Zou et al. Genome Biology
DOI 10.1186/s13059-017-1378-9

Genome Biology

RESEARCH Open Access

Adaptation of *Arabidopsis thaliana* to the Yangtze River basin

Yu-Pan Zou[1,2†], Xing-Hui Hou[1,2†], Qiong Wu[1], Jia-Fu Chen[1,2], Zi-Wen Li[1], Ting-Shen Han[1,2], Xiao-Min Niu[1,2], Li Yang[1], Yong-Chao Xu[1,2], Jie Zhang[1,2], Fu-Min Zhang[1], Dunyan Tan[3], Zhixi Tian[4], Hongya Gu[5,6] and Ya-Long Guo[1,2*]

Abstract

Background: Organisms need to adapt to keep pace with a changing environment. Examining recent range expansion aids our understanding of how organisms evolve to overcome environmental constraints. However, how organisms adapt to climate changes is a crucial biological question that is still largely unanswered. The plant *Arabidopsis thaliana* is an excellent system to study this fundamental question. Its origin is in the Iberian Peninsula and North Africa, but it has spread to the Far East, including the most south-eastern edge of its native habitats, the Yangtze River basin, where the climate is very different.

Results: We sequenced 118 *A. thaliana* strains from the region surrounding the Yangtze River basin. We found that the Yangtze River basin population is a unique population and diverged about 61,409 years ago, with gene flows occurring at two different time points, followed by a population dispersion into the Yangtze River basin in the last few thousands of years. Positive selection analyses revealed that biological regulation processes, such as flowering time, immune and defense response processes could be correlated with the adaptation event. In particular, we found that the flowering time gene *SVP* has contributed to *A. thaliana* adaptation to the Yangtze River basin based on genetic mapping.

Conclusions: *A. thaliana* adapted to the Yangtze River basin habitat by promoting the onset of flowering, a finding that sheds light on how a species can adapt to locales with very different climates.

Keywords: *Arabidopsis thaliana*, Population genomics, Adaptation, Yangtze River basin

Background

Global climate change has a profound influence on human health, food security, and biological diversity as it greatly taxes the ability of organisms to adapt to new environments [1–3]. A fundamental biological question that has recently emerged concerns how best to resolve the mismatch between organisms and human-altered environments. To avoid the tremendous cost of phenotype-environment mismatch, it is important to understand how organisms adapt to new habitats. The understanding of adaptation in constant environments, such as in serpentine soil using plants, or in experimental evolution using microorganisms, has progressed steadily [4, 5]. However, the mechanisms through which adaptation proceeds in heterogeneous natural environments are largely unknown. One of the major challenges in this area is that the genetic basis of adaptation to climate change is largely unknown.

Here, we use the plant model species *Arabidopsis thaliana* to address this fundamental question in the context of its adaptation in natural environments. *A. thaliana* is widely distributed across the temperate region in the northern hemisphere, including the Yangtze River basin, a region that is distant from its origin place of Europe/North Africa [6–9]. At several geographic scales in its native Eurasian range, *A. thaliana* demonstrates evidence of local adaptation [9–16]. Therefore, *A. thaliana* is a good model system to understand the mechanism of adaptation in natural environments at a global level [13, 16–19].

* Correspondence: yalong.guo@ibcas.ac.cn
†Equal contributors
[1]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
Full list of author information is available at the end of the article

A. thaliana originated in Europe/North Africa [8, 9, 20, 21] and the Yangtze River basin is the most south-eastern edge of *A. thaliana*'s native habitats [22, 23]. The environment of the Yangtze River basin is tremendously different compared with both its origin in Europe/North Africa and other regions between the Yangtze River basin and Europe/North Africa where *A. thaliana* is found. Of the 19 climate variables (Additional file 1: Table S1), the temperature seasonality (bio4) and the annual precipitation (bio12) are the most differentiated climate variables among the different regions (Additional file 2: Figure S1). Therefore, it is of great interest to know how this species could adapt to the faraway south-eastern habitats with such distinct environments.

Selective sweep scans and quantitative genetics provide robust and efficient approaches to identify genetic variants correlated with adaptation [19, 24–26]. To understand how this model species could adapt to this region, we performed population genomics analyses and genetic mapping for flowering time variation, one of the most important life history traits correlated with fitness. We found that the Yangtze River *A. thaliana* population is unique and diverged 61,409 years ago from its ancestor population with two independent waves of gene flows afterwards; it expanded across the Yangtze River basin over thousands of years. Genes that correlated with biological regulation processes, such as flowering time, immune and defense response processes could have contributed to the adaptation of the Yangtze River population. Our results highlight how a plant species could adapt to a new climate.

Results

The Yangtze River population is unique

We sequenced 118 strains of *A. thaliana* across north-western China (mainly from the Altai Mountains) to south-eastern China along the Yangtze River (Fig. 1a and Additional file 3: Table S2). Each strain was sequenced to at least 18× coverage (average = 31.97×), which amounts to 3772.59 × coverage in total. From these genome sequences, we called 2.66 million single nucleotide polymorphisms (SNPs) and 0.58 million indels (Additional file 2: Figure S2), using the Col-0 strain as the reference genome. The SNPs called from the 118 strains sequenced in this study and SNPs extracted from 103 geographically representative genomes of the 1001 Genomes Project (Additional file 4: Table S3 for the detail) [10, 14, 27] were integrated together to represent the worldwide strains (Fig. 1a).

To explore the relationship among samples, admixture analysis, phylogenetic analysis, and principal component analysis (PCA) were conducted. These analyses suggested that these 221 strains, with some intermediate strains, could be divided into three major groups (eastern Asia, central Asia, and European/North Africa [hereafter referred to as Europe]), roughly consistent with their geographical origin (Fig. 1b–d). Phylogenetic analyses using two close relatives, *Arabidopsis lyrata* and *Capsella rubella*, as outgroups suggested that the Iberian Peninsula and North Africa strains are located at the basal position of the phylogenetic tree and confirmed that they are relicts [7, 9] (Additional file 2: Figure S3). A small number of strains from different geographical regions formed a clade, which most probably reflects relicts or recent introduction. For example, for those strains grouped with Europe/North Africa samples, three strains from south-western China (Tibet and Yunnan provinces) could be relicts, while strains from USA, Japan, and New Zealand that clustered with European sample could be recent introductions (Additional file 2: Figure S3). In the following analysis, we excluded the outlier strains that could disturb the local adaptation analysis, based on both phylogenetic and PCA results (Fig. 1d and Additional file 2: Figure S3). In this way, the final subsets included 86 strains from the Yangtze River basin (hereafter referred to as popY), 25 strains from north-western China (popN) to represent the central Asian population, and 67 strains from Europe/North Africa (popE) (Fig. 1d and e; Additional file 3: Table S2 and Additional file 4: Table S3). Simulation analyses suggested that the sample size we selected from the Yangtze River population is large enough to cover all the possible genetic variants (Additional file 2: Figure S4).

