Does selection occur at the intermediate zone of two insufficiently isolated populations? A whole-genome analysis along an altitudinal gradient

Naofumi Yoshida (naofumi.yoshida.s2@dc.tohoku.ac.jp)
Tohoku University Graduate School of Life Sciences: Tohoku Daigaku Daigakuin Seimei Kagaku Kenkyuka https://orcid.org/0000-0003-0215-9608

Shin-Ichi Morinaga
Nihon Daigaku Seibutsu Shigen Kagakubu

Takeshi Wakamiya
Tohoku Daigaku Daigakuin Seimei Kagaku Kenkyuka

Yuu Ishii
Tohoku Daigaku Daigakuin Seimei Kagaku Kenkyuka

Shosei Kubota
Tokyo Daigaku Daigakuin Sogo Bunka Kenkyuka Kyoyo Gakubu

Kouki Hikosaka
Tohoku Daigaku Daigakuin Seimei Kagaku Kenkyuka

Research Article

Keywords: Local adaptation, selection, gene flow, phenotypic divergence, whole-genome sequences, homozygote

Posted Date: April 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-447152/v1

Abstract

Adaptive divergence would occur even between the insufficiently isolated populations when there is a great difference in the environments of their habitats. The individuals present in the intermediate zone of the two divergent populations are expected to have an admixed genetic structure due to gene flow. A selective pressure that acts on the genetically admixed individuals may limit the gene flow and promote the adaptive divergence. Here, we addressed a question whether the selection occurs in the genetically admixed individuals between the divergent populations and assessed its effects on the population divergence. We obtained the whole-genome sequences of a perennial montane plant, *Arabidopsis halleri*, which has clear phenotypic dimorphisms between altitudes, along an altitudinal gradient of 359–1,317 m with a high spatial resolution (mean altitudinal interval of 20 m). We found the zone where the highland and lowland genes were mixing. Using the F_{ST} approach, we found that 35 and 13 genes in the admixed zone had a high frequency of alleles that are accumulated in highland and lowland subpopulations, respectively, suggesting that these genes have been selected in the admixed zone. This selection might limit the gene flow and contribute to the adaptive divergence along the altitudes. In the single-nucleotide polymorphism (SNP)-based analysis, 3,000 out of 27,792 Altitude-Dependent SNPs had extremely high homozygosity in the admixed zone. In 84.7% of these SNPs, the frequency of homozygotes of highland alleles was comparable to that of lowland alleles, suggesting that these alleles are neutral but the heterozygotes were selectively eliminated. The distribution of highland and lowland alleles of these SNPs was not clearly separated between altitudes, implying that such selection did not impede the gene flow. We conclude that the selection occurring at the intermediate altitude affects the genetic structure in the admixed zone and adaptive divergence along the altitudes.

Introduction

Species having broad distributions would face different selections in each habitat and often obtain population-specific polymorphisms on various traits and genes as a result of their local adaptation to distinct environments (Pruisscher et al. 2018; Campbell-Staton et al. 2018). Such adaptive divergence is more likely to occur between the geographically isolated populations because geographical distance can provide a strong reproductive barrier and shape large environmental differences (Galloway and Fenster 2000; Kubota et al. 2015). Even with a short distance between the populations, adaptive divergence may occur if there is a great difference of environments between them (Skelly et al. 2004; Antonovics et al. 2006; Hämälä and Savolainen 2019). The classical expectation has indicated a negative role of gene flow in adaptive divergence; if the populations are not sufficiently isolated and experience strong gene flow, maladapted genetic variation would be introduced from one population to another and their local adaptation would be impeded (Lenormand 2002). However, if selective pressure in each environment is strong enough, those maladapted genetic variations would be removed (Bisschop et al. 2020). Therefore, on a small scale, the genetic structure would be established on a balance between the two competing evolutional powers, gene flow and selective pressure, in each habitat (Slatkin 1987).

Genetic exchange between two populations may mainly occur through individuals located in the intermediate zone of the two populations. The selection and gene flow around such admixed individuals in the intermediate zone may play important roles for the divergence of the two populations. Following scenarios may be considered on the genetic structure in the admixed individuals of the two populations *X* and *Y*. Scenario 1: If one of the alleles (allele *x*) of a gene is adaptive in population *X* and the other (allele *y*) is adaptive in population *Y*, but the two alleles have similar influence on the fitness in the intermediate zone, there is no selection on the gene and the accumulation of the two alleles in the admixed individuals is influenced mainly by gene flow. Scenario 2: If the allele *x* is adaptive not only in the population *X* but also in the intermediate zone whereas the allele *y* is adaptive only in the population *Y*, the allele *x* accumulates in the admixed individuals. Scenario 3: If an allele *z* is adaptive only in the intermediate zone and maladaptive in the populations *X* and *Y*, and it accumulates in the individuals located in the intermediate zone. Focusing on heterozygosity, genes belonging to the Scenario 1 may further be divided into three groups: fitness of genotypes *xx*, *xy*, and *yy* is similar to each other and they are randomly mixed (Scenario 1a), fitness of homogeneity genotypes *xx* and *yy* are similar to each other and higher than that of heterogeneity genotype *xy* (Scenario 1b), and fitness of heterogeneity genotype *xy* is higher than that of homogeneity genotypes (Scenario 1c). Under the Scenario 1c, the selection favoring heterozygotes would promote the admixture of polymorphism and prevent adaptive divergence of the two populations. Conversely, under the Scenario 2, the selection favoring one of the alleles would overwhelm gene flow, which restricts the admixture of polymorphism and acts as a potential driver of population adaptive divergence. Under the Scenario 3, the selection favoring peculiar genes to the environment at the intermediate zone would not relate with adaptive divergence between two edge populations, but would take an important role in the evolution in the intermediate population.

A number of studies have demonstrated that the genetic admixture occurred between phenotypically and/or genetically diverged populations (Ohtani et al. 2013; Richardson and Urban 2013; Le Moan et al. 2016; Puckett et al. 2016; Lipshutz et al. 2017), and individuals in such admixed populations often have intermediate phenotypes of the two populations (Stacy et al. 2016; Hendrick et al. 2016; Linnen et al. 2013). These results would be consistent with Scenario 1a. In contrast, other studies showed that the selection maintains lower recombination rates or genetic diversity on the particular genetic regions in the hybrid populations even under strong gene flow, which would be consistent with Scenario 1b or 2 (Comeault et al. 2015; Hämälä and Savolainen 2019). Heterosis, defined as a vigorous growth in the hybrid offspring of genetically distant individuals relative to their homozygous parents, has been shown in some species, which would be consistent with Scenario 1c (Facon et al. 2005; Li et al. 2018). However, which of these scenarios is applicable to genes in the intermediate population is poorly understood. In particular, the degree of heterozygosity and kind of genes that are selected in the intermediate zone have not been studied yet. Furthermore, a high spatial resolution sampling of genetic data is necessary to identify the zone where the admixture of gene occurs, but previous studies have not conducted such sampling.

The altitudinal adaptive divergence is one of the fascinating materials. There are steep environmental gradients along the altitudes to study the fine-scale local adaptation. For instance, temperature and the

length of growing season regularly decline with increasing elevation (Körner 2007). Many plant species show intraspecific variations along the altitudinal gradient. For example, with increasing altitude, *Metrosideros polymorpha* increases leaf mass per area, thereby enhancing tolerance to cold (Cordell et al. 1998). *Fallopia japonica* increases their flavonoid contents (Murai et al. 2015) and decreases the optimal temperature of photosynthesis (Machino et al. 2021). Such a small-scale altitudinal divergence has also been reported from a perennial montane plant, *Arabidopsis halleri* subsp. *gemmifera*. In Mt. Ibuki, a mountain located in Central Japan, *A. halleri* plants are distributed along the broad altitudinal gradient. Although the horizonal distance between the top and bottom populations is small (< 3 km), there are various phenotypic differences between highland and lowland ecotypes. Highland ecotypes are characterized by dense trichomes on the leaves, whereas lowland ecotypes have glabrous leaves (Fig. 1a, b). Physiological differentiations have also been reported for the tolerance to UV radiation, the response of biomass allocation to soil nutrient, and the water repellency of leaves (Wang et al. 2016, 2019; Aryal et al. 2018). Analyzing the whole-genome sequences, Kubota et al. (2015) found unidirectional allele frequency shifts along the altitudes in many genes; however, there is a relatively small genetic differentiation between highland hairy ecotype and lowland normal ecotypes (Ikeda et al. 2010; Kubota et al. 2015). The flowering time of lowland population is from the end of April to the end of May, whereas that of highland population is from the middle of May to the middle of June. In intermediate altitudes, plants with scarce trichomes on the leaf surface are often observed, suggesting that these plants have an intermediate phenotype between highland and lowland ecotypes. However, the genetic structure of plants inhabiting the intermediate altitudes individuals has not been studied yet.

In this study, we addressed a question how the selection and gene flow shape the genetic structure in the admixed individuals between two divergent populations. To answer this question, we sampled *A. halleri* individuals with a very high spatial resolution (every 20 m on average from 359 m to 1,317 m above the sea level, Fig. 1c) and analyzed their whole genome. First, we identified the areas where the genetic admixture of highland and lowland ecotypes mainly occurs. Second, we investigated allele frequency to find genes that selected in the admixed zone according to the above-mentioned scenarios. If an allele that is adaptive to the highland environment is also favored in the admixed zone, its frequency in the admixed zone may be similar to that in the highland but higher than that in the lowland (Scenario 2H in Fig. 2). *Vice versa* if an allele adaptive to the lowland environment is favored in the admixed zone (Scenario 2L). If there is an allele that is adaptive only in the admixed zone, its frequency may be higher in the intermediate zone than that in the highland and lowland (Scenario 3). If heterozygote of the two alleles are eliminated in the admixed zone, homozygote of each allele may be more frequent than the expected from the Hardy–Weinberg equilibrium (Scenario 1b). In contrast, if heterozygote is advantageous, heterozygote of the two alleles may be more frequent than the expected from the Hardy–Weinberg equilibrium (Scenario 1c). We investigated what kind of genes are selected in the admixed zone according to these scenarios.

Materials And Methods

Species and Study Sites

Arabidopsis halleri subsp. *gemmifera* is a diploid (2n = 16), self-incompatible, and perennial montane plant (Al-shehbaz and O'kane 2002; Kolnik and Marhold 2006). In Japan, this plant is distributed in a wide range of altitudinal and latitudinal gradients. Its leaves are generally glabrous, but the ecotypes in the highland areas in Mt. Ibuki and Mt. Fujiwara in central Japan have dense trichomes on the leaves and stems (Figs. 1a, b). A previous genome-wide association analysis suggested that the two highland ecotypes at Mt. Ibuki and Mt. Fujiwara evolved independently from each other (Kubota et al. 2015), though they have similar morphological characteristics.

