Genetic adaptations in the population history of Arabidopsis thaliana

Hirohisa Kishino¹, Reiichiro Nakamichi², and Shuichi Kitada³

¹University of Tokyo
²Japan Fisheries Research and Education Agency
³Tokyo University of Marine Science and Technology

May 8, 2023

Abstract

The Arabidopsis thaliana population has been exposed to unexperienced biotic and abiotic stresses as a result of range expansion or environmental change. To obtain a global picture of the genetic adaptations in the population history of A. thaliana, we constructed a database of the phenotypic adaptations (p-adaptations) and gene expression adaptations (e-adaptations). We analysed the dynamics of the allele frequencies at the 23,880 QTLs of 174 traits and 8,618 eQTLs of 1,829 genes with respect to the total SNPs in the genomes, and identified 650 p-adaptations and 3,925 e-adaptations (FDR=0.05). The population underwent large scale p-adaptations and e-adaptations along four lineages, the eastward migration to Central Asia and South Siberia, Russia, the northward migration to Sweden, the migration to Azerbaijan, and the migration of the German population to the United States. Extremely cold winters and short summers prolonged seed dormancy, and expanded the root system architecture. Low temperatures prolonged the growing season and low light intensity required the increased chloroplast activity. The subtropical and humid environment enhanced phytohormone signaling pathways in response to the biotic and abiotic stresses. Exposure to heavy metals selected for alleles underlying low heavy metal uptake from soil, lower growth rate, lower resistance to bacteria, and higher expression of photosynthetic genes were selected. The database of p-adaptations and e-adaptations, which complements studies focusing on specific aspects of adaptation, may be useful for future studies to understand the biological adaptations of A. thaliana throughout its population history.

Genetic adaptations in the population history of Arabidopsis thaliana

Running title: Genetic adaptations of Arabidopsis thaliana

Hirohisa Kishino¹-² Reiichiro Nakamichi³ Shuichi Kitada⁴

¹ Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo, Japan
² Research and Development Initiative, Chuo University, Tokyo, Japan
³ Japan Fisheries Research and Education Agency, Yokohama, Japan
⁴ Tokyo University of Marine Science and Technology, Tokyo, Japan

Correspondence: Hirohisa Kishino
Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo, Japan.
Email: kishino@g.ecc.u-tokyo.ac.jp

Abstract
The *Arabidopsis thaliana* population has been exposed to unexperienced biotic and abiotic stresses as a result of range expansion or environmental change. To obtain a global picture of the genetic adaptations in the population history of *A. thaliana*, we constructed a database of the phenotypic-adaptations (p-adaptations) and gene expression-adaptations (e-adaptations). We analysed the dynamics of the allele frequencies at the 23,880 QTLs of 174 traits and 8,618 eQTLs of 1,829 genes with respect to the total SNPs in the genomes, and identified 650 p-adaptations and 3,925 e-adaptations (FDR=0.05). The population underwent large scale p-adaptations and e-adaptations along four lineages, the eastward migration to Central Asia and South Siberia, Russia, the northward migration to Sweden, the migration to Azerbaijan, and the migration of the German population to the United States. Extremely cold winters and short summers prolonged seed dormancy, and expanded the root system architecture. Low temperatures prolonged the growing season and low light intensity required the increased chloroplast activity. The subtropical and humid environment enhanced phytohormone signaling pathways in response to the biotic and abiotic stresses. Exposure to heavy metals selected for alleles underlying low heavy metal uptake from soil, lower growth rate, lower resistance to bacteria, and higher expression of photosynthetic genes were selected. The database of p-adaptations and e-adaptations, which complements studies focusing on specific aspects of adaptation, may be useful for future studies to understand the biological adaptations of *A. thaliana* throughout its population history.

**KEYWORDS**

*Arabidopsis thaliana*, coadaptation, gene expression-adaptation, phenotypic-adaptation, population history

**INTRODUCTION**

In its history of range expansion and migration, a population has experienced biotic and abiotic stresses and has adapted to these stresses by selecting its alleles that can cope with them. In the case of insecticide resistance in *Drosophila*, a single allele saved the population (Daborn et al., 2002). On the other hand, the case of herbicide tolerance in the weedy morning glory revealed the cost of adaptation (Baucom & Mauricio, 2004). In plant-pathogen interactions, both plant-resistance pathways and pathogen’s virulence mechanisms are influenced by environmental factors such as temperature and humidity. The complex nature of adaptation is difficult to capture in experiments with controlled environmental conditions (Velásquez et al., 2018).