PopE has more SNPs, a total of 4,673,541, than either popY (n = 1,083,605) or popN (n = 975,715). PopE also has the highest number of private SNPs (n = 3,725,836) compared with popN (n = 273,787) and popY (n = 441,460). Furthermore, nucleotide diversity was highest in popE ($\pi = 6.09 \times 10^{-3}$), compared with popN ($2.78 \times 10^{-3}$) and popY ($2.08 \times 10^{-3}$) (Additional file 2: Figure S5). These results confirm that popE is the ancestral population [8, 9]. The *A. thaliana* samples that we studied make up three natural major groups, with popY from the Yangtze River basin being a uniform population.

The Yangtze River population was recently established

To clarify the genetic separation among populations of *A. thaliana*, we performed a multiple sequential Markovian coalescent (MSMC) analysis to estimate the relative cross coalescence rate [28]. By analyzing four haplotypes for each pair of populations, we found that all relative cross--coalescence rates between any two populations were similar and exhibited a gradual decline since the last glacial period (Fig. 2a). In contrast to the relative cross coalescence rates between popE and popN or popY, which completely diverged during the last glacial period, popN and popY diverged since then but with gene flow at two different periods, before separating completely about a few thousand years ago.

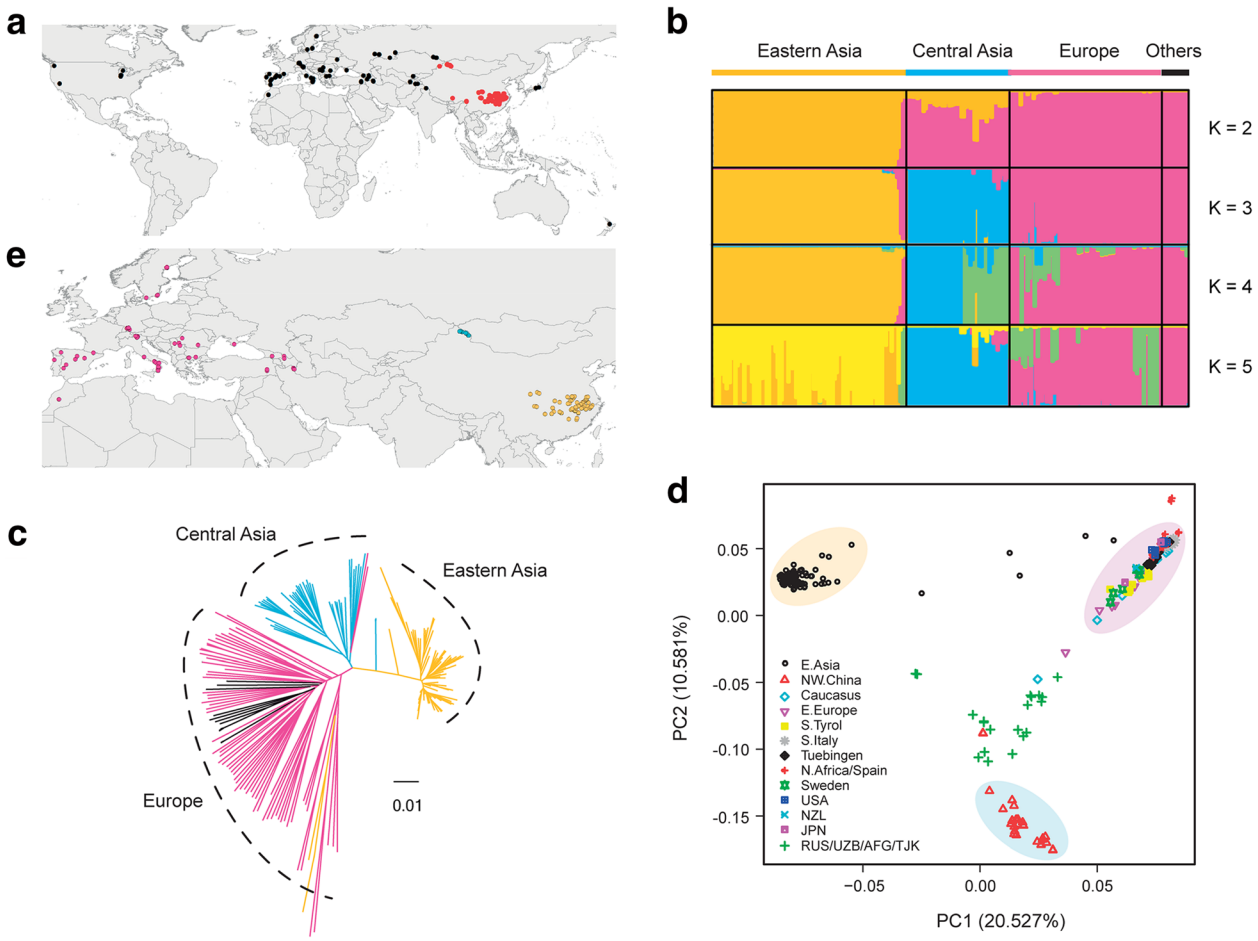

Fig. 1 Geographic locations and population structure of *A. thaliana*. **a** *Map* of the locations sampled (*black points*) and sequenced (*red points*) in this study. **b** Population structure of *A. thaliana* based on admixture analysis of strains from (**a**). "Others" indicates strains from USA, Japan, and New Zealand, which most probably reflects recent introduction given that *A. thaliana* originated in Europe. **c** Phylogenetic tree of *A. thaliana* strains from (**a**). *Black lines* indicate strains from USA, Japan, and New Zealand. **d** Principle component analysis (PCA) of *A. thaliana*. *Shaded areas* indicate the populations from different regions: *pink* refers to the European population (popE), *blue* refers to popN, and *yellow* refers to popY. **e** The locations of different populations color-coded as in (**d**)

To reflect the historical processes for the different populations, we calculated the distribution of coalescence times as conducted in a previous study [9]. Coalescence rate is an indication of relatedness, with higher ones indicating closer relationship and smaller population sizes. From the analysis of two haplotypes, the results suggested that, since the last glaciation, coalescence rates within popN and popY were much higher than that for popE; and coalescence rates between members of popN and popY were higher than those between popE and popN or popY (Fig. 2b).

Furthermore, we employed fastsimcoal2 [29] to infer the demographic history of the *A. thaliana* popN and popY populations, combining the findings with those of the aforementioned MSMC study. Four alternative models with different extents of gene flow and varying population sizes were investigated (Additional file 2: Figure S6). The best fit model had two waves of asymmetrical gene flow, which is consistent with the gene flow at two different periods in the MSMC analysis (Fig. 2a). Under the best model, popN and popY diverged 61,409 years ago from an ancient population of size 179,724 into sizes of 26,959 and 78,454, respectively (Fig. 2c, see Additional file 1: Table S4 for the detail). Gene flow existed at two time stages, between 18,652 and 14,307 years ago, and between 8440 and 7660 years ago, although both of these gene flow events were weak. Following that, since 7543 years ago, popY exhibited a notable expansion and reached the size of 45,089, and distributed across the Yangtze River basin, while popN went through a reduction to 45,617, about 7169 years ago.