We used plants growing in Mt. Ibuki, where the highland habitats are characterized by relatively low vegetation heights, bright environment near the ground, and heavy snow in winter, whereas the lowland habitats are characterized by dark forest floor and relatively mild winter weather (Honjo and Kudo 2019). We harvested the leaf samples from an individual plant at 48 positions along the altitude of Mt. Ibuki in 2007 and 2008 (Table S1). The lowest and highest sampling sites were at 359 m and 1,317 m above the sea level, respectively, and their horizontal distance was approximately 2.8 km (Fig. 1c). To avoid the sampling of same clones, the sampling positions were at least 3 m apart from each other. The mean interval of the altitude and horizontal distance between the sampled plants was 20.4 m and 59.6 m, respectively. We also used genome information reported in Kubota et al. (2015), which was obtained from *A. halleri* plants growing at altitude of 380, 600, 1,000 and 1,250 m (five plants per site) in 2009 and 2010.

DNA Extraction, Individual-based Sequencing, and Data Processing

We extracted the genomic DNA from the dried leaf samples of collected 48 individuals using the DNeasy Plant Kit (QIAGEN). Thereafter, we prepared the DNA libraries using the TruSeq Nano DNA Low Throughput Library Prep Kit (Illumina). We generated reads using the Illumina HiSeq X Ten, and obtained 270 Gb of data from the 48 samples. The genome size of *A. halleri* is estimated to be 250 Mb (Briskine et al. 2017), suggesting that the average coverage of our sequence data would be more than 22X. These raw read sequences are available in the DNA Data Bank of Japan Sequenced Read Archive under the accession number DRA010696. Furthermore, we added previously posed sequence reads of 20 individuals collected at the altitudes of 380, 600, 1,000 and 1,250 m on Mt. Ibuki (Kubota et al. 2015). We trimmed the low-quality reads (more than half of the nucleotides with quality score less than 30) using FASTX-toolkit v0.0.14 (http://hannonlab.cshl.edu/fastx_toolkit). After the trimming, we mapped the reads of total 68 individuals to the reference genome of *A. halleri* (Briskine et al. 2017) for each individual by the alignment algorism, BWA-MEM v0.6.2 (Li 2013) with the default parameters. We removed PCR duplication by SAMtools v0.1.8 (Li et al. 2009) rmdup. We employed SNP calling using SAMtools mpileup and bcftools v0.1.8 (Li 2011). We trimmed loci whose coverage depth was lower than 4 or higher than 200 by bcftools varFilter. After generating the personal SNP data, we combined them using vcf-merge (VCFtools v0.1.15) (Danecek et al. 2011).

Population Structure Analysis

To estimate the population structure of 68 individuals that were distributed continuously along the altitudes at Mt. Ibuki, we employed genetic clustering analysis with ADMIXTURE v1.3.0 (Alexander et al.

2009). For the population structure analysis, we only considered the SNPs that showed a minor allele frequency (MAF) > 3%. Furthermore, we removed the loci that were in the linkage disequilibrium with each other by plink –indep-pairwise 50 10 0.1 (plink v1.90b4) (Chang et al. 2015). For each value of *K* (the number of subpopulations) ranging 1 to 5, we performed independent runs. To determine the optimal number of subpopulations for the 68 individuals, we calculated the cross-validation (CV) error for each *K* value. The CV error would be minimized when the number of *K* was best or appropriate for the data. Based on this result, we divided the individuals into highland, intermediate, and lowland subpopulations.

Detecting altitude-dependent genomic region

We used F_{ST} approach to detect the genomic regions that accumulated in the intermediate subpopulation in relation to the Scenario 2 and 3. We required MAF > 5% for all SNPs. Using vcftools, we calculated the window-averaged F_{ST} values in each 10 kbp window between three combinations of subpopulations, highland and lowland, highland and intermediate, intermediate and lowland (F_{ST_HL}, F_{ST_HI} and F_{ST_IL}, respectively). We classified windows exhibiting extreme values of F_{ST} as outliers, defined as the higher 1% quantile for each test, which contained 207 genomic windows. Genomic windows of each outlier were expected to diverge significantly between subpopulations. First, we selected windows that had higher F_{ST_HL} values (included in the higher 1% quantile) as the altitude-dependent windows (ADW). Then, among ADWs, windows with extremely high F_{ST_IL} (included in the higher 1%) and low F_{ST_HI} (included in the lower 70%) were detected, which were considered to be consistent with the Scenario 2H (Table 1). Similarly, among ADWs, windows with extremely high F_{ST_HI} and low F_{ST_IL} were also detected, which were considered to be consistent with the Scenario 2L (Table 1). Alternatively, we selected windows that had lower F_{ST_HL} values. Among them, windows with higher F_{ST_IL} and F_{ST_HI} were sought, which were considered to be consistent with the Scenario 3 (Table 1). In the present study, we did not identify windows that are consistent with the Scenario 1 and that are included in the regions A and B in Fig. 2, because we were interested in the genes under the selective pressure in the intermediate subpopulation.

Homozygosity and Heterozygosity Analysis for Altitude-Dependent SNPs

We also used the latent factor mixed models (LFMM) (Frichot et al. 2013) to detect the Altitude-Dependent SNPs genetic variant at the SNP level. In this analysis, we considered only SNPs with MAF > 5% following the manual of LFMM. For the following analyses, we considered bi-allelic SNPs only. For each value of *K* (latent factor) = 1 and 2, we performed five independent runs. All the results were integrated by the Fisher's method (Fisher 1932), and SNPs whose false discovery rate (FDR) was below 0.05 were detected as Altitude-Dependent SNPs.

We sought Altitude-Dependent SNPs showing extremely high or low heterozygosity in the intermediate subpopulation in relation to the Scenario 1b and 1c. We calculated an index ΔH, which represents the difference between the expected and observed heterozygosity (H_{EXP} and H_{OBS}, respectively). We assumed that the allele frequency and genotype frequency of a neutral SNP in the intermediate subpopulation are in the Hardy–Weinberg equilibrium and those of SNPs selected in the intermediate subpopulation are

deviated from the equilibrium. Therefore, for each bi-allelic SNP, an expected allele frequency in the intermediate subpopulation (AF_{M_EXP}) was calculated as the average of observed allele frequencies in the highland and lowland subpopulations (AF_H and AF_L, respectively).

$$AF_{M\_EXP} = (AF_H + AF_L) / 2$$

Using AF_{M_EXP}, the expected value of heterozygosity (H_{EXP}) was calculated by assuming the Hardy–Weinberg equilibrium.

$$H_{EXP} = 2 \times (AF_{M\_EXP}) \times (1 - AF_{M\_EXP})$$

The differences between the observed (H_{OBS}) and expected heterozygosity in the intermediate subpopulation were calculated for each SNP.

$$\Delta H = (H_{OBS} - H_{EXP})$$

ΔH changes between −1.0 and +1.0 and is higher if the heterozygosity is large. We defined the Homozygote- and Heterozygote-selected SNPs that have ΔH value smaller and larger than 2.5% of total SNPs, respectively. Then, we sought the overlap between the Altitude-Dependent SNPs and the Homozygote- or Heterozygote-selected SNPs in the intermediate subpopulation, which were defined as Homozygote- and Heterozygote-accumulated Altitude-Dependent SNPs in the intermediate subpopulation, respectively.

To find genes that are consistent with the Scenario 1b (highland and lowland alleles have similar influence on the fitness in the intermediate zone but their heterozygotes have lower fitness than the homozygotes), we further assessed frequency of the homozygotes of the highland or lowland alleles in the intermediate subpopulation. We calculated an index ΔGF by the following equation with the genotype frequency of homozygote of highland and lowland alleles in the intermediate subpopulation (GF_{H_homo} and GF_{L_homo}, respectively).

$$\Delta GF = (GF_{H\_homo} - GF_{L\_homo}) / (GF_{H\_homo} + GF_{L\_homo})$$

ΔGF would be large (Max = 1.0) if the homozygote of highland allele was abundant in the intermediate subpopulation, and be smaller (Min = −1.0) if the homozygote frequency of lowland allele was large. In the scenario 1b, ΔGF would be neither extremely large nor small due to the selection not favoring one of the homozygotes of highland or lowland alleles.

Functional Annotation of Genes Including Candidate SNPs or Genetic Regions

We investigated functional genes included in the identified genomic regions by annotation to General Feature Format (GFF) file of the reference genome of *A. halleri* (Briskine et al. 2017). We considered genes that contained one or more SNPs within their coding regions. We also investigated whether the SNPs are synonymous replacement, non-synonymous replacement and intergenic variant, and removed

synonymous variants from candidate SNPs. A further functional annotation of detected genes in both methods was employed with gene ontology (GO) analysis by PANTHER 15.0 using the reference gene list of *A. thaliana*.

Results

Population structure

The ADMIXTURE analysis showed that the CV error was minimum when $K = 1$ (Fig.3a), indicating that the population was not clearly differentiated. However, the CV errors were similarly low when $K = 2$ or 3, suggesting that the population was weakly differentiated. When $K = 2$ was adopted, the divergence was clearly found along the altitude; individuals inhabiting altitude below 700 m were occupied by one group (blue in Fig. 3b), whereas those above 1000 m were occupied by the other (red). There were admixed individuals between their two ecotypes in the intermediate altitudes (Fig. 3b). When $K = 3$ was adopted, the third group was found mainly in the plants investigated by Kubota et al. (2015) (Fig. S1a). Further analysis revealed that the third group are included only when the size of bam file (compressed file to save alignment information of short reads mapped against reference sequence generated by Samtools) is smaller than 5 GB across all data (Fig. S1b). We considered this grouping as an artifact due to a variation in the data size and did not use in following analyses. Based on the result of $K = 2$, we defined following three subpopulations: lowland subpopulation (29 individuals at 359~687 m), which had lowland genotypes; intermediate subpopulation (16 individuals at 724~1,000 m), which had mixed genotypes; and highland subpopulation (23 individuals at 1,051~1,317 m), which had highland genotypes. The intermediate subpopulation was considered as the admixed subpopulation.

Genes matching the Scenarios 2 and 3

We obtained total 7,019,253 SNP loci by the whole-genome resequence. Selecting bi-allelic SNPs only, and eliminating SNPs with the extreme coverage depth (lower than 4 or higher than 200) or the low MAF in the 68 individuals (≤ 5%), we used 2,052,011 SNPs in the following analysis.