As a model plant species, the adaptation of *Arabidopsis thaliana* has been extensively studied for important traits and environments; flowering times (Izawa, 2007; Li et al., 2010; Ågren et al., 2017; Zan & Carlborg, 2019), seed dormancy (Alonso-Blanco et al., 2003), timing of germination (Zacchello et al., 2020), tolerance to freezing (Ågren & Schemske, 2012), response to salt, osmotic, and cold stress (Kreps et al., 2002; Busoms et al., 2021), water-use efficiency (Dittberner et al., 2018), and acclimation to light intensity (Stewart et al., 2015). In investigating the evolutionary potential of extreme drought adaptation, Exposito-Alonso et al., (2017) highlighted the importance of genetic variation in standing populations. To this end, it is important to understand when and where *A. thaliana* populations have experienced biotic and abiotic stresses and how they have responded to these stresses.

Although *A. thaliana* is selfing, the linkage disequilibrium decays rapidly, within 50 kb (Nordborg et al., 2005). This indicates the high accuracy of the population genomic inference on the genomic diversity, population structure, and population history. Approximate Bayesian Computation (ABC, Beaumont et al., 2002) with the spatial coalescent model was applied to the above sequence fragments from 76 European individuals and inferred from the longitudinal gradient of genomic polymorphism that the major migration in Europe occurred from the east about 10,000 years ago and spread westwards at a rate of about 0.9 km/year (François et al., 2008).

The 1001 Genomes Consortium presented a detailed map of variation in the genomes of 1,135 naturally inbred lines representing the native Eurasian and North African range and recently colonized North America (The 1001 Genomes Consortium, 2016). Pair-wise divergence analysis identified 26 outliers, which were reefered as relicts (from glacial refugium). Out of the 26 relicts, 22 were in the Iberian Peninsula, and one line each from the Cape Verde Islands, Canary Islands, Sicily, and Lebanon. Model-based inference of population structure
(Pritchard et al., 2000; Alexander et al., 2009) identified one group consisting of relicts and eight genetic clusters of non-relicts. While the polymorphism of relicts showed the sign of isolation by distance (IBD), that of non-relicts did not, possibly because of admixture with relicts. By carefully eliminating the effect of admixture with the relicts, Lee et al., (2017) suggested the Balkans or the Black Sea region as the most likely origin for the spread of the non-relicts. The widespread relicts were replaced by the non-relicts, which rapidly spread westwards, probably in association with agriculture and human commensal in general. To infer the pattern of introduction into the North American population and adaptation of immigrants, Shirsekar et al., (2021) collected 2,861 individuals from an area of 1,200 by 900 km in the Eastern United States. Their analysis of haplotype sharing with the global Afro-Eurasian collection inferred the ancestry profiles in detail.

The multiomic database generated by the 1001 Genomes Project gave rise to the databases collecting the information on the phenome/transcriptome/genome associations (Togninalli et al., 2020; Lan et al., 2021). To obtain a global picture of the genetic adaptations in the population history of *A. thaliana*, we constructed a database of the phenotypic-adaptations (p-adaptations) and gene expression-adaptations (e-adaptations) from the 1001 Genomes database. Utilizing the above databases, we studied the dynamics of allele frequencies at 23,880 QTLs for 174 traits and 8,618 eQTLs for 1,829 genes over population history. Traits such as the time to flowering and germination rates are quantitative traits, while those such as protist disease resistance and “growing includes September” are 0-1 scored qualitative traits. Here we refer to the SNPs associated with the trait values of both types as QTL. For each trait/gene expression, a p-adaptation/e-adaptation along an edge of the admixture tree (Pickrell & Pritchard, 2012) was identified as the set of changes in allele frequencies at its QTLs/eQTLs that are collectively predicted to significantly change the value of the trait/gene expression (Racimo et al., 2018). The constructed database of p-adaptations and e-adaptations, combined with the climatic information, can characterize the stress of the local environment. The information on the trait adaptations is directly interpretable. The co-adapted gene expressions may reflect the physiological requirements for adaptation to that stress. The integration of both types of the information provides an indication of where and when the *A. thaliana* population experienced the environmental stresses, how it was affected by these stresses and how it responded to them.

MATERIALS AND METHODS

Multi-omic data

The datasets of sampling locations and whole-genome SNP genotypes of the 1,135 accessions of *Arabidopsis thaliana* (Supplementary Figure S1) were downloaded from 1001 Genomes (The 1001 Genomes Consortium, 2016). The data of transcriptome expression values was obtained from NCBI GEO (GSE80744, GSE43858, GSE54680). The information of the phenotypes and the local environments, their QTLs was downloaded from AraPheno and AraGWAS (Togninalli et al., 2020, Fick & Hijmans, 2017, Ferrero-Serrano and Assmann, 2019). The information of transcriptome associated eQTLs was obtained from AtMAD (Lan et al., 2021). The annotations of QTL/eQTL associated genes were obtained from the databases of Entrez, GO and TAIR (Maglott et al., 2020, The Gene Ontology Consortium, 2020, Berardini et al., 2015) using R packages rentrez (Winter, 2017), AnnotationDbi (Pagès et al., 2023), org.At.tair.db (Carlson, 2019) and GO.db (Carlson, 2016). We totally obtained 49,973 QTLs, 16,672 eQTLs, 516 phenotypes and 2,879 expression genes from these public databases. Then, we screened the SNPs of QTLs and eQTLs by the criteria of non-indel and biallelic. We also screened the SNPs, phenotypes and expression genes by the criteria that there is no missing population (see below). By these screening, we obtained 23,880 QTLs, 8,618 eQTLs, 174 phenotypes and 1,829 expression genes for our study.