Ecological niche modeling (ENM) based on the *A. thaliana* distribution information (Additional file 5: Table S5) indicates that there were widely suitable

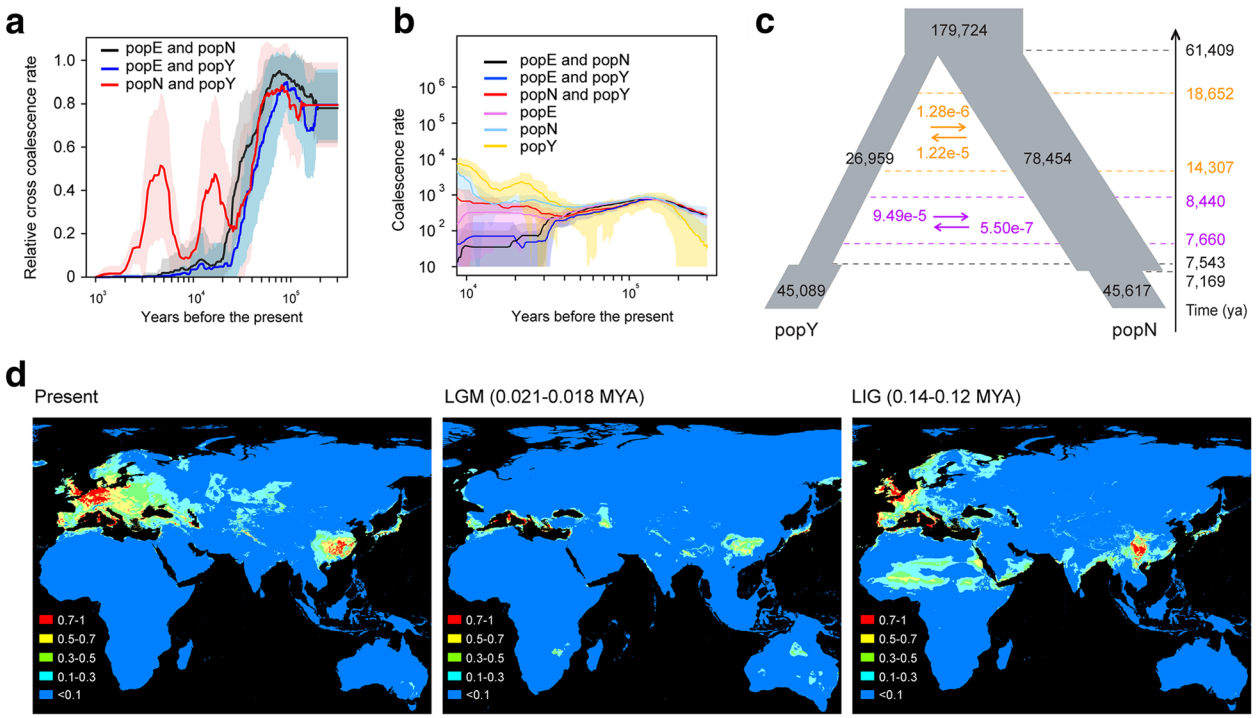

Fig. 2 Demographic history of *A. thaliana*. **a** Relative cross coalescence rate reveals the age and pace of divergence between two populations. The two populations are well-mixed if the relative cross coalescence rate is 1 and fully separated when the value is 0. *Solid lines* represent means and *shading* represents standard deviations (50 random samplings). **b** Coalescence rates for pairs of individuals within and between populations. **c** The best demographic model of the two populations of *A. thaliana*. The width of the *boxes* represents the relative effective population size and *arrows* represent the migration between popN and popY. **d** Predicated distributions of *A. thaliana* based on ecological niche modeling. Areas in different *colors* indicate the various possibilities (0–1) of suitable habitats for *A. thaliana*. LGM last glacial maximum, LIG last interglacial, MYA million years ago

habitats, roughly connected between the Yangtze River basin and the southern slopes of the Himalayas Mountains around the last interglacial period (Fig. 2d). This result revealed that the extant *A. thaliana* population of the Yangtze River basin could be derived from the eastward dispersion via the Himalayas, in agreement with previous proposals [22]. This observation is also supported by the phylogenetic results, in which samples from central Asia (including popN) are the most closely related lineage of popY (Additional file 2: Figure S3). In summary, we found that glacial cycle is one of the major determinants of the demographic history of *A. thaliana*. PopY diverged about 61,409 years ago from its ancestor and expanded across the Yangtze River basin thousands of years ago.

Pervasive selection and genomic signatures of local adaptation of the Yangtze River population

Abrupt geographical change in allele frequency is evidence of strong local adaptation [9]. To detect genes that are under positive selection and are important for adaptation, we searched the genomes for a selective sweep signal using a site frequency spectra (SFS)-based method (SweepFinder2) (Fig. 3) and a linkage disequilibrium (LD)-based method (OmegaPlus) (Additional file 2: Figure S7). The overlapped regions under selection between the two methods were regarded as the candidate regions of selection. In total, there were 530 protein-coding genes under positive selection (Fig. 3, see Additional file 6: Table S6 for the detail). These genes might have contributed to the adaptation of popY to the Yangtze River basin.

Gene Ontology (GO) analysis of the candidates under positive selection detected five significantly enriched biological process GO terms including immune response, innate immune response, immune system process, defense response, and biological regulation (false discovery rate [FDR] < 0.01; Additional file 2: Figure S8). The biological regulation processes comprised diverse genes, such as multiple gene candidates related to flowering (*SVP*, *DBP1*, *YAF9A*, *BLH3*, *VAL2*, *EBS*, *ATH1*) [30–37], response to temperature stress (*LCBK1*) [38], root hair development (*ZFP5*, *RSL4*, *WRKY6*) [39–41], and circadian period (*ARR4*) [42]. For the immune response genes, 19 genes were enriched in all of the four GO terms at the same time except the biological regulation GO term, of which nine are nucleotide-binding,

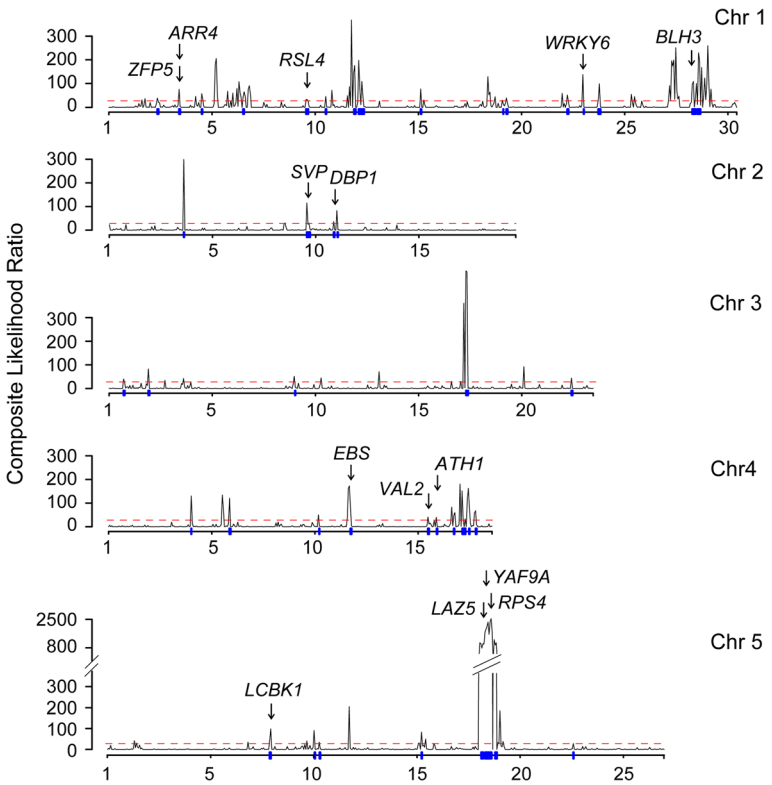

Fig. 3 Positive selection analysis in the Yangtze River basin population. *Dashed red line* indicates the cut-off of composite likelihood ratio and *vertical blue lines* across the *x-axis* indicate the overlapped regions that are under positive selection in both SweepFinder2 and OmegaPlus