We detected 24 out of 207 windows whose window-averaged F_{ST_HL} and F_{ST_IL} were large and $F_{ST_HI}$was small, which are expected to match the Scenario 2H (S2H windows) (Fig. 4, Table 1). Total 35 genes were included in 20 S2H windows (S2H genes) (Table S2). 828 out of 1,038 SNPs that located in coding regions of 35 S2H genes were non-synonymous variants. We also detected 8 windows whose window-averaged F_{ST_HL} and F_{ST_HI} were large and $F_{ST_IL}$was small, which are expected to match the Scenario 2L (S2L windows) (Fig. 4, Table 1). Total 13 genes were contained in 7 S2L windows (S2L genes) (Table S2). 477 out of 683 SNPs that located in coding regions of 13 S2L genes were non-synonymous variants. We could not detect any window whose F_{ST_LI} and F_{ST_HI} were large and $F_{ST_HL}$was small, which are expected to match the Scenario 3 (Fig. 4, Table 1). However, when we relaxed the thresholds of FST outliers from the higher 1% to the higher 1.67%, we found one window matching to the Scenario 3.

Genes matching the Scenarios 1b and 1c

The LFMM analysis detected 30,417 and 30,969 significant SNPs that were associated with altitude under the latent factor $K = 1$ and under $K = 2$, respectively (FDR < 0.05). Among these SNPs, 27,792 SNPs that overlapped between the results of $K = 1$ and 2 were defined as Altitude-Dependent SNPs.

We calculated the difference between the observed and expected heterozygosity (ΔH) for the whole-genome 2,052,011 SNPs. We defined the top 2.5% (51,300 SNPs) and bottom 2.5% SNPs (51,300 SNPs) as Heterozygote- and Homozygote-accumulated SNPs, respectively (Fig. 5a). In total, 3,000 SNPs of Homozygote-accumulated SNPs overlapped with Altitude-Dependent SNPs and 73 of those were synonymous variants and remaining 2,927 SNPs were defined as the Homozygote-accumulated Altitude-Dependent (HAAD) SNPs (Fig. 5b). There were no Heterozygote-accumulated SNPs that overlapped with Altitude-Dependent SNPs, suggesting that no genes match the Scenario 1c.

According to the index ΔGF, which represents the difference between the genotype frequencies of the homozygote of highland or lowland alleles (H or L alleles, respectively), HAAD SNPs were classified into the Highland- (ΔGF > 0.5; 348 SNPs), the Lowland- (ΔGF < −0.5; 97 SNPs), and the Coexisting-HAAD SNPs (−0.5 ≤ ΔGF ≤ 0.5; 2,482 SNPs) (Fig.6). Coexisting-HAAD SNPs are considered to match the Scenario 1b.

Based on the average of ΔGF of HAAD SNPs in each gene, we sought genes that tend to have the Coexisting-HAAD SNPs (−0.5 ≤ mean ΔGF ≤ 0.5) and identified 77 genes that are considered to match the Scenario 1b (S1b genes) (Table S3). 174 out of 2,482 Coexisting-HAAD SNPs located in the coding regions of 77 S1b genes. 32 out of 174 HAAD SNPs in S1b genes were missense variants and there was no nonsense variant (Table S3). 2,002 Coexisting-HAAD SNPs located in the intergenic region.

We also identified 22 genes that tend to have the Highland-HAAD SNPs (mean ΔGF > 0.5, the Highland-homozygote genes) and 3 genes that tend to have the Lowland-HAAD SNPs (mean ΔGF < −0.5, the Lowland-homozygote genes) (Table S3). Highland- and lowland-HAAD SNPs were expected to match the Scenario 2H and 2L, respectively, but there were only three Highland-HAAD SNPs that were contained in the S2H windows (Table 1) and no HSAD SNPs that were contained in the S2L windows.

We also identified genes whose coding regions located near (< 1 kb) the HAAD SNPs. 286 genes near the Coexisting-HAAD SNPs (Side-S1b genes), 64 genes near the Highland-HAAD SNPs (Side-Highland genes) and 14 genes near the Lowland-HAAD SNPs (Side-Lowland genes) were identified (Table S4). Because the different types of HAAD SNPs located near genomic position, some genes were included in several groups (11 genes belonged Side-S1b and Side-Highland genes, and 4 genes belonged Side-S1b and Side-Lowland genes).

Function of the pickup genes

We employed the functional annotation of candidate genes detected in the F_{ST}outlier analysis with GO terms, which describe the functions of gene products. The S2H genes included, for instance, *SAV6* (AT5G26680) implicated in “Response to UV” (Zhang et al. 2016) (Table S2). The genetic variants in this gene showed that the lowland allele was observed only in the lowland subpopulation, whereas the

highland alleles existed in the whole altitudes (Fig. 7a). 3 non-synonymous variants located in this gene and there was no missense or nonsense variant. The S2L genes included AT1G07590 assigned to “Response to cadmium ion”. The genetic variants in AT1G07590 showed that the lowland allele was more frequent than the highland allele in the intermediate subpopulation, and the homozygote of the highland allele was frequent only in the highland subpopulation (Fig. 7b). 7 out of 14 non-synonymous variants in AT1G07590 were missense variant and there was no nonsense variant. A variant represented in Fig. 7b was missense variant. GO terms that were assigned to S2H and S2L genes differed from each other in most cases. Exceptionally, only one GO term “Golgi vesicle-mediated transport” was assigned to both S2H and S2L genes (Table 2).

We also employed the functional annotation of the S1b genes. For instance, the S1b genes included *BAR1* (AT5G18360) implicated in “Immunity interacted in pathogen” (Laflamme et al. 2020) and *PDIL1-4* (AT5G60640) implicated in “Response to oxidative stress” (Sweetlove et al. 2002). 45 of 49 altitude-dependent SNPs in *BAR1* showed a similar pattern to each other that the heterozygotes were absent throughout the altitudinal gradient (Fig. 8a), while other 4 SNPs in *BAR1* and 2 SNPs in *PDIL1-4* showed the pattern that some heterozygotes were found in the lowland and highland subpopulations (Figs. 8b, c). 21 HAAD SNPs in *BAR1* and 2 SNPs in *PDIL1-4* were missense variants. To test whether the sets of 77 S1b genes and 286 Side-S1b genes locating near the Coexisting-HAAD SNPs (Tables S3, S4) were accumulated in particular biological processes, we conducted GO enrichment analysis. However, we could not find any strongly enriched term in both tests (Tables 2, S5). We found that some S1b genes were expected to relate with response to metal ion, for instance, “Response to zinc ion” and “Response to cadmium ion” (Table 2).

Discussion

Genetic structure in the intermediate subpopulation

We analyzed the individual-based whole-genome resequencing data of *Arabidopsis halleri* sampled along the altitude between 359 m and 1,317 m with very high spatial resolutions. This dataset enabled us to find the zone where the highland and lowland genes were mixing (Fig. 3b) and to investigate how gene flow and selective pressure shape the genetic structure in the intermediate subpopulation. The result of ADMIXTURE (Fig. 3b) suggests that the direct crossing between lowland and highland subpopulations might not occur so frequently, probably because of the less overlap of flowering time between the highland and lowland subpopulations. Therefore, the gene flow between the highland and lowland subpopulations is expected to occur through the intermediate subpopulation, which have relatively similar phenology to both highland and lowland subpopulations.

Our results suggest that there are various types of selection in the intermediate subpopulation of *Arabidopsis halleri*. We found 24 genomic windows whose allele frequency in the intermediate subpopulation is similar to that in highland but different from that in lowland subpopulation, which are consistent with the Scenario 2H (a gene that is adaptive to highland is also adaptive to the intermediate

zone). Similarly, we also found 8 genomic windows, which are consistent with the Scenario 2L (a gene that is adaptive to lowland is also adaptive to the intermediate zone). On the other hand, we could not find genetic windows in which one of alleles had high frequency only in the intermediate subpopulation (Fig. 4, Table 1). This result might reject scenario 3 that individuals with peculiar alleles to the intermediate zone are adaptive. However, when we relaxed the thresholds of F_{ST} outliers from the higher 1% to the higher 1.67%, we could find one window matching to the Scenario 3 (Fig. S2), implying that this scenario can occur. We also investigated Altitude-Dependent SNPs whose heterozygote had low frequency in the intermediate subpopulation and found that the 2,482 SNPs had similar frequencies of H and L alleles in the intermediate subpopulation (Coexisting-HAAD SNPs), which are consistent with the Scenario 1b. These alleles are considered to be neutral in the fitness in the intermediate subpopulation but their heterozygotes are negatively selected. In contrast, we did not find any Altitude-Dependent SNPs that had higher heterozygosity than the theoretical expectation (Fig. 5b). This result rejects the Scenario 1c that heterozygous individuals are advantageous (heterosis).

We found 35 and 13 genes that were contained in 2H and 2 L windows, respectively. These genes (2H and 2L genes) are considered to be related with adaptation in the intermediate subpopulation. *SAV6*, one of the 2H genes, is implicated in the responses to UV and *Arabidopsis* mutant of this gene showed hypersensitivity to UV radiation (Zhang et al. 2016). As UV stress is known to be greater in higher altitude (Wang et al. 2014), the variants in *SAV6* might relate with adaptation to UV stress. Our observations in *SAV6* is also consistent with the fact that the light environment of *A. halleri* habitat changes with altitude; the lowland subpopulation was covered by forest canopies, whereas the highland and intermediate subpopulations were exposed to direct light (hemisphere photographs are presented in Wang et al. 2019). Wang et al. (2016) reported that the response to enhanced UV was different between highland- and lowland ecotypes of *Alabidoposis halleri*, implying that the variant in *SAV6* is involved in the ecotypic differentiation in the UV response. AT1G07590, one of the 2L genes, is considered to relate with the responses to heavy metal stress (Sarry et al. 2006) (Table S2), suggesting that the genetic variants are related to the altitudinal gradient of the soil heavy metal concentration. Although we do not have information on the soil heavy metal concentrations at the studied site, it is known that the high altitudinal area of Mt. Ibuki is characterized by calcareous soil, whereas non-calcareous area exists at lower altitudes (Honjo and Kudo 2019), which may be related to heavy metal concentrations. Our results thus suggest that the soil heavy metal concentration may be similar between the lowland and intermediate zones, whereas that in highland soil differs from others.

We found 77 S1b genes, which are consistent with the Scenario 1b (the homozygous genotypes of both highlands and lowlands are similarly adaptive in the intermediate subpopulation, but the heterozygous genotype is less adaptive and eliminated). In some SNPs such as those in *BAR1*, we did not find any heterozygote not only in the intermediate subpopulation but also in other subpopulations (Fig. 8a). This observation implies that the heterozygote of this SNP has some adverse effects on the gene function. The homozygotes of H and L alleles coexisted in the intermediate subpopulation, suggesting that they are neutral in the intermediate zone. In contrast, in some genes such as those in *PDIL1-4*, the heterozygote

was found rarely in the intermediate subpopulations but found frequently in the highland subpopulations (Fig. 8c). This phenomenon suggests that their heterozygous genotype is maladapted only in the intermediate zone. However, there is a possibility that we could not find heterozygote genotype by chance because of the small sample size. A larger sample size may be necessary to judge which of these hypotheses is true.