Estimation of the population history

To estimate the history of range expansion, admixture, and introgression, we applied Treemix (Pickrell & Pritchard, 2012) to the set of the current populations of *Arabidopsis thaliana*. For the population units, we first followed the assignment of the 1,135 accessions to the genetic clusters by ADMIXTURE (k=9, Alexander et al., 2009, The 1001 Genomes Consortium, 2016). The genetic clusters consist of one relict group and eight non-relict groups, broadly corresponding to geography, labelled as Asia, Central Europe,
Germany, Italy-Balkan-Caucasus, Northern Sweden, Southern Sweden, Spain, and Western Europe. To get a global view of local adaptations, we further subdivided them into countries: Armenia (ARM), Austria (AUT), Azerbaijan (AZE), Bulgaria (BUL), Czech Republic (CZE), Spain (ESP), France (FRA), Georgia (GEO), Germany (GER), Italy (ITA), Kyrgyzstan (KGZ), Lithuania (LTU), Netherlands (NED), Portugal (POR), Romania (ROU), Russian Federation (RUS), Serbia (SRB), Switzerland (SUI), Slovakia (SVK), Sweden (SWE), Tajikistan (TJK), United Kingdom of Great Britain and Northern Ireland (UK), United States of America (USA), and Uzbekistan (UZB). For this analysis, we used whole-genome SNPs filtered by four criteria, non-indel, biallelic, missing rate <1% and allele frequency >1%. As a result, the allele frequencies of 37,718 filtered whole-genome SNPs were analyzed. We focused on the populations from which at least three individuals were sampled. As a result, 1,073 accessions from 46 populations were included for the analysis.

In the Treemix analysis, we assumed that the relicts were an outgroup of the other non-relicts. These relicts were not sampled from historical sites, but were identified as outlier accessions in the distribution of genetic distances between accessions (the 1001 Genomes Consortium, 2016, Supplementary Figure S2). Regardless of the nomenclature, the set of relicts can be a valid outgroup that is genetically distinct from the non-relicts.

Constructing selection matrices

Adaptation of traits and gene expression was detected by applying PolyGraph (Racimo et al., 2018). In total, the dynamics of allele frequencies at 23,880 QTLs of 174 traits and 8,618 eQTLs of 1,829 genes were analyzed. For each of the traits and gene expressions, PolyGraph contrasts the variation in allele frequencies at QTLs and eQTLs with that at the neutral SNPs, and estimates the selection parameters along the edges of the admixture tree. The selection parameters express the directional changes in allele frequencies at the set of QTLs and eQTLs in relation to the trait values and the levels of gene expressions, beyond the level of genetic drift.

The required input of the eQTL effect sizes in PolyGraph were available from the eQTL dataset. For the QTLs, we regressed the trait value on the alleles of the QTL, for each trait and for each QTL. To account for the effect of population structure, we included the first five principal components of the genotype data of whole-genome SNPs as covariates. For the neutral SNPs, we adopted the 37,718 filtered whole-genome SNPs analyzed by Treemix to estimate the population history. As PolyGraph analyses the variation in allele frequencies between populations, we could not include the QTL/eQTL SNPs that were missing in all individuals within any of the populations. We did not perform PolyGraph analysis for some traits and gene expressions, whose QTLs or eQTLs were all excluded by this filtering.

Since PolyGraph estimates the selection parameters in a Bayesian framework, we obtained the z-value of the selection parameters as the ratio of the posterior mean and the posterior standard deviation from the MCMC samples. Assuming normality of the estimates, we obtained the p-values. For each of the traits and gene expressions, we obtained a map of significant adaptation on the admixture tree using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995). We constructed two types of selection matrices, listing the z-values and the 0-1 values indicating significance. The rows represent the edges of the admixture tree and the columns represent the traits and gene expressions.

Co-adaptation and enrichment analysis

To examine the pattern of co-adaptation, we listed the traits and gene expressions that adapted along each edge of the admixture graph. The size of co-adaptations was measured as the sum of each row of the significance matrix. We characterized the large-scale co-adaptations by enrichment analysis of STRING (Szklarczyk et al., 2023). We also performed enrichment analysis of the causal genes to characterize the molecular background of the QTLs. We used the R package rbioapi (Rezwani et al., 2022) to access the STRING database.