leucine rich repeat (NB-LRR) genes, including the well--known genes *RPS4* and *LAZ5*. RPS4 interacts with another NB-LRR protein RRS1-R and triggers defense response [43, 44]. *LAZ5* encodes a TIR-class NB-LRR gene and could activate cell death [45, 46]. Overall, the selection scan suggested that genes enriched in biological regulation processes, such as flowering time, immune response, and defense response, could play an important role during the establishment of the Yangtze River population.

Genomic regions associated with flowering time variation

Given that some outlier loci from genome-wide selection scans might not be actually adaptive [47] and adaptation to the new climate could involve different traits [1], association between fitness related traits and genomic variation is a robust way to validate genes that are found by selection scans [48, 49]. Flowering time is an important fitness trait and there was huge flowering time variation within or between popY and popN (Additional file 3: Table S2). In particular, popY is significantly early flowering than popN (Fig. 4a).

To clarify the genetic basis of flowering time variation, we constructed F_2 population (1158 plants in total) using two extreme accessions with contrasting flowering time (3-2 flowered after 50.33 days and 29-8 after 24.87 days), and identified *SVP* as the causal locus (Fig. 4b). To identify the causal gene, 86 plants of F_2 individuals were used in the analysis. Quantitative trait locus (QTL) mapping identified two QTLs on chromosomes 2 and 5 that were responsible for the flowering time variation and the locus on chromosome 2 explained a larger fraction of the flowering time variation compared with that on chromosome 5 (60.9% vs 21.6%; Fig. 4b). To fine-map the locus on chromosome 2, we analyzed 184 early-flowering F_2 plants and narrowed the candidate region to 130 kb (Additional file 1: Table S7). Within this region between the two accessions, there are only four polymorphisms in four different genes that induced amino acid changes, which are assumed to be functionally important [50]. Only one of these four genes, *SHORT VEGETATIVE PHASE* (*SVP*, AT2G22540) is a well-known negative regulator of the onset of flowering that could be degraded at high temperature and promote flowering [51, 52]. We divided the 98 accessions of popN and popY with the flowering time data, into two different haplotypes according to the non-synonymous polymorphism. There is significant difference in the flowering time between the two haplotypes (Fig. 4c).

The non-synonymous polymorphism between the two haplotypes leads to one amino acid substitution (Ala^{32}/Val^{32}) in exon1 located in the MADS-box domain,

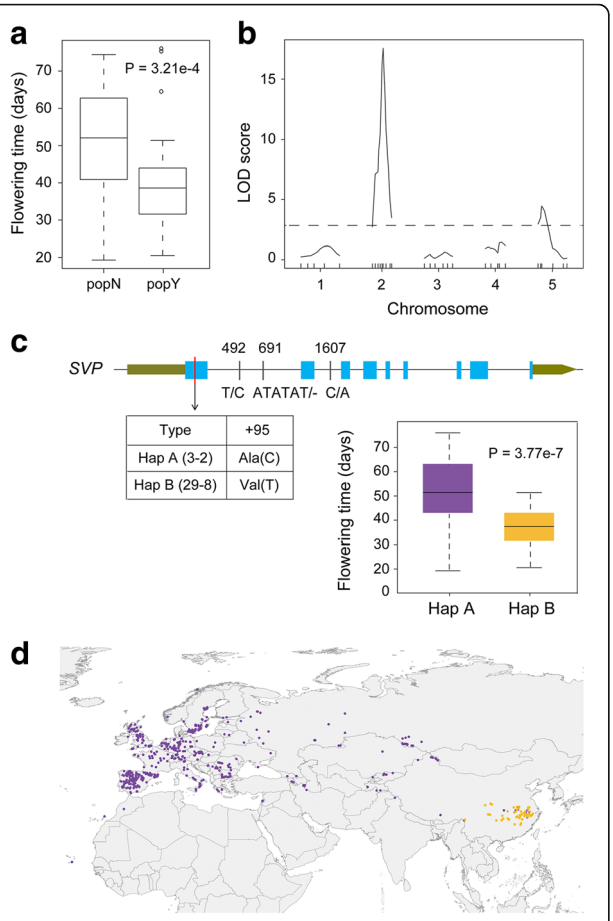

Fig. 4 Genomic regions associated with flowering time variation. **a** Flowering time variation between popY and popN. **b** QTL mapping analysis of flowering time between 3-2 and 29-8. The *dashed horizontal line* indicates the LOD threshold for QTLs (LOD = 2.9). **c** Sequence variation of *SVP* between 3-2 and 29-8; association between haplotypes and flowering time among 98 accessions. **d** Distribution of haplotypes across the world

which has been demonstrated to generate a loss-of-function (LOF) allele and could promote flowering [34] (Fig. 4c). Within the 881 genomes from the 1001 Genomes Project and the 118 genomes sequenced in this study (see Additional file 7: Table S8 for the details), we found that the amino acid substitution (Ala^{32}/Val^{32}) only existed in the Yangtze River region and was almost fixed, consistent with the scenario of positive selection on the *SVP* gene (Fig. 4d). However, this mutation has been identified in the natural accessions of Pakistan and Japan [34] that are not included in the present study. We concluded that the amino acid mutation of the *SVP* gene should have contributed to the adaptation to the Yangtze River basin.

Discussion

Global climate change has had a tremendous impact on the fitness of various organisms, mainly due to the lagging adaptation to climate change [53]. Understanding the adaptation of plants to new environments is a robust and practical way to understand the mechanisms behind this mismatch [3, 54]. In particular, it is largely unknown which kind of molecular processes or mechanisms are the determinant factors during adaptation process. To fully clarify the complete picture of local adaptation is challenging and complicated, as the process involves different factors, including identifying the genomic loci under selection, the phenotypes that selection is acting upon, and the external conditions driving the selection [55]. The classic scan of genes under positive selection and the mapping of genes correlated with the adaptive traits, such as flowering time, are robust ways to identify genes correlated with adaptation [9, 55].