In this study, 210 out of 1,038 SNPs in S2H genes and 206 out of 683 SNPs in S2L genes were synonymous variants, which seem to be neutral in terms of selection (Moutinho et al. 2020). In the SNP-based method, 73 of 2,927 HAAD SNPs were synonymous variants, and the large proportion (2,002 out of 2,482) of Coexisting-HAAD SNPs located in the intergenic region. It would suggest that many candidate SNPs do not relate with adaptation but are hitchhiked with the actual genetic target of natural selection. Furthermore, our results might include some errors due to small sample size. However, we found 174 out of 2,482 Coexisting-HAAD SNPs located in coding region of 77 S1b genes, and the 9 genes contained 32 missense variants (Table S3). We also found that 190 missense variants located in 35 S2H genes and 141 missense variants located in 13 S2L genes. In genes containing missense variants, such as AT1G07590, *BAR1* and *PDIL1-4*, amino-acids substitution and change in the protein structure might occur. Mutation occurring in UTR or splice region would also regulate the gene expression and relate with the phenotypic divergence (Mayr 2017, Guan et al. 2017). Therefore, S2H, S2L and S1b genes containing non-synonymous variants might contribute to the adaptive divergence between the highland and lowland populations. The mutation in intergenic regions would also change the expression levels of genes that locate near the mutation sites (Ochiai et al. 2014). HAAD SNPs in the intergenic region might locate in the promoter regions of Side-S1b, Side-Highland or Side-Lowland genes and be the target of selection. We need to further assess whether the candidate genes actually relate with altitudinal adaptation by improving the amount of data and/or using reverse genetics.

Is selection at the intermediate altitudes potentially a driver of adaptive divergence between highland and lowland subpopulations?

A part of the results supported our hypothesis that the selective pressure at the intermediate zone could constrain the gene flow between the two populations and act as a potential driver of population adaptive divergence. In genetic variants in AT1G07590, one of the S2L genes, the homozygote of H allele was found mainly in the highland subpopulation and rare in the intermediate and lowland subpopulations, whereas the homozygote of L allele was found in both intermediate and lowland subpopulations (Figs. 7b). This pattern suggests that the instruction of H alleles to lower altitudes is prevented by the selection favoring the homozygote of L alleles at the intermediate subpopulation. Theoretical studies have suggested that the divergent selection due to environmental differences between the habitats can reduce the effect of gene flow and lead to the population adaptive divergence in particular environments during the speciation process, especially ecological speciation (reviewed in Nosil et al. 2005). Although supporting examples for this hypothesis were reported in the previous empirical studies that focused on the populations experiencing migration from neighboring populations (Nosil et al. 2005), there seems no empirical study that focused on an intermediate zone located between the two different environments.

Our results of AT1G07590 suggest that selection to a particular genomic region at the intermediate altitudes has acted as a driver of adaptive divergence between the highland and lowland subpopulations.

In genetic variant in *SAV6*, one of the S2H genes, the homozygotes of L alleles were rarely observed in the intermediate and highland subpopulation, whereas the homozygotes of H alleles existed in the whole altitudes (Figs. 7a). This pattern would be partly inconsistent with the hypothesis that the selection at the intermediate altitudes act as a driver of adaptive divergence. This gene might possess a high homozygosity at the intermediate subpopulation because of the maladaptation of the L alleles to the intermediate and highland subpopulations, whereas constraint by the selection might be weak for the gene flow on the H alleles between the altitudes.

In S1b genes, both the H and L alleles tended to exist in the whole altitudes, and the differences of allele frequency between the altitudes were relatively unclear (Figs. 8a-c). The patterns of S1b genes would suggest that the heterozygotes were negatively selected at the intermediate altitude but the selective pressure on each ecotype allele at highland and lowland altitudes was not strong, which was inconsistent with our expectation. Therefore, the selection to these genes might not prevent the gene flow between altitudes and to be a driver of adaptive divergence.

Conclusions

In this study, we investigated a fine-scale local adaptation along the altitudinal gradient using the whole-genome sequence of individuals sampled with a very high spatial resolution. We detected the genetically admixed zone between highland and lowland ecotypes in the intermediate altitudes. Using the F_{ST} approach, we detected genomic regions that suggest the existence of selective pressure in the admixed zone locating in the intermediate altitudes. Focusing on the genotype frequency of Altitude-Dependent SNPs, we also found many SNPs that have maintained a high homozygosity in the admixed zone. In some genes detected by the above two approaches, the distribution of genotypes was clearly separated above and below the intermediate zone, suggesting that the selection might prevent the admixture of highland and lowland genotypes by the gene flow over the intermediate altitudes. We suggest that selective pressures at the intermediate zone would partly contribute to the divergence between the highland and lowland populations.

Declarations

Acknowledgments

We would like to thank Masakado Kawata, Wataru Iwasaki, Takashi Makino and Kousuke Hanada for support in this research project. This study was partly supported by JST CREST Grant Number JPMJCR11B3 and KAKENHI (No. 25291095, 17H03727, 20H03317).

Funding:

This study was partly supported by JST CREST Grant Number JPMJCR11B3 and KAKENHI (No. 25291095, 17H03727, 20H03317).

Conflicts of interest / Competing interests

Not applicable

Ethics approval

Not applicable

Consent to participate

approve

Consent for publication

approve

Availability of data and material

The raw read sequences in present study are available in the DNA Data Bank of Japan Sequenced Read Archive under the accession number DRA010696.

Code availability

Not applicable

Author Contributions:

K. H., S.M. and N.Y. designed the research, S. M. performed material sampling, N. Y. performed the experiments and data analysis, T. W. and Y. I. and S. M. contributed the experiment and analysis, and N.Y. wrote the manuscript with comments from other authors.

References

Al-Shehbaz I A, O'Kane SL Jr (2002) Taxonomy and phylogeny of *Arabidopsis* (Brassicaceae). The Arabidopsis Book 1

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19: 1655–1664

Antonovics J (2006) Evolution in closely adjacent plant populations X: long-term persistence of prereproductive isolation at a mine boundary. Heredity 97: 33–37

Aryal B, Shinohara W, Honjo MN, Kudoh H (2018) Genetic differentiation in cauline-leaf-specific wettability of a rosette-forming perennial *Arabidopsis* from two contrasting montane habitats. Ann Bot 121: 1351–1360

Bisschop G, Setter D, Rafajlović M, Baird SJE, Lohse K (2020) The impact of global selection on local adaptation and reproductive isolation. Phil Trans R Soc B 375: 20190531

Briskine RV, Paape T, Shimizu-Inatsugi R, Nishiyama T, Akama S, Sese J, Shimizu KK (2017) Genome assembly and annotation of *Arabidopsis halleri*, a model for heavy metal hyperaccumulation and evolutionary ecology. Mol Ecol Resour 17: 1025–1036

Campbell-Staton SC, Bare A, Losos JB, Edwards SV, Cheviron ZA (2018) Physiological and regulatory underpinnings of geographic variation in reptilian cold tolerance across a latitudinal cline. Mol Ecol 27: 2243–2255

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK - rising to the challenge of larger and richer datasets. GigaScience 4: 7

Comeault AA, Flaxman SM, Riesch R, Curran E, Soria-Carrasco V, Gompert Z, Farkas TE, Muschick M, Parchman TL, Schwander T, Slate J, Nosil P (2015) Selection on a Genetic Polymorphism Counteracts Ecological Speciation in a Stick Insect. Curr Biol 25: 1975–1981

Cordell S, Goldstein G, Mueller-Dombois D, Webb D, Vitousek PM (1998) Physiological and morphological variation in *Metrosideros polymorpha*, a dominant Hawaiian tree species, along an altitudinal gradient: the role of phenotypic plasticity. Oecologia 113:188–196

Danecek P, Auton A, Abecasis G, Albers C A, Banks E, DePristo M A, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools, Bioinformatics 27: 2156–2158

Frichot E, Schoville SD, Bouchard G, François O (2013) Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. Mol Biol Evol 30: 1687–1699

Facon B, Jarne P, Pointier JP, David P (2005) Hybridization and invasiveness in the freshwater snail *Melanoides tuberculata*: hybrid vigour is more important than increase in genetic variance. J Evol Biol 18: 524–535

Fisher RA (1932) Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh

Galloway LF, Fenster CB (2000) Population differentiation in an annual legume: local adaptation. Evolution 54: 1173–1181

Guan H, Dong Y, Liu C et al. (2017) A splice site mutation in *shrunken1-m* causes the *shrunken 1* mutant phenotype in maize. Plant Growth Regul 83 429–439

Hendrick MF, Finseth FR, Mathiasson ME, Palmer KA, Broder EM, Breigenzer P, Fishman L (2016) The genetics of extreme microgeographic adaptation: an integrated approach identifies a major gene underlying leaf trichome divergence in Yellowstone *Mimulus guttatus*. Mol Ecol 25: 5647–5662

Hämälä T, Savolainen O (2019) Genomic Patterns of Local Adaptation under Gene Flow in *Arabidopsis lyrate*. Mol Biol Evol 36: 2557–2571

Honjo MN, Kudoh H (2019) *Arabidopsis halleri*: a perennial model system for studying population differentiation and local adaptation. AoB Plants 11, plz076

Ikeda H, Setoguchi H, Morinaga S (2010) Genomic Structure of Lowland and Highland Ecotypes of *Arabidopsis halleri* subsp. *gemmifera* (Brassicaceae) on Mt. Ibuki. Acta Phytotax Geobot 61: 21–26

Kolnik M, Marhold K (2006). Distribution, chromosome numbers and nomenclature conspect of *Arabidopsis halleri* (Brassicaceae) in the Carpathians. Biologia 61: 41–50

Körner C (2007) The use of 'altitude' in ecological research. Trends Ecol Evol 22: 569–574

Kubota S, Iwasaki T, Hanada K, Nagano AJ, Fujiyama A, Toyoda A, et al. (2015) A Genome Scan for Genes Underlying Microgeographic-Scale Local Adaptation in a Wild *Arabidopsis* Species. PLoS Genet 11: e1005361

Laflamme B, Dillon MM, Martel A, Almeida RND, Desveaux D, Guttman DS (2020) The pan-genome effector-triggered immunity landscape of a host-pathogen interaction. Science 367: 763-768

Le Moan A, Gagnaire PA, Bonhomme F. (2016) Parallel genetic divergence among coastal–marine ecotype pairs of European anchovy explained by differential introgression after secondary contact. Mol Ecol 25: 3187–3202

Lenormand T (2002) Gene flow and the limits to natural selection. Trends Ecol Evol 17: 183–189

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987–2993

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv: 1303.3997

Li Y, Stift M, van Kleunen M (2018) Admixture increases performance of an invasive plant beyond first-generation heterosis. J Ecol 106: 1595–1606

Linnen CR, Poh YP, Peterson BK, Barrett RDH, Larson JG, Jensen JD, Hoekstra HE (2013) Adaptive Evolution of Multiple Traits Through Multiple Mutations at a Single Gene. Science 339: 1312–1316