RESULTS

Range expansion, migration, and introgression from relicts
Figure 1 shows the history of population splitting and admixture by TreeMix. The maximum log likelihood value increased with the number of admixture events, but it was difficult to determine the number of admixtures. This was partly because the graph was estimated by maximizing the composite likelihood. As the global pattern of non-migrant edges was qualitatively unchanged (Supplementary Figure S3), we used the graph with ten admixtures for the subsequent analyses of multi-trait and multi-gene expression co-adaptations. It contains four events of introgression from relict to non-relict and seven admixtures between non-relicts, allowing us to assess how these events increased genetic variation in populations and contributed to adaptation to the local environment (Lee et al., 2017).

The horizontal scale of the length of an edge represents the amount of genetic drift, and are interpreted as the ratio of the number of generations to the effective population size. The population in Iberia was consistently clustered with relicts in the admixture trees with different settings for the number of admixtures. The other populations were mostly grouped in the genetic clusters by ADMIXTURE: Italy-Balkan-Caucasus, Western Europe, Germany and Sweden, Central Europe, and the populations in Asia derived from Central Europe. Their vertically straight arrangement indicates their spread in a short period of time, reflecting the history of human activity dispersal (François et al., 2008, Lee et al., 2017). The long edges towards Georgia and Azerbaijan, the United States of America, Sweden, and Asian countries may reflect population bottlenecks in migrations.

The splitting of populations within the same genetic clusters can be interpreted as the migration into an open niche. For example, the splitting between the United States and the United Kingdom in the Western European cluster corresponds to the migration from Great Britain to the American continent. Similarly, the edge to the United States in the German genetic cluster corresponds to the migration from the European continent. In addition, the United States received the immigration from Central Europe. This picture of multiple sources of introduction from across Eurasia (Shirsekar et al., 2021) could not be captured if the countries were treated as the unit of the population (Supplementary Fig. S4).

Identified history of p-adaptations and e-adaptations

We identified 650 phenotypic adaptations (p-adaptations) and 3,925 gene expression adaptations (e-adaptations) in the population history (FDR=0.05, Supplementary Table S1. All the α values (selection parameters) and their z-values are found in Supplementary Data S01 - S04). The identified p-adaptations and e-adaptations at each edge of the admixture graph and the enrichment analysis can be found in Supplementary Data S05 and S06.

Four lineages with large-scale trait and adaptations and gene expression adaptations

Among these, large scale p-adaptations occurred along the two lineages (Figure 2(a)). One is diverged from central Europe and leads to the Asian admixture group: Central Asia (Kyrgyzstan and Tajikistan), and southern Siberia, Russia. The other is the lineage towards Sweden. Three lineages (Figure 2(b)) have also undergone large scale e-adaptation. One is the above lineage towards the Asian admixture group. The others are the lineage from Germany to the United States and the lineage in the Italy-Balkan-Caucasus group east to Azerbaijan. In particular, the eQTLs of more than a hundred genes changed their allele frequencies at each lineage edge in response to the environmental stress (Supplementary Table S2 and Figure S5). In this paper, we focus on these four lineages.

Eastward migration into Central Asia and Southern Siberia, Russia

The *Arabidopsis thaliana* population underwent large-scale adaptation of traits and gene expression along the seven edges leading into Kyrgyzstan, Tajikistan, and Russia (Figure 2, Supplementary Figure S5, Data S05 and S06). The Russian samples in the Asian admixture group were collected in southern Siberia (Supplementary Figure S6). To characterize the adaptations along the lineage, we collected the traits and gene expressions that were adapted on more than half of the edges.

Remarkable p-adaptations are seed dormancy, leaf chlorosis, root mass density and root branching in response to phytohormonal signaling pathways (Figure 3(a), Supplementary Table S2). Cold winters and low
precipitation characterize the climate of the sampling sites in southern Siberia (Supplementary Figure S7). In response to the very cold mean temperature (-21 in January), the population prolonged seed dormancy and delayed germination (Alonso-Blanco et al., 2003; Bewley, 1997). The short growing season required efficient growth, as summer annuals germinate in spring and seed in late summer (Footitt et al., 2013; Huang et al., 2014). The climates of Kyrgyzstan and Tajikistan are very similar, characterized by cold winters (-6.4 to -9.3 from December to February) and high rainfall in spring (314 to 406 mm from March to May), followed by hot and dry summers (28 to 31 and 39 to 41% humidity in July and August) (Supplementary Fig. S7). Plants in the populations reduced evapotranspiration in May and June. They adapted to the dry summer by expanding the root system architecture (RSA): the rate of proliferation and differentiation, the direction of root growth, and the formation of lateral roots (Ristova et al., 2018). It also enabled rapid growth in a short growing season by efficiently taking up the water and nutrients needed for growth.