The present study revealed the demographic history of *A. thaliana* at the global level of its natural habitats and indicates that the Yangtze River population is a unique population that diverged 61,409 years ago and expanded recently to the Yangtze River basin. This knowledge is a great opportunity to address how plants adapt to the diverse habitats in natural environments. We found that biological regulation processes, such as flowering time, immune and defense response processes could be important in this adaptation process. Particularly, *SVP* LOF mutation has been under positive selection and is nearly fixed in the Yangtze River population. Given that *SVP* is an important gene to allow plants to respond to ambient temperature changes in the context of global climate change [56], it must play an important role in the adaptation of the plant to the Yangtze River basin, the most south-eastern of *A. thaliana*'s native habitats. Consistently, during the range expansion of an invasive plant *Lythrum salicaria*, earlier flowering is important for the adaptation [54]. Many more studies are necessary to reveal the genetic basis of adaptation; for example, further analyses of the genes under positive selection in this study will be insightful for understanding the genetic basis of adaptation, mapping another QTL on chromosomes 5, and characterizing the mechanism behind the flowering time variation between the two accessions (3-2 and 29-8). In addition, given that we found that there are gene flows between popN and popY at two different periods (Fig. 2a), it would be intriguing to know to what extent these gene flows have contributed to adaptation. Overall, this study greatly progresses our understanding of the adaptation in plants by exploring the genetic variations and adaptation of the worldwide samples of *A. thaliana*.

Conclusions

Adaptation is a robust way to deal with the challenge of global climate change. Examining recent range expansion

aids our understanding of how organisms evolve to overcome environmental constraints. Our results suggest that *A. thaliana* dispersed thousands of years ago to the Yangtze River basin, the most south-eastern edge of its native habitats. In addition, we demonstrate that flowering time variation related genes and immune response genes, particularly *SVP*, have contributed to the adaptation to the Yangtze River basin. This study highlights the importance of adaptation and demonstrates the genetic basis of adaptation in plants.

Methods

Plant materials and resequencing

A total of 118 strains were collected from north-western China and south-western China along the Yangtze River basin to eastern China [57] (Additional file 3: Table S2). Genomic DNA was extracted from the seedlings using the CTAB method [58]. Paired-end sequencing libraries with insert size around 500 bp were constructed. One hundred base-pair paired-end reads were sequenced using Illumina HiSeq 2000 for 91 samples and 150 bp paired-end reads were sequenced using Illumina HiSeq X Ten for the other 27 samples. For flowering time measurements, at least 11 plants were sowed for each strain in the greenhouse at 20 °C and 40–65% humidity with a 16-h photoperiod. Flowering time was assayed as the day of the first flower anthesis and the average of flowering time from each strain was regarded as flowering time [59].

Identification of SNPs and indels

Paired-end reads were mapped to the TAIR10 reference genome (www.arabidopsis.org) using Burrows–Wheeler Alignment tool (version 0.6.2) [60], allowing up to 4% mismatches and one gap. Next, the rmdup function of Samtools (version 0.1.8) [61] was used to remove reads that were duplicated in library preparation or sequencing. Finally, reads were locally realigned with the Genome Analysis Toolkit (GATK version 2.1.8) [62] Indel Realignment tool that performs realignment around indels to avoid alignment errors. SNPs and indels were called using the UnifiedGenotyper tool packaged in GATK with default parameters. Extra filtration steps were applied to the raw SNPs and indels using the built-in function VariantFiltration, including quality (Q) ≥ 30, mapping quality (MQ) ≥ 20, quality-by-depth ratio (QD) ≥ 10, ReadPosRankSum ≥ − 8.0, depth coverage (DP) ≥ 3, probability of strand bias (FS) ≤ 10.0 (FS ≤ 200.0 for indels), and no more than three SNPs within 10 bp.

Population genetics analysis

Besides the 118 strains sequenced in this study, 103 published strains were included for analysis [10, 14, 27] (Additional file 4: Table S3) and thus 221 strains in total were used in the study. The biallelic SNPs with information in at least 219 strains (in total, 1.97 million SNPs) were used to perform the population genetics analyses. ADMIXTURE [63] was used to estimate the genetic ancestry of each sample, specifying a range of 2–5 hypothetical ancestral populations. PCA was performed with EIGENSOFT (version 4.2) [64]. The unrooted neighbor-joining tree was constructed with PHYLIP (version 3.695) [65]. In addition, a neighbor-joining tree using the third codon site of 16,047 orthologous genes across the three closely related species, *A. thaliana* (221 strains), *Arabidopsis lyrata* (MN47) [66], and *Capsella rubella* (MTE) [67], was constructed, with MN47 and MTE as the outgroups. Orthologous genes among *A. thaliana*, *A. lyrata*, and *C. rubella* were identified by InParanoid [68] with default parameters. Nucleotide diversity π, Watterson's estimator θ, and F_{ST} were calculated in a 200-kb sliding window with a step size of 10 kb.

Demographic and ecological niche analyses

The demographic history of *A. thaliana* was inferred using the MSMC model [28] based on two or four haploid genomes with default parameters. As the *A. thaliana* plant self-fertilizes, the genome of each strain can be considered as a haplotype sequence when heterozygous sites are excluded. Only homozygous SNP sites without missing data were used in the analysis. For two haplotypes, two strains were randomly extracted from the same population (popE, popN, or popY) or two populations (one haplotype from each population). For four haplotypes, four strains were randomly extracted from the same population or two different populations (two haplotypes from each population). In each analysis, 50 rounds of random samplings were performed to estimate the mean and standard deviation of the relative cross coalescence rate or the coalescence rates along the evolutionary time.

Fastsimcoal2 [29] was used to infer the demographic parameters of popY and popN. First, the site frequency spectra (SFS) was computed for the 399,165 non-coding SNPs that have no missing site in any of the samples. Four alternative models with different extents of gene flow and varying population sizes were compared, using Akaike's information criterion (AIC) and Akaike's weight of evidence [29]. The timespans of the gene flow were set according to the observations in Fig. 2a and effective population sizes were set according to the results of Fig. 2b. The best parameter estimates under each model were obtained from 50 independent runs with a minimum of 100,000 and a maximum of 1,000,000 coalescent simulations as well as 10–40 cycles of the likelihood maximization algorithm. SFS entries with support from < 10 SNPs were ignored [29]. The 95% confidence intervals for each parameter were computed based on 100 parametric bootstrapping datasets simulated according to the estimations under

the best model, using fastsimcoal2 again. In this study, the generation time (g) was set as one year and the mutation rate was considered to be 7×10^{-9} per base per generation [69] and the recombination rate as 3.6 cM/Mb [70].

To reconstruct the potential distribution pattern of *A. thaliana* worldwide, ENM analysis was employed to predict the distribution of *A. thaliana* during three periods, including the present time, the time of last glacial maximum (LGM; 0.021–0.018 MYA) and the time of last interglacial (LIG; 0.14–0.12 MYA). In total, 291 geo-referenced and non-overlapped occurrence records of *A. thaliana* from our own field works and published articles [9, 10, 14, 27] were used; these records covered nearly the whole native ranges of *A. thaliana* in the world (Additional file 5: Table S5). The 19 environmental variables of the three periods used to perform ENM analysis were downloaded from the WORLDCLIM database (www.worldclim.org). Since the existence of strongly related environmental variables may over-fit models during ENM analysis, environmental variables were filtered so that no two variables had a pairwise Pearson correlation coefficient $r > 0.7$ or < -0.7 (Additional file 1: Table S9). As a result, 11 environmental variables were used for the subsequent analysis (Additional file 1: Table S10). Ecological niche models were constructed using the present variables and projected for the other two historical variable datasets via maximum entropy in Maxent 3.3.3 [71] with default settings as in our previous study [72]. To identify the most significant climate variable that contributes to the distribution of *A. thaliana*, we performed PCA on the 19 environmental variables using R (www.r-project.org).

