Discovery of a high-altitude ecotype and ancient lineage of *Arabidopsis thaliana* from Tibet

Liyan Zeng[1,2]†, Zhuoya Gu[1]†, Min Xu[3,4]†, Ning Zhao[3]†, Weidong Zhu[3], Takahiro Yonezawa[3,5], Tianmeng Liu[3], Lha Qiong[3], Tashi Tersing[3, 6], Lingli Xu[1], Yang Zhang[7], Rongyan Xu[1], Ningyu Sun[1], Yanyan Huang[1], Jiankun Lei[8], Liang Zhang[8], Feng Xie[9], Fang Zhang[10], Hongya Gu[11], Yupeng Geng[12], Masami Hasegawa[1,5], Ziheng Yang[13], M. James C. Crabbe[14,15], Fan Chen[10, *], Yang Zhong[1,3,16, *]

1. Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, School of Life Sciences, Fudan University, Shanghai 200433, China

2. Shanghai Public Health Clinical Center, Fudan University, Shanghai 201508, China

3. Institute of Biodiversity Science and Geobiology, College of Sciences, Tibet University, Lhasa 850000, China

4. Institute of Forest Inventory, Planning and Research of Tibet Autonomous Region, Lhasa 850010, China

5. Institute of Mathematical Statistics, Midori-cho 10-3, Tachikawa, Tokyo 190-8562, Japan

6. Tibet Museum of Natural Science, Lhasa 850000, China

7. Department of Bioengineering, University of Illinois at Urbana-Champaign, Champaign, IL, 61801, USA

8. School of Computer Science, Fudan University, Shanghai 200433, China

9. School of Urban Rail Transportation, Soochow University, Suzhou 215131, China

10. Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China

11. School of Life Sciences, Peking University, Beijing 100871, China

12. School of Life Sciences, Yunnan University, Kunming 650091, China

13. Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London WC1E 6BT, United Kingdom

14. Department of Zoology, University of Oxford, Tinbergen Building, South Parks Road, Oxford, OX1 3PS, United Kingdom

15. Institute of Biomedical and Environmental Science & Technology, Department of Life Sciences, University of Bedfordshire, Park Square, Luton, LU1 3JU, United Kingdom
16. Shanghai Center for Bioinformation Technology, Shanghai 201203, China

†These authors contributed equally to this work
*Correspondence and requests for the materials should be addressed to F. Chen (fchen@genetics.ac.cn) or Y. Zhong (yangzhong@fudan.edu.cn).

Arabidopsis thaliana has long been a model species for dicotyledon study, and was the first flowering plant to get its genome completed sequenced [1]. Although most wild *A. thaliana* are collected in Europe, several studies have found a rapid *A. thaliana* west-east expansion from Central Asia [2]. The Qinghai-Tibet Plateau (QTP) is close to Central Asia and known for its high altitude, unique environments and biodiversity[3]. However, no wild-type *A. thaliana* had been either discovered or sequenced from QTP. Studies on the *A. thaliana* populations collected under 2000 m asl have shown that the adaptive variations associated with climate and altitudinal gradients [4]. Hence a high-altitude *A. thaliana* provides a precious natural material to investigate the evolution and adaptation process.

Here, we present the genome of a new ecotype of *A. thaliana* collected in the Gongga County, Tibet (4200 m asl) (Fig. 1a), to demonstrate its evolutionary history and adaptation to high-altitude regions. The Tibetan samples were identified as *A. thaliana* by comparing the nuclear internal transcribed spacer (ITS), four chloroplast genes (*mat*K, *rbc*L, *rpo*B, and *rps*16), and three chloroplast intergenic spacers (IGS, *trn*L-*trn*F, and *trn*T-*trn*L) with *A. thaliana* (Col-0) and *A. lyrata* (Supplementary Fig. 1). This is the first report that an *A. thaliana* population has been collected in the QTP over 4000 m asl and identified by molecular analysis. Moreover, the new Tibetan ecotype (herein referred to as "Tibet-0") is diploid ($2n$=10) according to karyotype analysis of its pollen mother cells during meiosis (Supplementary Fig. 2 online), suggesting that the ploidy of the Tibet-0 is stable and capable of further sequence analysis.

We then conducted genome-wide resequencing of Tibet-0 with a mean coverage of 40x of the reference genomes Col-0 and TAIR10, by using Illumina Hiseq2000 (Supplementary Tables 5, 6, online). We compared Tibet-0 with 47 other *A. thaliana* ecotypes that have been genome-wide sequenced, and found that Tibet-0 was of high divergence, including a higher proportion of SNPs (Supplementary Tables 7-9, online). Evolutionary relationships between Tibet-0 and other ecotypes were evaluated by the following two independent approaches based on 5611 single-copy orthologues in 47 *A. thaliana* ecotypes including 26 relicts and 21 non-relicts defined by the 1001 Genomes Consortium [5]. The first approach is the phylogenetic method. The genealogy among the individuals was inferred based on the concatenated genomic data, and Tibet-0 was placed at the root of the *A. thaliana* populations with high support value (Fig. 1b) [5, 6]. It makes Tibet-0 the most ancestral lineage.