Lipshutz SE, Overcast IA, Hickerson MJ, Brumfield RT, Derryberry EP (2017) Behavioural response to song and genetic divergence in two subspecies of white-crowned sparrows (*Zonotrichia leucophrys*). Mol Ecol 26: 3011–3027

Machino S, Nagano S, Hikosaka K (2021) The latitudinal and altitudinal variations in the biochemical mechanisms of temperature dependence of photosynthesis within *Fallopia japonica*. Environ Exp Bot, in press. doi.org/10.1016/j.envexpbot.2020.104248

Mayr C (2017) Regulation by 3'-Untranslated Regions. Annu Rev Genet 51: 171-194

Moutinho AF, Bataillon T, Dutheil JY (2020) Variation of the adaptive substitution rate between species and within genomes. Evol Ecol 34: 315–338

Murai Y, Setoguchi H, Kitajima J, Iwashina T (2015) Altitudinal Variation of Flavonoid Content in the Leaves of *Fallopia japonica* and the Needles of *Larix kaempferi* on Mt. Fuji. Nat Prod Commun 10: 407-411

Nosil P, Vines TH, Funk DJ (2005) Reproductive isolation caused by natural selection against immigrants from divergent habitats. Evolution 59: 705–719

Ochiai H, Miyamoto T, Kanai A, Hosoba K, Sakuma T, Kudo Y, Asami K, Ogawa A, Watanabe A, Kajii T, Yamamoto T, Matsuura S (2014) TALEN-mediated single-base-pair editing identification of an intergenic mutation upstream of *BUB1B* as causative of PCS (MVA) syndrome. PNAS 111 (4): 1461-1466

Ohtani M, Kondo T, Tani N, Ueno S, Lee LS, Ng KKS, Muhammad N, Finkeldey R, Na'iem M, Indrioko S, Kamiya K, Harada K, Diway B, Khoo E, Kawamura K, Tsumura Y (2013) Nuclear and chloroplast DNA phylogeography reveals Pleistocene divergence and subsequent secondary contact of two genetic lineages of the tropical rainforest tree species *Shorea leprosula* (Dipterocarpaceae) in South-East Asia. Mol Ecol 22: 2264–2279

Pruisscher P, Nylin S, Gotthard K, Wheat CW (2018) Genetic variation underlying local adaptation of diapause induction along a cline in a butterfly. Mol Ecol 27: 3613–3626

Puckett EE, Park J, Combs M, Blum MJ, Bryant JE., Caccone A, Costa F, Deinum EE, Esther A, Himsworth CG, Keightley PD, Ko A, Lundkvist Å, McElhinney LM, Morand S, Robins J, Russell J, Strand TM, Suarez O, Yon L, Munshi-South J (2016) Global population divergence and admixture of the brown rat (*Rattus norvegicus*). Proc R Soc B 283: 20161762

Richardson JL, Urban MC (2013) Strong selection barriers explain microgeographic adaptation in wild salamander populations. Evolution 67: 1729–1740

Sarry JE, Kuhn L, Ducruix C, Lafaye A, Junot C, Hugouvieux V, Jourdain A, Bastien O, Fievet JB, Vailhen D, Amekraz B, Moulin C, Ezan E, Garin J, Bourguignon J (2006) The early responses of *Arabidopsis thaliana* cells to cadmium exposure explored by protein and metabolite profiling analyses. Proteomics 6: 2180-2198

Skelly DK (2004) Microgeographic countergradient variation in the wood frog, *Rana sylvatica*. Evolution 58: 160–165

Slatkin M (1987) Gene flow and the geographic structure of natural populations. Science 15: 787–792

Stacy E, Johansen J, Sakishima T et al. (2016) Genetic analysis of an ephemeral intraspecific hybrid zone in the hypervariable tree, *Metrosideros polymorpha*, on Hawai'i Island. Heredity, 117: 173–183

Sweetlove L, Heazlewood J, Herald V, Holtzapffel R, Day D, Leaver C, Millar A (2002) The impact of oxidative stress on *Arabidopsis* mitochondria. Plant J 32: 891-904

Wang Q, Hidema J, Hikosaka K (2014) Is UV-induced DNA damage greater at higher elevation?. Am J Bot 101: 796-802

Wang Q, Nagano S, Ozaki H, Morinaga S, Hidema J, Hikosaka K (2016). Functional differentiation in UV-B-induced DNA damage and growth inhibition between highland and lowland ecotypes of two *Arabidopsis* species. Environ Exp Bot 131: 110–119

Wang Q, Daumal M, Nagano S et al. (2019) Plasticity of functional traits and optimality of biomass allocation in elevational ecotypes of *Arabidopsis halleri* grown at different soil nutrient availabilities. J Plant Res 132: 237–249

Zhang Y, Wen C, Liu S, Zheng L, Shen B, Tao Y (2016) Shade avoidance 6 encodes an Arabidopsis flap endonuclease required for maintenance of genome integrity and development. Nucleic Acids res 44: 1271–1284

Tables

Table 1 The definition for searching genomic regions or single-nucleotide polymorphisms (SNPs) matching each scenario of selection and the numbers of detected genes or SNPs. The values of F_{ST} are calculated for three combinations of two subpopulations; between highland and lowland, between intermediate and lowland, and between highland and intermediate (F_{ST_HL}, F_{ST_IL} and F_{ST_HI}, respectively).

Name	Condition	Number	Scenario
S2H window	Top 1% F_{ST_HL} × Top 1% F_{ST_IL} × Bottom 70% F_{ST_HI}	24 windows	2H
S2L window	Top 1% F_{ST_HL} × Top 1% F_{ST_HI} × Bottom 70% F_{ST_IL}	8 windows	2L
S3 window	Top 1% F_{ST_HI} × Top 1% F_{ST_IL} × Bottom 70% F_{ST_HL}	0 window	3
S2H gene	Included in S2H windows × Coding region containing SNPs	35 genes	2H
S2L gene	Included in S2L windows × Coding region containing SNPs	13 genes	2L
Altitude-Dependent SNP	FDR < 0.05 in LFMM at K =1 and K =2	27,792 SNPs	1
Homozygote-accumulated SNP	Bottom 2.5% ΔH	51,300 SNPs	
Heterozygote-accumulated SNP	Top 2.5% ΔH	51,300 SNPs	
Homozygote-accumulated Altitude-Dependent (HAAD) SNP	Homozygote-accumulated SNP ×Altitude-Dependent SNP	2927 SNPs	1b
Heterozygote-accumulated Altitude-Dependent SNP	Heterozygote-accumulated SNP ×Altitude-Dependent SNP	0 SNP	1c
Highland-HAAD SNP	ΔGF > 0.5	348 SNPs	
Lowland-HAAD SNP	ΔGF < −0.5	97 SNPs	
Coexisting-HAAD SNP	$-0.5 \leq \Delta GF \leq 0.5$	2,482 SNPs	1b
S1b gene	$-0.5 \leq$ mean ΔGF ≤ 0.5	77 genes	1b
Side-S1b gene	Locating near (< 1 kb) Coexisting-HAAD SNPs	286 genes	
Highland-homozygote gene	mean ΔGF > 0.5	22 genes	2H
Side-Highland gene	Locating near (< 1 kb) Highland-HAAD SNPs	64 genes	
Lowland-homozygote gene	mean ΔGF < −0.5	3 genes	2L
Side-Lowland gene	Locating near (< 1 kb) Lowland-HAAD SNPs	14 genes	
	S2H genes × Highland-homozygote gene	3 SNPs (2 genes)	2H

	S2L genes × Lowland-homozygote gene	0 SNP	2L

Table 2 The list of gene ontology (GO) terms and the number of candidate genes associated with each term. The columns of Scenario 2H and 2L represent the results of the S2H and S2L genes, respectively. The column of Scenario 1b represents the result of the S1b genes and not consider the Side-S1b genes. In this table, only the GO terms comprising two or more candidate genes that detected by F_{ST} or SNP-based approach have been shown.

	F_{ST} approach		SNP-based approach
GO terms	Scenario 2H	Scenario 2L	Scenario 1b
Negative regulation of biological process	0	0	6
Developmental process	6	0	0
Regulation of developmental process	3	0	3
Regulation of flower development	2	0	2
Regulation of DNA-templated transcription, elongation	0	0	2
Negative regulation of gene expression	0	0	3
Formation of plant organ boundary	1	0	1
Regulation of cell population proliferation	2	0	2
Catabolic process	4	0	0
Oxoacid metabolic process	0	0	6
Carboxylic acid metabolic process	3	0	5
Cytokinin biosynthetic process	1	0	1
Peptidyl-amino acid modification	2	0	0
Dephosphorylation	2	0	0
Cellular localization	0	0	6
Organelle organization	0	3	0
Membrane docking	0	1	2
Protein transport	0	0	5
Golgi vesicle-mediated transport	1	1	0
Cellular response to stress	0	2	0
Response to zinc ion	0	0	2
Response to cadmium ion	0	0	3
Response to endoplasmic reticulum stress	0	0	2