Selection test and functional annotation

SweepFinder2 is an effective program that implements a powerful likelihood-based method for detecting recent positive selection or selective sweeps. SweepFinder2 is the first method that accounts for the effects of negative selection on diversity when searching for adaptive alleles [73]. SweepFinder2 scanned for positive selection in the folded site frequency spectrum (fSFS) for popY. The parameter -g was set to 50,000. In total, 10,000 1-Mb simulation datasets were generated as the null datasets based on the demographic parameters from the best model using Fastsimcoal2. The 10,000 simulation datasets were used to calculate a statistical cut-off with the same parameters as for the real data, allowing for a false-positive rate of 0.01%. After filtering with the threshold of 27.85, the neighbor sweep targets were merged to sweep regions.

To increase the ability to detect selective sweeps, OmegaPlus (version 2.3.0; a LD-based method) was used [74]. The ω statistic was computed at 10 kb intervals. The minwin and maxwin parameters were set to 10 kb and 100 kb, respectively. As in the Fastsimcoal2 method, 10,000 simulation datasets were used to calculate a statistical cut-off. The sweep targets adjacent to each other were subsequently merged to sweep regions after filtering with the cut-off ($\omega > 11.92$). The overlap regions of the two methods were computed and those regions were regarded as the confident selective sweep regions. Genes within these regions were regarded as genes under selection. The software Cytoscape with the BiNGO plugin was used for GO analysis [75].

QTL mapping

For QTL mapping of flowering time variation, 1158 individuals of F_2 plants generated from 3-2 (female) and 29-8 (male) were used. Markers were identified based on the resequencing data, in which indel and SNP markers were called with Pindel (version 0.2.5a3) and GATK (version 2.1.8), respectively [76]. The genotype information of 32 markers across the whole genome, with an average density of 3.75 Mb/marker (Additional file 1: Table S11) and the flowering time of 86 F_2 individuals, were used to perform QTL analysis using the R/qtl package with default parameters implemented in R (http://www.R-project.org).

Statistical analysis

Statistical analyses were performed in R (www.r-project.org).

Additional files

Additional file 1: Table S1. Environmental variables used in the ecological analysis. **Table S4.** Demographic parameters results from fastsimcoal2. **Table S7.** Fine mapping of causal locus on chromosome 2 with 184 early flowering plants from the F_2 population of 3-2 × 29-8. **Table S9.** Significant Pearson's correlation ($r > 0.7$ or < -0.7) between bioclimatic variables in the distribution of all samples (outlined by green), with the retained variables during the ecological niche modeling analysis in red. **Table S10.** The contributions of the 11 bioclimatic variables (abbreviation in parentheses) to the Maxent models during ENM analysis for total and geographic populations, respectively, with their permutation importance index indicated in parentheses. Values corresponding to the three most significant variables are in red. **Table S11.** Markers used in the QTL mapping analysis. (DOCX 54 kb)

Additional file 2: Figure S1. PCA analysis of ecological differentiation among strains of *A. thaliana* based on 19 environmental variables using two discriminant principal components (PC) based on 291 geo-referenced and non-overlapped occurrence records of *A. thaliana* (Additional file 5: Table S5). Bio4 and bio12 are the most significant factors differentiating PC2 and PC1, respectively. All the pairwise comparisons of bio4 or bio12 among the three regions are significant ($P < 0.001$). **Figure S2.** Sequence variation of the 118 strains sequenced in this study. Intergenic represents intergenic region; gene represents genic region from ATG to TGA/TAA/TAG, including introns; 5′ UTR represents the untranslated region before start codon ATG; CDS represents coding sequence region; 3′ UTR refers the untranslated region after stop codon (TGA/TAA/TAG); total represents whole genome in total. **Figure S3.** Phylogenetic tree of all the 221 *A. thaliana* strains with outgroups. Numbers nearby a branch indicate the bootstrap value > 50% with 100 replicates. Strains from different regions were color-coded, pink: European strains, blue: central Asia strains, yellow: eastern Asia strains, green: others indicate strains from USA, Japan, and New Zealand, most probably reflects recent introduction, given *A. thaliana* originated in Europe. **Figure**

S4. Saturation analysis of the 30 times random samplings of the Yangtze River population (popY) based on the recovery of the number of total SNPs. **Figure S5.** Genetic variation among different populations. **Figure S6.** Different models of demographic history between the two populations of *A. thaliana*. Model 4 is the best fit model; see Table S4 for the detailed demographic parameters for each model. **Figure S7.** Selection scans of genes under positive selection based on LD-based method (OmegaPlus). The dashed red line indicates the threshold of 0.01% based on simulation data sets. **Figure S8.** Over-representation (FDR < 0.01) of GO annotation categories in gene sets under selection. (DOCX 2316 kb)

Additional file 3: Table S2. Samples sequenced in this study. (XLSX 21 kb)

Additional file 4: Table S3. Previously sequenced samples that are used in this study. (XLSX 16 kb)

Additional file 5: Table S5. Geographic information of the representative samples used in the ecological niche modeling (ENM) analysis. The geographic region was divided via longitude range and latitude range: Europe (192 samples) = E8.54 ~ 38.28 and N31.47 ~ 61.36; central Asia (48 samples) = E42.22 ~ 90.34 and N37.29 ~ 58.01; eastern Asia (51 samples) = E98.63 ~ 120.37 and N26.75 ~ 33.15. (XLSX 21 kb)

Additional file 6: Table S6. Genes under positive selection based on the overlapping of the two different methods. (XLSX 14 kb)

Additional file 7: Table S8. The 999 accessions used for the *SVP* gene haplotype analysis. (XLSX 51 kb)

Abbreviations
ENM: Ecological niche modelling; GO: Gene Ontology; MSMC: Multiple sequential Markovian coalescent; PCA: Principal component analysis; popE: Europe/North Africa population; popN: North-western China population; popY: Yangtze River basin population; QTL: Quantitative trait locus

Acknowledgements
We would like to thank Detlef Weigel, Song Ge, Wolfgang Busch, Marco Todesco, Li-Jia Qu, Tao Sang, and Yufei Wang for valuable comments and discussions about this work; especially Detlef Weigel for his insightful revision of the draft; Quan Long, Christian Huber, and Alexander Lipka for their valuable suggestions on the data analyses; and Jian Wang for his help with the sample collection. In particular, we thank the anonymous reviewers for their help improving the manuscript.

Funding
This work was supported by National Natural Science Foundation of China grants 91231104, 31222006, and 31470331 (YLG); and the 100 Talents Program of the Chinese Academy of Sciences (YLG).