Enrichment analysis of e-adaptations identified PR1, PNP-A, AED1, RLP23 whose functions are the systemic acquired resistance in response to oomycetes and detection of molecules of fungal origin (Figure 4(a), Supplementary Table S2, Supplementary Data S07). The mean maximum temperature near the sampling sites in July is 24 in Novosibirsk in the southern Siberia, and 29 to 31 in Kyrgyzstan and Tajikistan (Supplementary Figure S7). Fungal and bacterial growth rates had optimum temperatures around 25–30. Growth rates decrease at either higher or lower temperatures. The decrease is more drastic for bacteria at higher temperatures, while it is more drastic for fungi at lower temperatures. As a result, fungi are better adapted to low temperature conditions than bacteria (Pietikäinen et al., 2005). Oomycetes become most aggressive against lucerne when a cool, wet spring is followed by an early, warm, dry summer (Abbas et al 2022). The identified p-adaptation of leaf chlorosis is the penetration resistance in response to microbial invasion (Johansson et al., 2014). It was also an adaptation to the short growing season. By closing the path to the leaves, more nutrients are directed to the seeds and seed maturation is accelerated as the leaves turn yellow during the reproductive process (Huang et al., 2014).

Migration northwards to Sweden

In Sweden, A. thaliana has adapted to cold winters by delaying flowering time (Figure 3(b), Supplementary Table S3, Data S05 and S07). It behaves as winter annual, germinating in autumn, overwintering as a vegetative rosette and flowering in spring or summer (Li et al 2010; Ågren & Schems, 2012; Footitt et al., 2013). The Swedish ecotype is exposed to subfreezing mean monthly minimum temperatures for several winter months (Supplementary Figure S7), and has a higher freezing tolerance. The other major adaptation is to acquire sufficient carbon to support reproduction during the growing season. The Swedish ecotype has extended its life cycle with earlier seedling establishment (early September) and late seed maturation (late June). As a result, it experiences similar temperatures to the Italian ecotype at corresponding developmental stages (Demmig-Adams et al., 2022).

Enrichment analysis of e-adapted genes identified ATH13, DIS1, FAD6, FTSH8, KAC2, NDA1, SIGE, VLN2, VLN3 that have the role of actin filament in response to light stimulus (Figure 4(b), Supplemental Table S3, Data S06 and S07). Actin filaments play an essential role in organelle movement in plants (Kadota et al., 2009). Chloroplasts move in response to external stimuli, in particular light. Low light induces a chloroplast accumulation response so that light is captured efficiently for photosynthesis. Strong light induces an avoidance response to avoid light damage. Overexpression of F-actin genes and p-adaptation of an increased number of leaves allowed efficient photosynthetic activity in response to the low irradiance.

Although the Central Asia and Sweden both have cold winters, A. thaliana has adopted contrasting strategies. The average minimum temperature is lower in Bishkek, Kyrgyzstan (-9.3) and Dushanbe, Tajikistan (-8.2) than that of Stockholm, Sweden (-3.9), where majority of samples were collected in these regions. Bishkek and Dushanbe are dry (39 – 45% humidity in July and August), while Stockholm is humid throughout the year (69 – 87% humidity) (Supplementary Figure S7). Summer is cooler and shorter in Stockholm (maximum temperature is 21 in July) than in Bishkek (29) and Dushanbe (31). Sunlight is stronger in Central Asia. These environmental conditions led to different adaptation strategies, summer annuals with rapid growth in Central Asia and winter annuals with longer life cycles in Sweden.
Migration to Azerbaijan

A large-scale e-adaptation was identified along the lineage in the Italy-Balkan-Caucasus group eastwards to Azerbaijan (Figure 2, Supplementary Figure S5). The p-adaptations imply the climatic change experienced by the population on the way to Azerbaijan (Figure 3(c), Supplementary Table S4, Data S05). The climate of Azerbaijan is very diverse, with nine of the world’s eleven climate zones, including semi-arid, temperate, continental and tundra zones. This results in significant variations in annual temperature and precipitation across the country (https://climateknowledgeportal.worldbank.org/).

Enrichment analysis identified the plant hormone signaling pathways and enhanced responses to external stimuli (ACS6, ERF017, ERF016, PILS7, CAF1b, ECS1, ERF13, GH3.3, HLS1, IAA13, IAA19, IAA5, NHL3, PDF1.3, PROPEP3, RRTF1, RTM3, TIP, WRKY40, WRKY46) (Figure 4(c), Supplementary Table S4, Data S06). Plant hormones regulate various physiological processes, including plant defense. Among them, jasmonate (JA) and salicylic acid (SA) are important defense-related phytohormones. Ethylene (ET), abscisic acid (ABA), auxin, gibberellins (GAs), cytokinins (CKs), and brassinosteroids (BRs), are also involved in defense responses (Kazan & Manners, 2009, Berens et al., 2017).