However, since this phylogenetic approach assumes that all gene loci have the same genealogy, coalescent method was also applied as a cross check [7]. In this method, 2788 single-copy orthologues were independently analyzed, and the distributions of the tMRCAs (the time to the most recent common ancestor) for these genes were estimated. If Tibet-0 is the most basal lineage among the *A. thaliana* populations and Tibet-0 specific alleles has generally older histories than others, tMRCAs excluding Tibet-0 will be smaller than tMRCAs of all *A. thaliana* populations. Otherwise, if there is no such genetic structure and Tibet-0 specific

alleles are included within the genetic diversity of other *A. thaliana* populations, tMRCAs excluding Tibet-0 will be equal to the tMRCAs of all *A. thaliana* populations. To examine the differences among the distributions, the tMRCAs were first estimated based on 48 *A. thaliana* (tMRCA48). Subsequently, each ecotype was excluded once, and the tMRCAs of 47 remaining *A. thaliana* were estimated (tMRCA47: there are 48 combinations of tMRCA47). Finally, the relative tMRCAs (tMRCA47/tMRCA48) were estimated. Fig. 1C illustrates the distributions of the relative tMRCAs. When Tibet-0 was excluded, the distribution of the relative tMRCAs ($tMRCA47_{excluding\ Tibet\text{-}0}$/tMRCA48) significantly shifted (t test, p = 9.77E-32), while the average of $tMRCA47_{excluding\ one\ ecotype\ other\ than\ Tibet\text{-}0}$/tMRCA48 showed no significant change (Fig. 1c). These findings confirm that Tibet-0 has the most ancestral positions among *A. thaliana* populations.

To understand the correlation between the evolution of *A. thaliana* and major geological events, especially Tibetan uplifts, the divergence time between Tibet-0 and other ecotypes were estimated. Since there is no suitable fossil calibrations within *A. thaliana*, the divergence time between *A. lyrata* and *A. thaliana* was estimated based on the genomic data in the framework of whole land plant evolution with reliable fossil records, and it was estimated to be about 9 million years ago (Fig. 1D, Supplementary Fig. 3). Then, the time of the common ancestor of *A. thaliana* was estimated by multiplying the divergence time between *A. lyrata* and *A. thaliana* and the ratio of the divergence time between *A. lyrata* and *A. thaliana*. The divergence time between Tibet-0 and other ecotypes was found to be 126 – 149 Ka (kili annum: thousand years ago). Interestingly, the Gonghe movement, which was the last phase of Tibetan uplift, isolated the Qinghai Lake and raised the QTP to its present height algo began at about 15 Ka [8]. Besides, the divergence time of Tibet-0 and other ecotypes is in the middle Pleistocene from 781 to 126 Ka.

A. thaliana has been widely used in studies of plant biology. By collecting and sequencing *A. thaliana* collected from the QTP over 4200 m asl, we have found that the Tibet-0 is a new and divergent ecotype that isolated from other *A. thaliana* ecotypes since the last uplift of the QTP. After 126 – 149 thousands years evolution in the extreme plateau environment, Tibet-0 possesses a distinctive genome with a high proportion of SNPs compared to other ecotypes. According to the strongly negatively skewed Tajima's D of 5611 single-copy orthologues, a recent selective sweep or population expansion might have occurred in the *A. thaliana*, which is consistent with previous studies (Supplementary Fig. 4, online) [10]. Considering the ancestral position of Tibetan populations as well as the subsequent selective sweep or population expansion, possibly in the Last Glacial Period, suggested by the negative Tajima's D, we suppose that some mutations might have emerged in the ancient *A. thaliana* population located around the QTP, and then spread to most other populations. Following step is investigating phenotypic traits of Tibet-0 to study the adaptive evolution of *A. thaliana* to high altitudes. As a new model plant, the Tibet-0 from QTP would provide an invaluable material for further study.

Author Contributions

F. C. and Y. Zhong conceived the project. L. Zeng, Z. G., T. Y., F. C. and Y. Zhong contributed to the design of the project and extensive discussions. M. X., N. Z., W. Z., L. Q. and T. T. collected samples from Tibet. L. Zeng and H. G. helped with sample identification. L. X., R. X., F. X., J. L., L. Z., Z. G., N. Z., Y. H., T. Y., M. H., F. Z., F. C., Y. G., L. Zhang, Y. Zhang, Z. Y., M. J. C. C. and Y. Zhong performed the common garden experiments, sequence analyses and evolutionary analyses. L. Zeng, Z. G., Y. Zhang, M. J. C. C., N. S., F. C. and Y. Zhong wrote the manuscript. Other authors revised the manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (91131901), the specimen platform of China (teaching specimens sub-platform) and PSCIRT project.

Availability of data and materials

The genomic DNA of Tibet-0 has been deposited in the Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/sra/) under accession number SRP052218.

Reference:

1. Arabidopsis Genome I, Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* (2000),**408**: 796-815.
2. Yin P, Kang J, He F, Qu LJ, Gu H, The origin of populations of Arabidopsis thaliana in China, based on the chloroplast DNA sequences. *BMC plant biology* (2010),**10**: 22.
3. Liu S WN, Duan K, Xiao C, Ding Y, Recent progress of glaciological studies in China. *Journal of Geographical Sciences* (2004),**14**: 401-10.
4. Suter L, Ruegg M, Zemp N, Hennig L, Widmer A, Gene regulatory variation mediates flowering responses to vernalization along an altitudinal gradient in Arabidopsis. *Plant physiology* (2014),**166**: 1928-42.
5. Consortium G, 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* (2016),**166**: 481-91.
6. Gan X *et al.*, Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* (2011),**477**: 419-23.
7. Nakagome S, Mano S, Hasegawa M, Comment on "Nuclear genomic sequences reveal that polar bears are an old and distinct bear lineage". *Science* (2013),**339**: 1522.
8. Li J, Late Cenozoic intensive uplift of Qinghai-Xizang Plateau and its impacts on environments in surrounding area. *Quaternary Science* (2001),**21**: 381-91.
9. Cohen KM GP, Global chronostratigraphical correlation table for the last 2.7 million years. *Subcommission on Quaternary Stratigraphy* (2011),**31**: 243-7.
10. Shimizu KK *et al.*, Darwinian selection on a selfing locus. *Science* (2004),**306**: 2081-4.