/j.1365-294X.2010.04613.x
1048 Yakimowski, S. B., & Rieseberg, L. H. (2014). The role of homoploid hybridization in
1049 evolution: a century of studies synthesizing genetics and ecology. *American*
1050 *Journal of Botany*, *101*(8), 1247–1258. https://doi.org/10.3732/ajb.1400201
1051 Yang, Y. (2010). Cytotype distribution and polyploid origins of *Allium mairei* Lévl.
1052 (Alliaceae). Kunming Institute of Botany, Chinese Academy of Sciences, 1–140.
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064

1065 **Figure Legend**

1066

1067 **Figure 1 Distribution, morphology and phylogenetic relationship of *Allium*
1068 *tetraploideum*.** (A-B) Distribution and morphology of *A. tetraploideum, A.*
1069 *cyathohorum*, and *A. farreri* in the Qinghai-Tibet Plateau (QTP). (C–D) The
1070 significantly larger stomata and seeds of *A. tetraploideum* compared to *A.*
1071 *cyathophorum* and *A. farreri*. (E) Phylogenetic relationship of *A. tetraploideum* in the
1072 genus *Allium* based on chloroplast genomes and transcriptomes respectively. *
1073 indicates branches with bootstraps <70%; all others have ≥95% support.

1074

1075 **Figure 2 Phylogenetic relationships and polyploid origin of *Allium tetraploideum*
1076 inferred from chloroplast genomes and transcriptomes.** (A) A phylogenetic tree of
1077 subgenus *Cytahophora* based on chloroplast genomes, with shared single nucleotide
1078 polymorphisms (SNPs) between *A. tetraploideum* and the others are indicated, as well
1079 as those between *A. caythophorum* and the other diploid species. (B) The integrated
1080 single copy gene (SCG) trees with bootstrap supporting values >70%. (C)
1081 PhyloNet-based network for *A. tetraploideum* and two potential parents based on the
1082 540 SCG trees and 807 SCG trees. (D–E) Tree-topology statistics when examining
1083 different species (as indicated in (B)) of the subgenus with different homologs of *A.*
1084 *tetraploideum* and SCG of the diploids using Twisst. (F) Integrated density plots of Ks
1085 values of paralogs of *A. tetraploideum* and homologs of *A. tetraploideum* - *A.*
1086 *cyathophorum* (2:1) and *A. tetraploideum* - *A. farreri* (2:1). Ks peak (0.042) between *A.*
1087 *cyathophorum* and *A. farreri* is indicated in blue. The two grey rectangles indicate two
1088 shared Ks peaks (0.066–0.107 and 0.036–0.049) among all compared Ks plots. A third
1089 Ks peak only found in *A. tetraploideum* is at 0.025–0.026.

1090

1091 **Figure 3 Evolutionary origin of *A. tetraploideum* based on population genetic
1092 analyses of Sanger sequences at multiple loci and SLAF-seq data.** (A)
1093 Maximum-likelihood (ML) trees of 10 nuclear genes. (B) ML tree for subg.
1094 *Cyathophora* using SLAF-seq SNPs. The colors of lines in A and B denote the same
1095 species. The numbers at nodes indicate bootstrap support values. (C) Genetic
1096 differentiation (F_{ST}) within each species and between each of the three species pairs
1097 based on SLAF-seq SNPs. (D) PCA plot based on SLAF-seq SNPs among all
1098 individuals of *A. cyathophorum*, *A. farreri* and *A. tetraploideum*, with the first two axes
1099 (PC1 and PC2) explaining 10.43% and 4.46% of the variation, respectively. (E)
1100 Population structure of the three species based on SLAF-seq SNPs with $K = 3$. The bars
1101 represent the sampled individuals, and the colors of bars indicate assignment
1102 probabilities. (F) The best-fitting evolutionary model among *A. cyathophorum* (CC), *A.*
1103 *farreri* (FF), and *A. tetraploideum* (CFGG) using *Fastsimcoal2*, with the participation
1104 of a ghost lineage (GG). The inferred split ages are shown in circles in units of millions