Figures

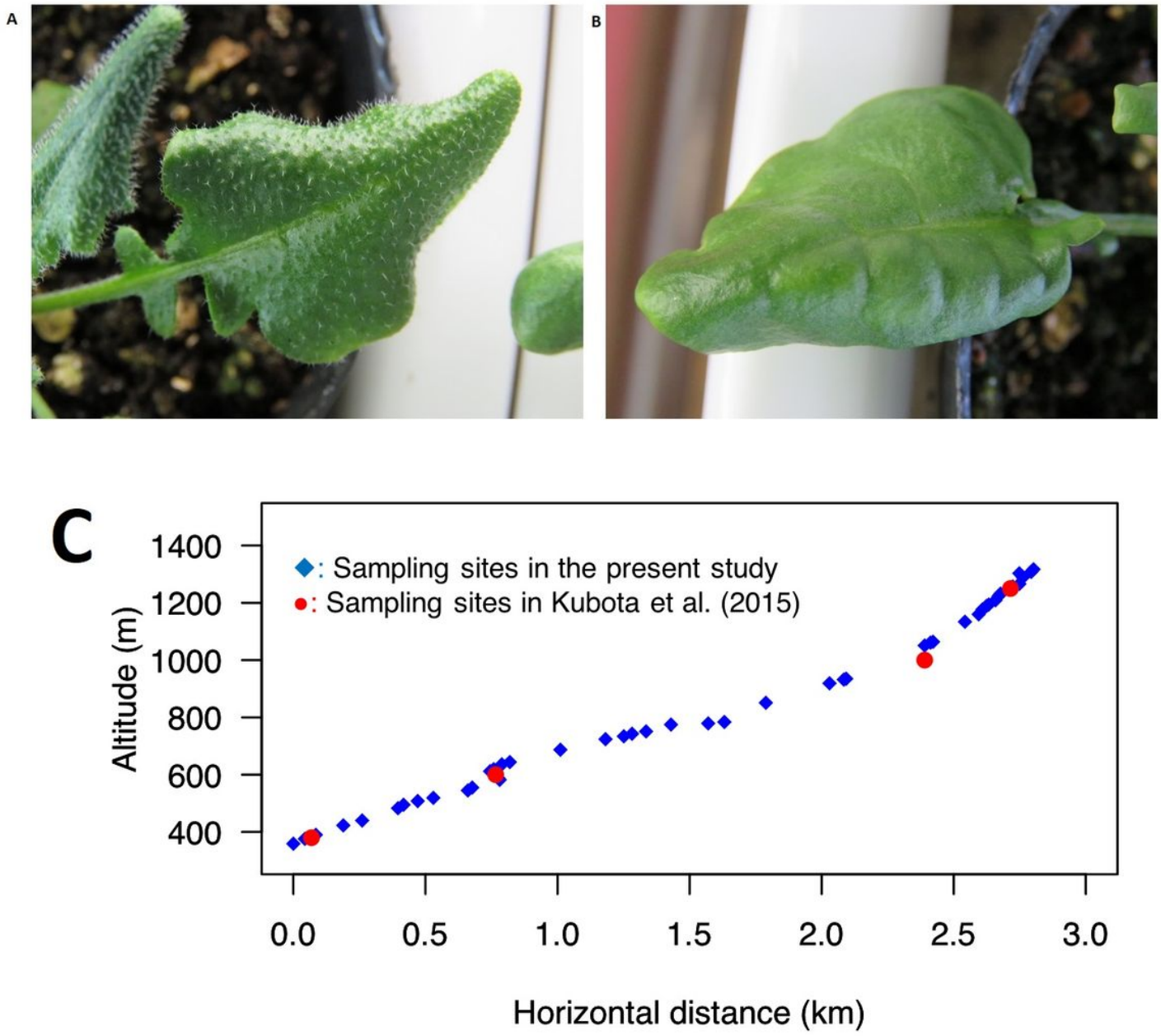

Figure 1

The leaves of (a) highland (hairy) and (b) lowland ecotypes (glabrous) of Arabidopsis halleri which were grown from the seeds in the same growth chamber. (c) The sampling point of individuals analyzed in this study. The horizontal axis represents the horizonal distance from the bottom sampling point (359 m above sea level) to each point.

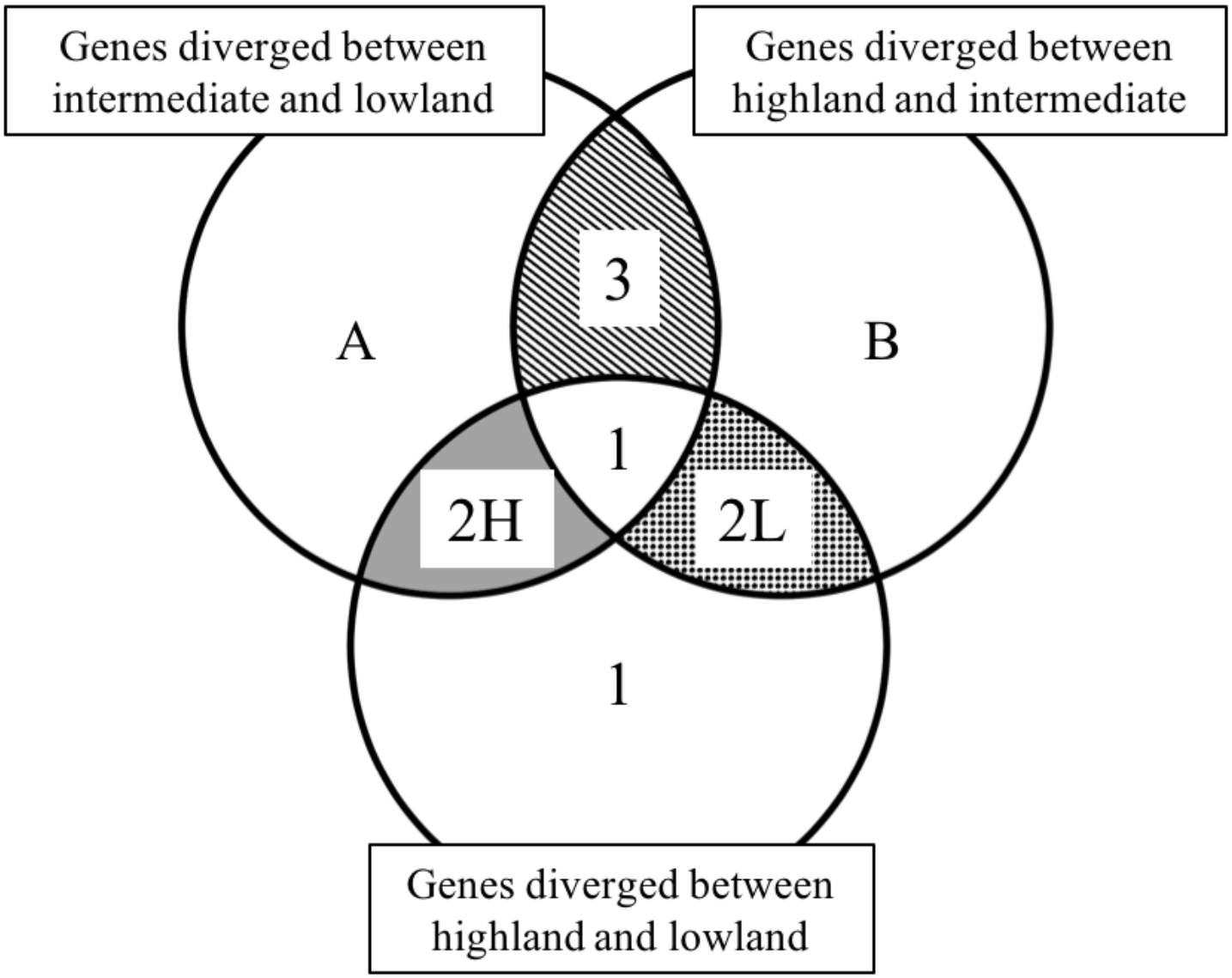

Figure 2

The outline of detection of outlier genomic windows matching each selective scenario. Each circle represents genomic windows that was significantly different between the two subpopulations. Under the Scenario 2H, genomic windows are expected to diverge between highland and lowland and between intermediate and lowland, but not between highland and intermediate subpopulations. Under the Scenario 2L, genomic windows are expected to diverge between highland and lowland and between highland and intermediate, but not diverge between intermediate and lowland subpopulations. Genomic windows that expected to match the Scenario 1 will be contained in the circle of genes that diverge between highland and lowland subpopulations except genes matching the Scenarios 2H and 2L. Under the scenario 3, genomic windows are expected to diverge between highland and intermediate and between intermediate and lowland, but not between highland and lowland subpopulations. The area A denotes the genomic windows that significantly diverge between the intermediate and lowland subpopulations but not in other combinations of subpopulations. The area B denotes the genomic

windows that diverge between the highland and intermediate subpopulations but not in other combinations of subpopulations.

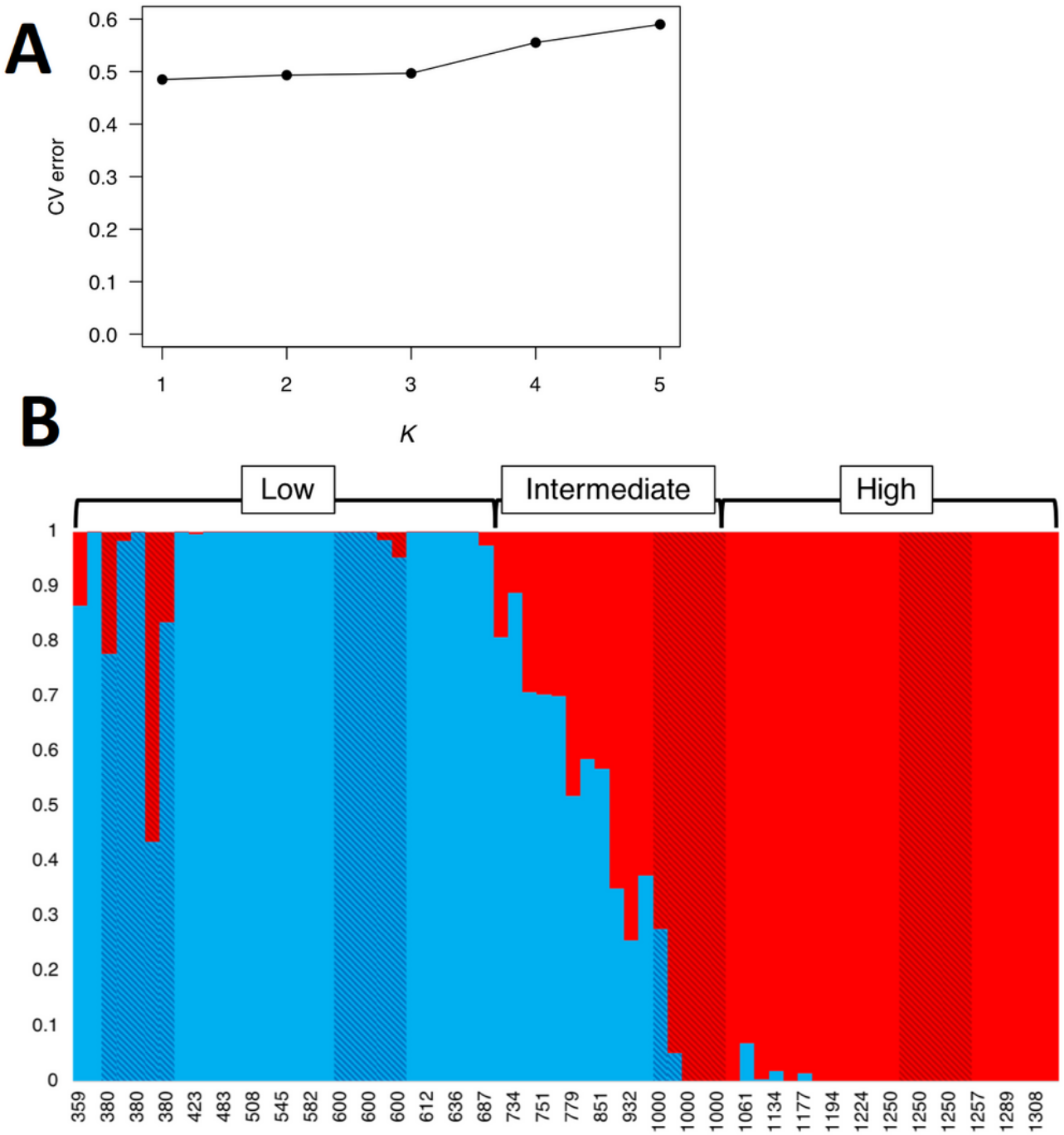

Figure 3

(a) The cross-validation (CV) error as a function of the number of subpopulations (K). (b) The population structure when K=2. Each bar represents one individual. The shaded bars represent individuals sampled previously in Kubota et al. (2015), and the other bars represent individuals sampled in this study. The

vertical axis represents the estimated membership in a particular genetic cluster, and the horizonal axis represents altitude where each individual was sampled.