These adaptations reflect the diversity of insects and microbes generated by the local climate. The sample was taken from the south-eastern corner (Supplementary Figure S8). Lankaran, the sampling site in Azerbaijan, has a hot-summer Mediterranean climate (Csa) and is the wettest province in the country: the mean temperature in July exceeds 26degC, the mean annual precipitation is 1200 mm, and the mean number of rainy days is 110 days; 42% of the annual precipitation and 32% of the rainy days occur between September and November (35 days, 540 mm, Supplementary Figure S7). Based on the latitudinal proximity to the central Italy and similar annual patterns of temperature, precipitation, daylight hours and humidity (figure not shown), the germination and flowering time in Lenkaran are expected to be similar to those in Italy from October to December and February to April (Agren & Schemske, 2012). The population in Lenkaran would be affected by high rainfall before germination and during germination and growth of seedlings. The plant fungal pathogen *Fusarium oxysporum* is the causal agent of root rot or wilt diseases in several plant species. Affected plants (hosts) are mostly from the tropical and subtropical areas, probably because wilt symptoms are more pronounced at elevated temperatures (Berrocal-Lobo & Molina, 2008).

Migration of German population into the United States

The sample from the United States was mainly of Germany (Supplementary Figure S9). The adapted traits of the accompanying weeds were days to flowering, bacterial disease resistance, leaf cadmium concentrations, and rolled leaf (Figure 3(d), Supplementary Table S5, Data S05). The selection for a reduction in leaf cadmium concentrations indicated that the plants were exposed to heavy metals. The enrichment analysis of the QTL-coding genes of the cadmium concentrations in leaves (Cd111) indicated that P-type ATPase, transmembrane transporters reduced their activities (Supplementary Table S6, Data S05, S07). Cadmium is mainly released from nickel-cadmium batteries, fossil fuel combustion, coating and plating, and cement production (Ruchi Bharti & Sharma, 2022). In fact, the samples were mainly collected from the coastal areas of Lake Michigan and near New York City (Supplementary Figure S9). The state of Michigan produces cement and has a mining industry.

The heavy metal stress selected the alleles with lower germination rate, lower leaf number, slower growth and flowering, and rolled leaves, providing a chance for succession in an improved environment (Li et al., 2005). Notably, alleles with weak resistance to bacteria were selected. This is probably because the heavy metals affected not only the plant individuals but also plant pathogens. As a result, the plants had a reduced risk of bacterial infection. As disease resistance comes at some cost (van Hulten et al., 2006), alleles with lower resistance were selected in the absence of infection risk.

Enrichment analysis of the e-adapted genes identified the selection of alleles with higher expression of the photosynthesis-related genes (AT2G12905, ATPA, ATPH, MATK, ORF31, PB, PETD, PETG, PSAB, PSAJ, PSBE, PSBF, PSBI, PSBJ, PSBK, PSBL, PSBT, RBCL, SIGE, YCF10) (Figure 4(d), Supplemental Table S5, Data S06). *Arabidopsis* is a C3 plant that performs C3-type photosynthesis, as do algae such as
Chlorella and many plants such as rice, wheat, soybean, rapeseed, and spinach. Unlike C4 plants, photosynthesis in C3 plants is not carried out by the division of labour between leaf pith cells and vascular sheath cells. They are therefore less able to capture $\text{CO}_2$ under conditions that tend to close their stomata, such as high temperatures and drought. Photosynthesis is less efficient in climates that are hostile to plants, such as high temperatures, drought, low $\text{CO}_2$, and low nitrogen soils (Brown, 1978, Waller & Lewis, 1979). In addition, the heavy metals weaken the photosynthetic activities. Populations have adapted to this activity-depressing pressure by selecting for alleles with higher expression of photosynthetic-related genes.

### Selection on the haplotypes

Although *A. thaliana* is a selfing weed, linkage disequilibrium decays rapidly, within 50 kb (Nordborg et al., 2005). Notably, however, we further identified selection on the haplotypes of the adaptive QTLs and eQTLs. Selection was particularly strong along the migration lineage from Germany to the United States. Here we describe the selection on the haplotypes of the QTLs and eQTLs, focusing on the observation of this lineage.