1105 of years before present. The red and green labels indicate the extant *A. cyathophorum*
1106 and *A. farreri*, and blue label and red "X" indicate the extinct lineages.
1107
1108 **Figure 4. GISH and FISH analyses on the mitotic chromosomes of *A.***
1109 ***tetraploideum* (2n = 4x = 32) and related extant diploids.** (A) Four-fluorochrome
1110 GISH with gDNAs of the four extant diploid species (2n = 2x = 16) probed onto the
1111 chromosomes of *A. tetraploideum*, showing 22 chromosomes stained: chromosomes
1112 1–6 (pale-rose) = four diploid species; chromosomes 7–10 (yellow) = *A. cyathophorum*
1113 + *A. farreri*; chromosomes 11–16 (red) = *A. cyathophorum*, chromosomes 17–20 (green)
1114 = *A. farreri* and chromosomes 21–22 (greenish-yellow) = *A. spicatum*. (B–E) Separate
1115 monochromatic images from (B) *A. cyathophorum* (stained red), (C) *A. farreri* (stained
1116 green), (D) *A. spicatum* (stained greenish-yellow) and (E) *A. mairei* (stained rose). (F)
1117 Merging two-monochromatic images from *A. cyathophorum* and *A. farreri* with ten
1118 chromosomes (yellow) both stained by *A. cyathophorum* + *A. farreri*. (G–J) FISH with
1119 45S rDNA probes (green) and 5S rDNA probes (red) in (G) *A. cyathophorum*, (H) *A.*
1120 *farreri* and (I) *A. tetraploideum*, and (J) their karyotypes. Scale bar is in 10 µm.
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146

1147 **Supplementary Material**
1148 **Figure S1** Cytology and sequencing sampling of the species within subg.
1149 *Cyathophora*.
1150 **Figure S2** Gaussian mixture model (top) and SiZer results (bottom) for Ks
1151 distributions.
1152 **Figure S3** Schematics of the 15 different demographic models analyzed using
1153 Fastsimcoal2.
1154 **Figure S4** Molecular dating analysis of subg. *Cyathophora* with a constant mutation
1155 rate.
1156 **Figure S5** Niche estimations for *A. tetraploideum*, *A. cyathophorum* and *A. farreri*.
1157 **Figure S6** Density plots for *A. tetraploideum*, *A. cyathophorum* and *A. farreri* using
1158 each bioclimatic variable.
1159 **Figure S7** Comparisons of seed germination and seedling length.
1160 Figure **S8** Coalescence tree of *Allium* based on CDs extracted from transcriptomes.
1161 **Figure S9** Previous reports on the population genetics of subg. *Cyathophora*.
1162 **Figure S10** Hypothesized phylogenetic relationships between the two extant species
1163 and one extinct progenitor of *A. tetraploideum*.
1164 **Figure S11** Density plot of K_S values using R *mclust* package.
1165 **Figure S12** Phylogenies among the diploid species of subg. *Cyathophora* for ten genes.
1166 **Figure S13** The best fitting model for the origin of *A. tetraploideum*.
1167 **Figure S14** Reciprocal GISH patterns between the diploids of subg. *Cyathophora*.
1168 **Figure S15** Inferring chromosomal constitutions using GISH and FISH results.
1169 **Figure S16** Origin pathways of the tetraploid *A. tetraploideum* from two current
1170 species and at least one extinct diploid lineage.
1171 **Table S1** Sampling information.
1172 **Table S2** Detailed information for 47 vetted localities for *A. tetraploideum*, *A.*
1173 *cyathophorum* and *A. farreri*.
1174 **Table S3** Bioclimatic variables extracted from World-Clim data set.
1175 **Table S4** The principle components (PC) loading scores.
1176 **Table S5** GenBank accession numbers for 48 chloroplast genomes.
1177 **Table S6** GenBank accessions of transcriptomes used in this study.
1178 **Table S7** Bayesian Information Criterion (BIC) values to Ks distribution by *Mclust*.
1179 **Table S8** Primer information for 10 gene loci and their amplified protocols.
1180 **Table S9** ANOSIM and ADONIS2 tests for the three species along different factors.
1181 **Table S10** Comparisons of seed germination percentage and rate and seedling length
1182 for *A. tetraploideum*, *A. cyathophorum* and *A. farreri*.
1183 **Table S11** Tree topologies within the single copy gene trees.
1184 **Table S12** Genome networks using single copy gene trees in PhyloNet.
1185 **Table S13** The divergence time estimation based on the K_S peak values.
1186 **Table S14** Gaussian mixture modeling and SiZer results for K_S distributions by species
1187 to species.
1188 **Table S15** The best-fitting model parameters.