Availability of data and materials
The genome sequences of the 118 strains of *A. thaliana* reported in this paper have been deposited in the NCBI Sequence Read Archive (SRA) under accession number SRP062811 [57] and of the 103 strains published under accession number (SRA029270, SRP012869, and SRA012474) [10, 14, 27]. Sequence data of *SVP* have been deposited in the GenBank under accession numbers MF663187 and MF663188 [77].

Authors' contributions
YLG conceived the study; HG provided the biological materials; YPZ, XHH, TSH, XMN, LY, and JFC performed the experiments; YPZ, XHH, QW, ZWL, TSH, YCX, FMZ, DT, ZT, HG, and YLG analyzed and interpreted the data; YPZ, XHH, QW, and YLG wrote the paper with contribution from all authors. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China. [2]University of Chinese Academy of Sciences, Beijing 100049, China. [3]Xinjiang Key Laboratory of Grassland Resources and Ecology, College of Grassland and Environment Sciences, Xinjiang Agricultural University, Urümqi, China. [4]State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China. [5]State Key Laboratory for Protein and Plant Gene Research, College of Life Sciences, Peking University, Beijing 100871, China. [6]The National Plant Gene Research Center, Beijing 100101, China.

Received: 1 November 2017 Accepted: 12 December 2017
Published online: 28 December 2017

References
1. Hoffmann AA, Sgro CM. Climate change and evolutionary adaptation. Nature. 2011;470:479–85.
2. Carroll SP, Jorgensen PS, Kinnison MT, Bergstrom CT, Denison RF, Gluckman P, et al. Applying evolutionary biology to address global challenges. Science. 2014;346:1245993.
3. Scheffers BR, De Meester L, Bridge TC, Hoffmann AA, Pandolfi JM, Corlett RT, et al. The broad footprint of climate change from genes to biomes to people. Science. 2016;354:aaf7671.
4. Barrick JE, Lenski RE. Genome dynamics during experimental evolution. Nat Rev Genet. 2013;14:827–39.
5. Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV. Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. Nat Genet. 2010;42:260–3.
6. Kang JQ, Zhang HT, Sun TS, Shi YH, Wang JQ, Zhang BC, et al. Natural variation of *C-repeat-binding factor (CBFs)* genes is a major cause of divergence in freezing tolerance among a group of *Arabidopsis thaliana* populations along the Yangtze River in China. New Phytol. 2013;199:1069–80.
7. Durvasula A, Fulgione A, Gutaker RM, Alacakaptan SI, Flood PJ, Neto C, et al. African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. Proc Natl Acad Sci U S A. 2017;114:5213–8.
8. Lee CR, Svardal H, Farlow A, Exposito-Alonso M, Ding W, Novikova P, et al. On the post-glacial spread of human commensal *Arabidopsis thaliana*. Nat Commun. 2017;8:14458.
9. The 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell. 2016;166:481–91.
10. Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, et al. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet. 2011;43:956–63.
11. Fournier-Level A, Korte A, Cooper MD, Nordborg M, Schmitt J, Wilczek AM. A map of local adaptation in *Arabidopsis thaliana*. Science. 2011;334:86–9.
12. Hancock AM, Brachi B, Faure N, Horton MW, Jarymowycz LB, Sperone FG, et al. Adaptation to climate across the *Arabidopsis thaliana* genome. Science. 2011;334:83–6.
13. Horton MW, Hancock AM, Huang YS, Toomajian C, Atwell S, Auton A, et al. Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. Nat Genet. 2012;44:212–6.
14. Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A, et al. Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. Nat Genet. 2013;45:884–90.
15. Dubin MJ, Zhang P, Meng D, Remigereau MS, Osborne EJ, Paolo Casale F, et al. DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. Elife. 2015;4:e05255.
16. Kawakatsu T, Huang SS, Jupe F, Sasaki E, Schmitz RJ, Urich MA, et al. Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. Cell. 2016;166:492–505.
17. Koornneef M, Alonso-Blanco C, Vreugdenhil D. Naturally occurring genetic variation in *Arabidopsis thaliana*. Annu Rev Plant Biol. 2004;55:141–72.