Each QTL/eQTL has H-alleles (higher value/higher expression) and L-alleles (lower value/lower expression). We simply counted the proportions of H-alleles for each of the identified traits and gene expressions for each individual. As above, the enrichment analysis of e-adaptations identified 20 photosynthesis-related genes, AT2G12905, ATPA, ATPH, MATK, ORF31, PB, PETD, PETG, PSAB, PSJ, PSBE, PSBF, PSBI, PSBJ, PSBK, PSBL, PSBT, RBCL, SIGE, YCF10. They had a total of 108 eQTLs, widely distributed on chromosomes 1 and 2. Of these, 51 significantly changed allele frequencies (FDR=0.05) along this lineage.

Out of 123 individuals in the US sample, 84 individuals (68%) had H-alleles at all eQTLs (Figure 5 (a)). We call this haplotype as a full haplotype. These 84 individuals were all from the German admixture group (Supplementary Figure S10). In sharp contrast, in the sample from outside the United States, 834 out of 1032 individuals (82%) had L-alleles (lower expression) at all eQTLs (Figure 5 (b)). However, in the German sample, no individual had more than a few H-alleles. No individual in the UK sample had the full haplotype either, although a few individuals had H-alleles at about 80% of the eQTLs.

The haplotypes of photosynthesis-related eQTLs received a strong selection for increasing H-alleles in the coastal area of Lake Michigan, the urban areas near New York City, and the arid area of California (Figure 6(a)). At the same time, these areas showed strong selection on alleles of numerous QTLs for decreasing heavy metal uptake activity (Figure 6(b)), increasing the length of growing season (Figures 6(c) and (d)), increasing rolled leaf (Figure 6(e)), and decreasing bacterial resistance activity (Figure 6(f)).

### DISCUSSION

We identified 650 p-adaptations and 3,925 e-adaptations in the population history of *A. thaliana* (Supplementary Table S1, Data S05 and S06). In this paper, we focused on the four lineages that experienced a large-scale p-adaptations and e-adaptations and characterised the adaptations (Supplementary Data S07). In these lineages, the relative fitness of the alleles at many QTLs and eQTLs changed. The populations also seemed to optimize the site frequency spectrum of these loci to fit to the environment. In the eastward migration to Central Asia and southern Siberia, the extremely cold winter and short summer prolonged seed dormancy, and expanded the root system architecture. In the northward migration to Scandinavia, low temperatures prolonged the growing season and low light intensity required the increased chloroplast activity. The subtropical and humid environment of south-eastern Azerbaijan enhanced phytohormone signaling pathways in response to the biotic and abiotic stresses.

The immigrants from Germany adapted to the heavy metal exposure and arid environment of the United States by selecting the eQTL alleles that increase photosynthetic activity, the QTLs that reduce the growth rate and those that reduce the weight of bacterial resistance. Figure 6 suggests that the population had completed the full set of adaptations by the time it reached the east coast and that the adaptive haplotypes spread westwards. However, these haplotypes were not present in the German sample. Instead, these haplotypes may have been introduced by some of the immigrants from England (Supplementary Figure S10). An alternative tempting scenario is that an environmental change in Germany wiped out the haplotypes, but
we could find no evidence for it. It is still very difficult to get a clear picture of adaptation at the molecular level. Focused study of the North American A. thaliana population revealed that multiple introductions into a non-native range have allowed a rapid adaptation of the colonizing species by increasing haplotypic diversity through admixture (Shirsekar et al., 2021). It is interesting to scrutinize the large-scale p-adaptations and e-adaptations observed in our analysis with reference to result from the focused study with enhanced fine scale population genomic data.

Enrichment analysis was useful to characterize the sets of e-adaptations in response to the local environments. It was also useful to characterize the p-adaptations. Alleles at the QTLs in the coding regions may affect the traits mainly by altering the physical and chemical properties of the relevant proteins, while those in the intergenic regions may affect the traits by altering the gene expression. In this paper, we did not fully analyze the latter type of QTLs to characterize the molecular mechanisms behind the p-adaptations, but the enrichment analysis of the QTL-accommodating genes could interpret the p-adaptations of cadmium concentration in leaves in the United States as the reduced activities of the transmembrane transporters.

It should be noted that the local environment may not necessarily select for the alleles of the key genes, as its changes may have deleterious effects on the pathways. For example, in the anthocyanin biosynthesis pathways, upstream genes evolved more slowly than the downstream genes (Rausher et al., 1999). It should also be noted that comparing phenotypes and gene expressions between populations does not necessarily identify adaptation, because the observed phenotypes and gene expressions are the result of a balance between the environmental stresses and the genetic responses. The directional changes in the allele frequencies at the QTLs/eQTLs provide direct evidence of population adaptations.