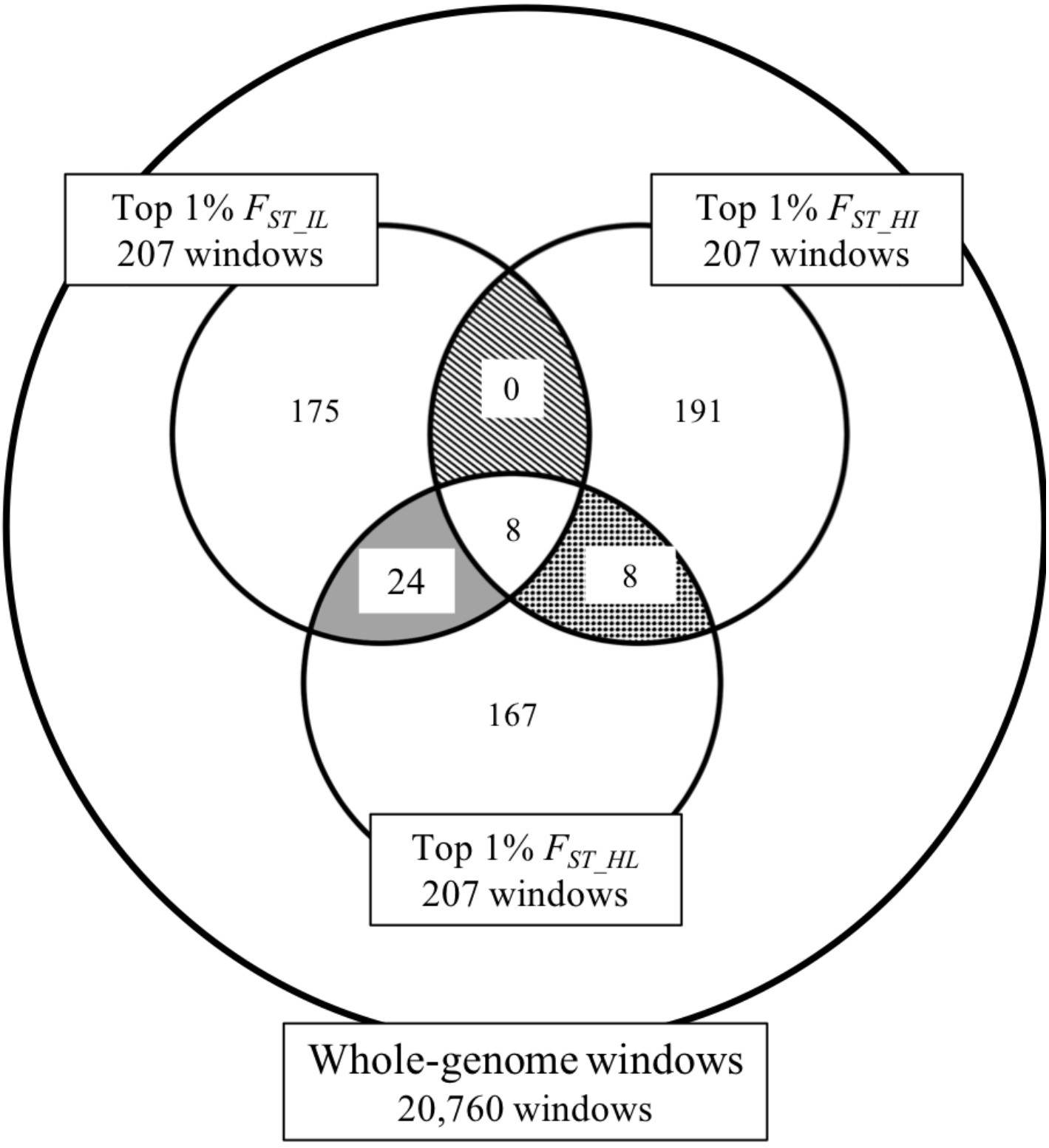

Figure 4

The numbers of genomic windows detected by the FST approach. Each small circle in a large circle represents outliers whose statistic values of FST was extremely large (the higher 1%) in each calculation.

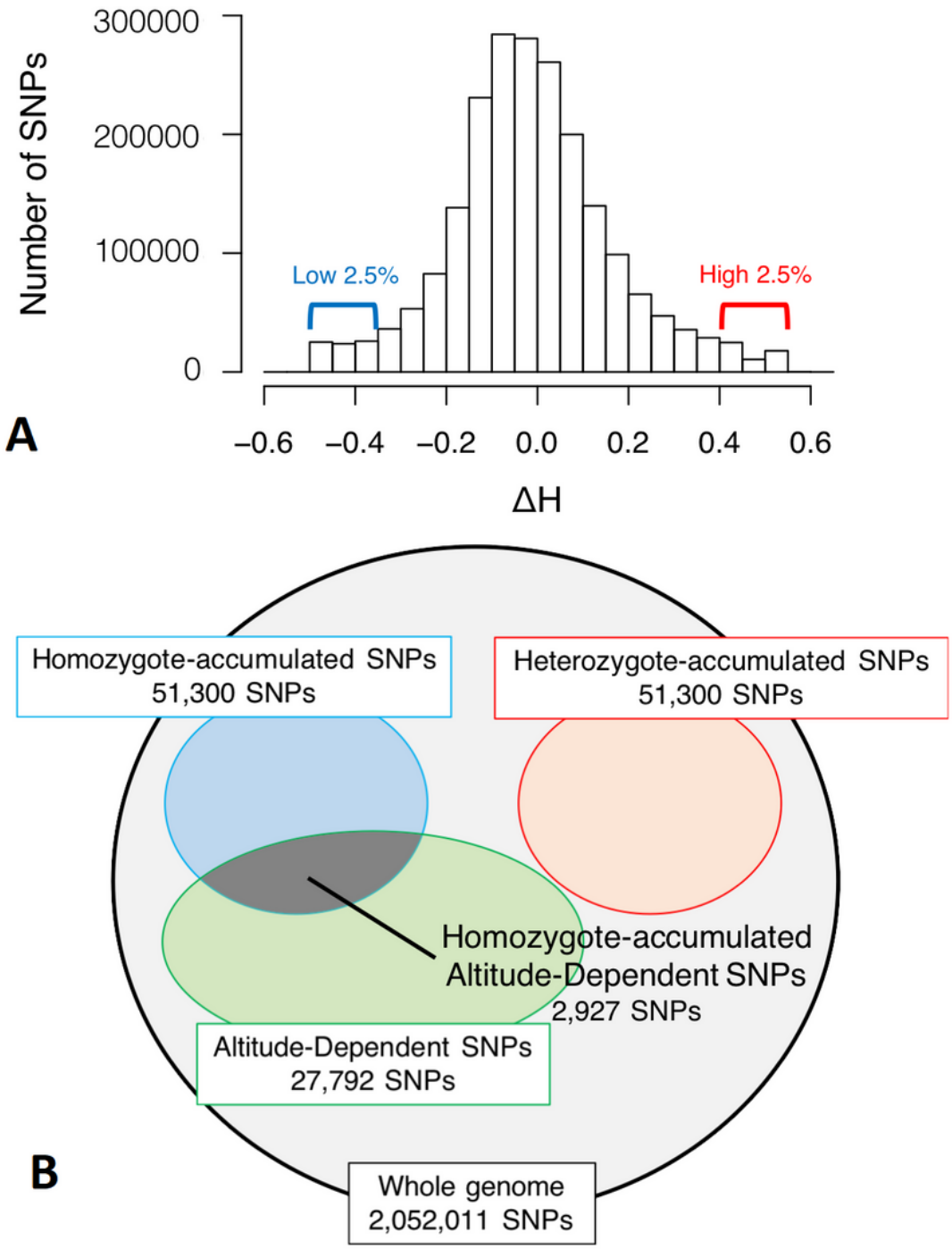

Figure 5

(a) The histogram of the differences between the observed and expected heterozygosity in the intermediate subpopulation (ΔH). The vertical axis is the number of single-nucleotide polymorphisms (SNPs). The top 2.5% (51,300 SNPs) and bottom 2.5% SNPs are defined as Heterozygote- and Homozygote-accumulated SNPs, respectively. (b) The number of the Altitude-Dependent SNPs defined by the latent factor mixed models (FDR < 0.05), and Heterozygote- and Homozygote-accumulated SNPs. The

overlap between the Altitude-Dependent SNPs and Homozygote-accumulated SNPs are defined as Homozygote-accumulated and Altitude-Dependent (HAAD) SNPs.