18. Weigel D. Natural variation in Arabidopsis: from molecular genetics to ecological genomics. Plant Physiol. 2012;158:2–22.
19. Weigel D, Nordborg M. Population genomics for understanding adaptation in wild plant species. Annu Rev Genet. 2015;49:315–38.
20. Beck JB, Schmuths H, Schaal BA. Native range genetic variation in *Arabidopsis thaliana* is strongly geographically structured and reflects Pleistocene glacial dynamics. Mol Ecol. 2008;17:902–15.
21. Platt A, Horton M, Huang YS, Li Y, Anastasio AE, Mulyati NW, et al. The scale of population structure in *Arabidopsis thaliana*. PLoS Genet. 2010;6: e1000843.
22. 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 Biol. 2010;10:22.
23. He F, Kang D, Ren Y, Qu LJ, Zhen Y, Gu H. Genetic diversity of the natural populations of *Arabidopsis thaliana* in China. Heredity (Edinb). 2007;99:423–31.
24. Gaut B. *Arabidopsis thaliana* as a model for the genetics of local adaptation. Nat Genet. 2012;44:115–6.
25. Scheinfeldt LB, Tishkoff SA. Recent human adaptation: genomic approaches, interpretation and insights. Nat Rev Genet. 2013;14:692–702.
26. Ellegren H. Genome sequencing and population genomics in non-model organisms. Trends Ecol Evol. 2014;29:51–63.
27. Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, et al. Patterns of population epigenomic diversity. Nature. 2013;495:193–8.
28. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nat Genet. 2014;46:919–25.
29. Excoffier L, Dupanloup I, Huerta-Sanchez E, Sousa VC, Foll M. Robust demographic inference from genomic and SNP data. PLoS Genet. 2013;9: e1003905.
30. Gómez-Mena C, Pineiro M, Franco-Zorrilla JM, Salinas J, Coupland G, Martinez-Zapater JM. Early bolting in short days: an Arabidopsis mutation that causes early flowering and partially suppresses the floral phenotype of leafy. Plant Cell. 2001;13:1011–24.
31. Pineiro M, Gomez-Mena C, Schaffer R, Martinez-Zapater JM, Coupland G. EARLY BOLTING IN SHORT DAYS is related to chromatin remodeling factors and regulates flowering in Arabidopsis by repressing *FT*. Plant Cell. 2003;15: 1552–62.
32. Proveniers M, Rutjens B, Brand M, Smeekens S. The Arabidopsis TALE homeobox gene *ATH1* controls floral competency through positive regulation of *FLC*. Plant J. 2007;52:899–913.
33. Zacharaki V, Benhamed M, Poulios S, Latrasse D, Papoutsoglou P, Delarue M, et al. The Arabidopsis ortholog of the YEATS domain containing protein *YAF9a* regulates flowering by controlling H4 acetylation levels at the *FLC* locus. Plant Sci. 2012;196:44–52.
34. Mendez-Vigo B, Martinez-Zapater JM, Alonso-Blanco C. The flowering repressor *SVP* underlies a novel *Arabidopsis thaliana* QTL interacting with the genetic background. PLoS Genet. 2013;9:e1003289.
35. Yuan W, Luo X, Li Z, Yang W, Wang Y, Liu R, et al. A cis cold memory element and a trans epigenome reader mediate Polycomb silencing of *FLC* by vernalization in Arabidopsis. Nat Genet. 2016;48:1527–34.
36. Zhai H, Ning W, Wu H, Zhang X, Lu S, Xia Z. DNA-binding protein phosphatase AtDBP1 acts as a promoter of flowering in Arabidopsis. Planta. 2016;243:623–33.
37. Zhang L, Zhang X, Ju H, Chen J, Wang S, Wang H, et al. Ovate family protein1 interaction with BLH3 regulates transition timing from vegetative to reproductive phase in Arabidopsis. Biochem Biophys Res Commun. 2016;470:492–7.
38. Huang X, Zhang Y, Zhang X, Shi Y. Long-chain base kinase1 affects freezing tolerance in *Arabidopsis thaliana*. Plant Sci. 2017;259:94–103.
39. An L, Zhou Z, Sun L, Yan A, Xi W, Yu N, et al. A zinc finger protein gene *ZFP5* integrates phytohormone signaling to control root hair development in Arabidopsis. Plant J. 2012;72:474–90.
40. Marzol E, Borassi C, Denita Juarez SP, Mangano S, Estevez JM. RSL4 takes control: multiple signals, one transcription factor. Trends Plant Sci. 2017;22:553–5.
41. Stetter MG, Benz M, Ludewig U. Increased root hair density by loss of *WRKY6* in *Arabidopsis thaliana*. PeerJ. 2017;5:e2891.
42. Salome PA, To JP, Kieber JJ, McClung CR. Arabidopsis response regulators ARR3 and ARR4 play cytokinin-independent roles in the control of circadian period. Plant Cell. 2006;18:55–69.
43. Le Roux C, Huet G, Jauneau A, Camborde L, Tremousaygue D, Kraut A, et al. A receptor pair with an integrated decoy converts pathogen disabling of transcription factors to immunity. Cell. 2015;161:1074–88.
44. Sarris PF, Duxbury Z, Huh SU, Ma Y, Segonzac C, Sklenar J, et al. A plant immune receptor detects pathogen effectors that target WRKY transcription factors. Cell. 2015;161:1089–100.
45. Palma K, Thorgrimsen S, Malinovsky FG, Fiil BK, Nielsen HB, Brodersen P, et al. Autoimmunity in Arabidopsis *acd11* is mediated by epigenetic regulation of an immune receptor. PLoS Pathog. 2010;6:e1001137.
46. Munch D, Teh OK, Malinovsky FG, Liu Q, Vetukuri RR, El Kasmi F, et al. Retromer contributes to immunity-associated cell death in Arabidopsis. Plant Cell. 2015;27:463–79.
47. Lachance J, Tishkoff SA. Population genomics of human adaptation. Annu Rev Ecol Evol Syst. 2013;44:123–43.
48. Jeong C, Di Rienzo A. Adaptations to local environments in modern human populations. Curr Opin Genet Dev. 2014;29:1–8.
49. Weinig C, Ewers BE, Welch SM. Ecological genomics and process modeling of local adaptation to climate. Curr Opin Plant Biol. 2014;18:66–72.
50. Yano K, Yamamoto E, Aya K, Takeuchi H, Lo PC, Hu L, et al. Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nat Genet. 2016;48:927–34.
51. Lee JH, Ryu HS, Chung KS, Pose D, Kim S, Schmid M, et al. Regulation of temperature-responsive flowering by MADS-box transcription factor repressors. Science. 2013;342:628–32.
52. Pose D, Verhage L, Ott F, Yant L, Mathieu J, Angenent GC, et al. Temperature-dependent regulation of flowering by antagonistic FLM variants. Nature. 2013;503:414–7.
53. Wilczek AM, Cooper MD, Korves TM, Schmitt J. Lagging adaptation to warming climate in *Arabidopsis thaliana*. Proc Natl Acad Sci U S A. 2014;111: 7906–13.
54. Colautti RI, Barrett SC. Rapid adaptation to climate facilitates range expansion of an invasive plant. Science. 2013;342:364–6.
55. Fan S, Hansen MEB, Lo Y, Tishkoff SA. Going global by adapting local: A review of recent human adaptation. Science. 2016;354:54–9.
56. Lee JH, Yoo SJ, Park SH, Hwang I, Lee JS, Ahn JH. Role of *SVP* in the control of flowering time by ambient temperature in Arabidopsis. Genes Dev. 2007; 21:397–402.
57. Zou YP, Hou XH, Wu Q, Li ZW, Han TS, Niu XM, et al. Adaptation of *Arabidopsis thaliana* to the Yangtze River basin. NCBI SRA. 2017;BioProject Accession: PRJNA293798; SRP062811.
58. Doyle JJ, Doyle JL. A rapid DNA isolation procedure from small quantities of fresh leaf tissues. Phytochem Bull. 1987;19:11–5.
59. Guo YL, Todesco M, Hagmann J, Das S, Weigel D. Independent *FLC* mutations as causes of flowering-time variation in *Arabidopsis thaliana* and *Capsella rubella*. Genetics. 2012;192:729–39.
60. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
61. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
62. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.
63. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64.
64. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38:904–9.
65. Feisenstein J. PHYLIP-phylogeny inference package (version 3.2). Cladistics. 1989;5:164–6.
66. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, et al. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet. 2011;43:476–81.
67. Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F, Guo YL, et al. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. Nat Genet. 2013;45:831–5.
68. Berglund AC, Sjolund E, Ostlund G, Sonnhammer EL. InParanoid 6: eukaryotic ortholog clusters with inparalogs. Nucleic Acids Res. 2008;36: D263–266.
69. Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, et al. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science. 2010;327:92–4.
70. Salomé PA, Bomblies K, Fitz J, Laitinen RA, Warthmann N, Yant L, et al. The recombination landscape in *Arabidopsis thaliana* F_2 populations. Heredity (Edinb). 2012;108:447–55.

71. Phillips SJ, Anderson RP, Schapire RE. Maximum entropy modeling of species geographic distributions. Ecol Model. 2006;190:231–59.
72. Han TS, Wu Q, Hou XH, Li ZW, Zou YP, Ge S, et al. Frequent introgressions from diploid species contribute to the adaptation of the tetraploid Shepherd's purse (*Capsella bursa-pastoris*). Mol Plant. 2015;8:427–38.
73. DeGiorgio M, Huber CD, Hubisz MJ, Hellmann I, Nielsen R. SweepFinder2: increased sensitivity, robustness and flexibility. Bioinformatics. 2016;32:1895–7.
74. Alachiotis N, Stamatakis A, Pavlidis P. OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. Bioinformatics. 2012;28:2274–5.
75. Maere S, Heymans K, Kuiper M. *BiNGO*: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. Bioinformatics. 2005;21:3448–9.
76. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009;25:2865–71.
77. Zou YP, Hou XH, Wu Q, Li ZW, Han TS, Niu XM, et al. Adaptation of *Arabidopsis thaliana* to the Yangtze River basin. NCBI GenBank. 2017; GenBank accession number:MF663187-MF663188.