Based on the database of QTLs (Togninalli et al., 2020) and eQTLs (Lan et al., 2021) and the population genomic data (The 1001 Genomes Consortium, 2016), we constructed a database of p-adaptations and e-adaptations in the population history of A. thaliana, which were characterized with reference to the information on the local environments. It complements studies focusing on specific aspects of adaptation. Although traits without identified QTLs and genes without identified eQTLs were not included in the analysis, it provided us a clue for understanding the complex history of biological adaptation. The information on p-adaptation is directly interpretable. The information on e-adaptation may provide a less biased picture of physiological adaptation. Both have been essential in characterizing the adaptations in the four lineages. For example, in the northward migration to Scandinavia, the p-adaptations confirmed that the population adapted to the cold winter by extending the growing season. Importantly, the e-adaptations indicated a physiological adaptation to the low sunlight intensity in the northern area. By increasing the expression of actin filaments, the chloroplasts in the leaves were able to move smoothly in response to the sunlight. The current study could not integrate the p-adaptations and e-adaptations into a unified framework. The database of p-adaptations and e-adaptations, which are publicly available (supplementary Data), may be useful for future studies to obtain a complete picture of the biological adaptations of A. thaliana throughout its population history.

AUTHOR CONTRIBUTIONS

H.K. conceived, designed the study, and wrote the manuscript. R.N. constructed screened datasets for this study and performed a high-throughput analysis of population history to generate the selection matrices. H.K. statistically analyzed the selection matrices and constructed the database of the p-adaptations and e-adaptations. R.N. and H.K implemented the scripts of this study, and R.N. revised the scripts for publication. H.K. and S.K. characterized the identified p-adaptations and e-adaptations. S.K. contributed to the interpretation of the results and reviewed the manuscript. All authors approved the final version.

ACKNOWLEDGEMENTS

This work was supported by the Japan Society for the Promotion of Science Grant-in-Aid for Scientific Research (KAKENHI) No. 22K11950.

CONFLICT OF INTEREST STATEMENT
The authors declare no conflict of interest.

**DATA AVAILABILITY STATEMENT**

The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures, tables, supplementary material, and supplementary data. The screened data for this study (genotype, trait, environment and gene expression), the estimated population history and selection parameters, the identified phenotypic adaptations (p-adaptations) and gene expression adaptations (e-adaptations), with the result of gene annotation and enrichment analysis, and the scripts for the analysis, are available at https://zenodo.org/record/7903202.

**ORCID**

Hirohisa Kishino https://orcid.org/0000-0002-3244-359X  
Reiichiro Nakamichi https://orcid.org/0000-0001-5789-7689  
Shuichi Kitada https://orcid.org/0000-0001-5838-0374

**REFERENCES**


Johansson, O. N., Fantozzi, E., Fahilberg, P., Nilsson, A. K., Buhot, N., Tor, M., & Andersson, M. X. (2014). Role of the penetration-resistance genes PEN1, PEN2 and PEN3 in the hypersensi-


Figure 1 The admixture graph of *Arabidopsis thaliana*. The history of population divergence and admixture was estimated by Treemix. The graph assumes ten admixture events. The admixture graph was defined by the genetic clusters of ADMIXTURE (k=9, The 1001 Genomes Consortium 2016) subdivided by country (see Method).

Figure 2 The number of p-adaptations and e-adaptations along the edges of the admixture graph. (a) The number of p-adaptations, (b) The number of e-adaptations. The edge colour of the admixture tree ranges from blue (minimum) to red (maximum). The edge width represents the migration weight.

Figure 3 The p-adaptations along the four lineages. Mean selection coefficients within the lineages are shown. (a) diverged from central Europe and led to the Asian group, which then led to Russia, (b) led to Sweden, (c) in the Italy-Balkan-Caucasus group led east to Azerbaijan, (d) led to the United States.

Figure 4 The enrichment analysis of e-adaptations along the four lineages. Mean selection coefficients within the lineages are shown. (a) diverged from central Europe and led to the Asian group, which then led to Russia, (b) led to Sweden, (c) in the Italy-Balkan-Caucasus group led east to Azerbaijan, (d) led to the United States.

Figure 5 Haplotypes of the photosynthesis-related eQTLs, whose allele frequencies changed significantly (FDR=0.05) along the lineage to the United States. Each horizontal line of dots represents an individual’s haplotype. (a) individuals from the United States sample, (b) a high-expression allele (H-allele) and a low-expression allele (L-allele) from the mixed population of northern Europe and the Middle East, (c) a high-expression allele (H-allele) and a low-expression allele (L-allele) from the Mediterranean basin, (d) an H-allele (higher expression) and an L-allele (lower expression). Red and blue dots represent the H-alleles and L-alleles respectively.

Figure 6 The geographical map of the USA sample. The colours of the points represent the proportion of H-alleles (blue for 0% and red for 100%) for (a) eQTLs of bolting height (number of days required for the bolt height to reach 5 cm), (b) eQTLs of leaf color, (c) QTLs of rolled leaf, (d) QTLs of bolting height, (e) QTLs of apical dominance, (f) QTLs for bacterial resistance.