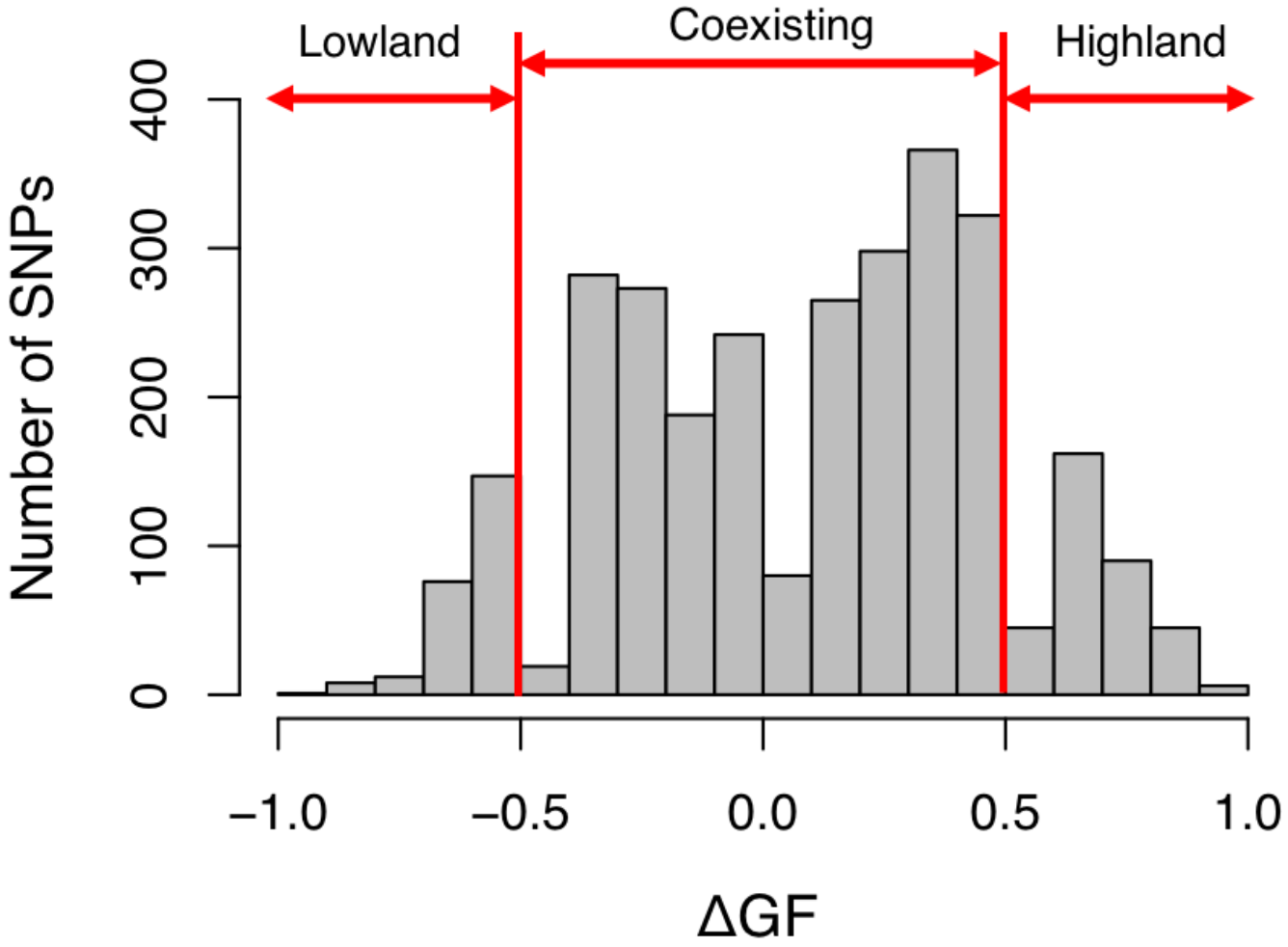

Figure 6

The histogram of the index for assessing which of the highland or lowland alleles are more frequent in the intermediate subpopulation (ΔGF) in the Homozygote-accumulated Altitude-Dependent single-nucleotide polymorphisms (HAAD SNPs). The higher values of ΔGF mean that the homozygote of highland allele is more abundant at intermediate altitudes. HAAD SNPs are classified into the Highland- ($\Delta GF > 0.5$), the Lowland- ($\Delta GF < -0.5$), and the Coexisting- HAAD SNPs ($-0.5 < \Delta GF < 0.5$).

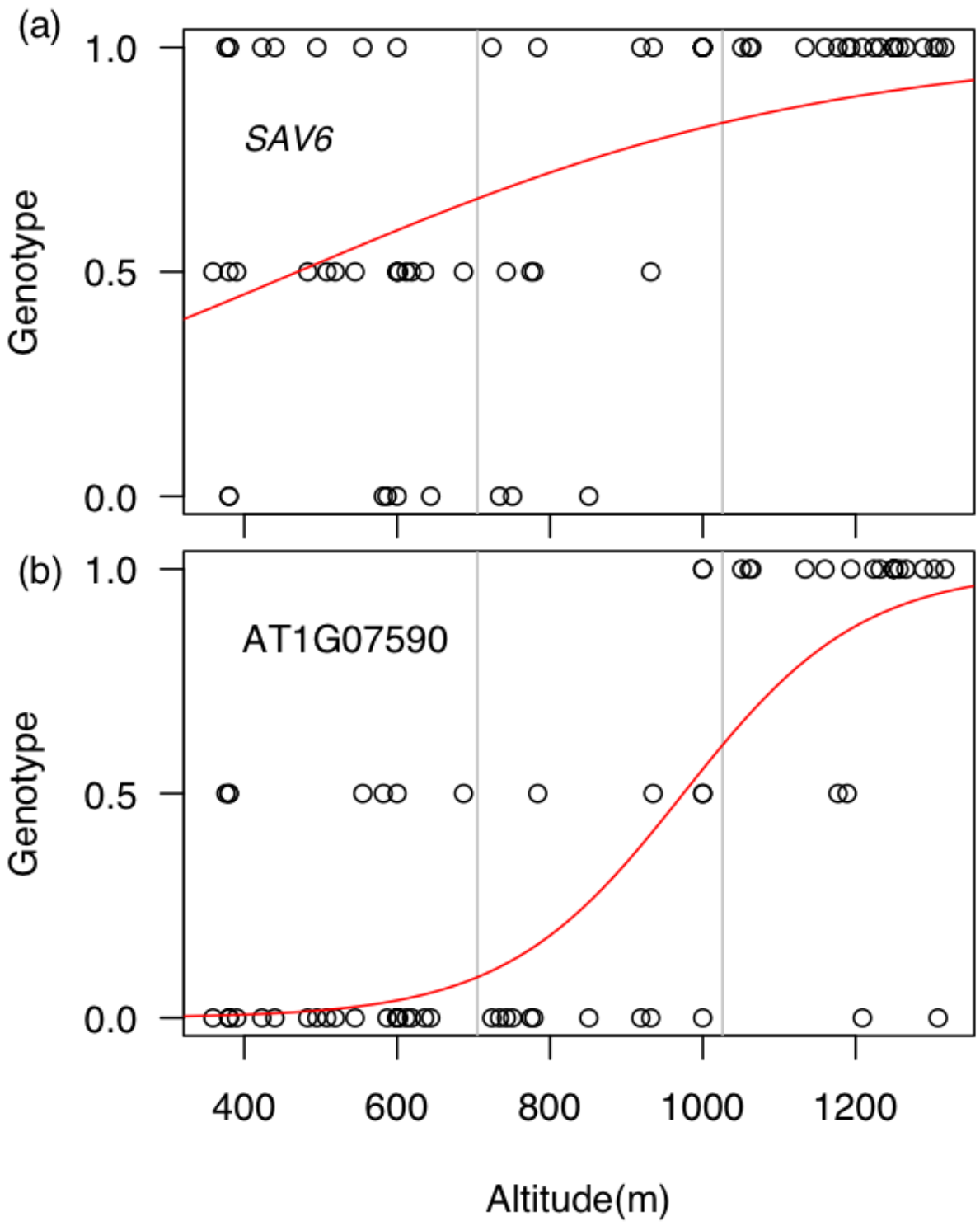

Figure 7

The genotype pattern of variants contained in S2H and S2L genes along the altitude. Each circle represents an individual. The vertical axis represents its genotype (homozygote of lowland allele = 0, heterozygote = 0.5, and homozygote of highland allele = 1). The curvilinear is logistic regression.

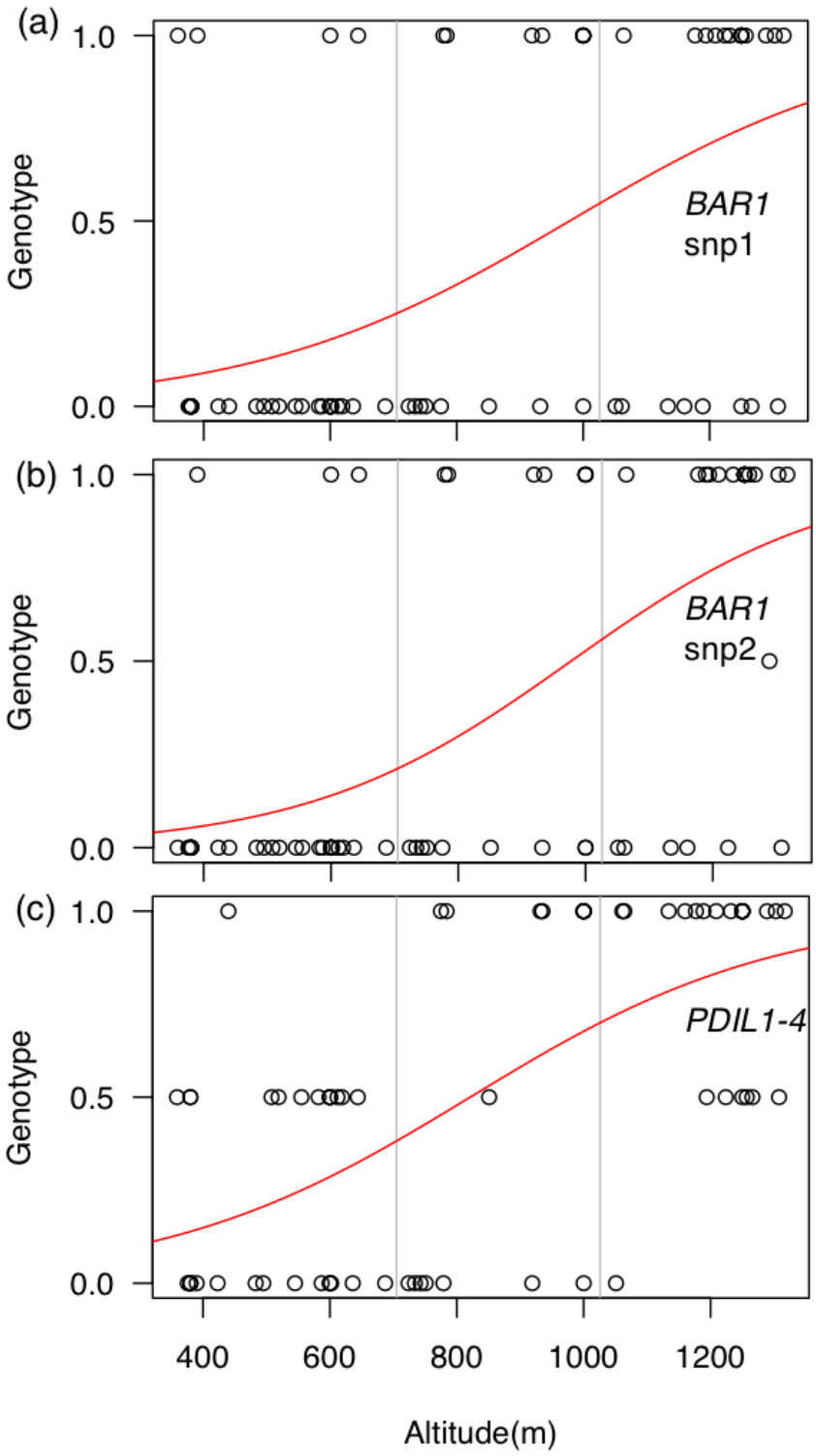

Figure 8

The genotype pattern of Homozygote-accumulated Altitude-Dependent single-nucleotide polymorphisms (HAAD SNPs) along the altitude. Each circle represents an individual. The vertical axis represents its genotype (homozygote of lowland allele = 0, heterozygote = 0.5, and homozygote of highland allele = 1). The curvilinear is logistic regression.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementalfile2.xlsx
- supplementalfile1.pdf