Natural *CMT2* variation is associated with genome-wide methylation changes and temperature seasonality

Xia Shen[1,3,4], Jennifer De Jonge[2], Simon Forsberg[1], Mats Pettersson[1], Zheya Sheng[1], Lars Hennig[2] and Örjan Carlborg[1]*

Affiliations: [1]Swedish University of Agricultural Sciences, Department of Clinical Sciences, Division of Computational Genetics, Box 7078, SE-750 07 Uppsala, Sweden. [2]Swedish University of Agricultural Sciences, Department of Plant Biology, Box 7080, SE-750 07 Uppsala, Sweden. [3]Karolinska Institutet, Department of Medical Epidemiology and Biostatistics, Box 281, SE-171 77 Stockholm, Sweden. [4]University of Edinburgh, MRC Institute of Genetics and Molecular Medicine, EH4 2XU, Edinburgh, United Kingdom.

*To whom correspondence should be addressed. Email: orjan.carlborg@slu.se

ABSTRACT

As *Arabidopsis thaliana* has colonized a wide range of habitats across the world it is an attractive model for studying the genetic mechanisms underlying environmental adaptation. Here, we used public data from two collections of *A. thaliana* accessions to associate genetic variability at individual loci with differences in climates at the sampling sites. We use a novel method to screen the genome for plastic alleles that tolerate a broader climate range than the major allele. This approach reduces confounding with population structure and increases power compared to standard genome-wide association methods. Sixteen novel loci were found, including an association between Chromomethylase 2 (*CMT2*) and temperature seasonality where the genome-wide CHH methylation was different for the group of accessions carrying the plastic allele. *Cmt2* mutants were shown to be more tolerant to heat-stress, suggesting genetic regulation of epigenetic modifications as a likely mechanism underlying natural adaptation to variable temperatures, potentially through differential allelic plasticity to temperature-stress.

AUTHOR SUMMARY

A central problem when studying adaptation to a new environment is the interplay between genetic variation and phenotypic plasticity. *Arabidopsis thaliana* has colonized a wide range of habitats across the world and it is therefore an attractive model for studying the genetic mechanisms underlying environmental adaptation. Here, we study two collections of *A. thaliana* accessions from across Eurasia to identify loci associated with differences in climates at the sampling sites. A new

genome-wide association analysis method was developed to detect adaptive loci where the alleles tolerate different climate ranges. Sixteen novel such loci were found including a strong association between Chromomethylase 2 (*CMT2*) and temperature seasonality. The reference allele dominated in areas with less seasonal variability in temperature, and the alternative allele existed in both stable and variable regions. Our results thus link natural variation in *CMT2* and epigenetic changes to temperature adaptation. We showed experimentally that plants with a defective *CMT2* gene tolerate heat-stress better than plants with a functional gene. Together this strongly suggests a role for genetic regulation of epigenetic modifications in natural adaptation to temperature and illustrates the importance of re-analyses of existing data using new analytical methods to obtain deeper insights into the underlying biology from available data.

INTRODUCTION

Arabidopsis thaliana has colonized a wide range of habitats across the world and it is therefore an attractive model for studying the genetic mechanisms underlying environmental adaptation [1]. Several large collections of *A. thaliana* accessions have either been whole-genome re-sequenced or high-density SNP genotyped [1-7]. The included accessions have adapted to a wide range of different climatic conditions and therefore loci involved in climate adaptation will display genotype by climate-at-sampling-site correlations in these populations. Genome-wide association or selective-sweep analyses can therefore potentially identify signals of natural selection involved in environmental adaptation, if those can be disentangled from the effects of other population genetic forces acting to change the allele frequencies. Selective-sweep studies are inherently sensitive to population-structure and, if present, the false-positive rates will be high as the available statistical methods are unable to handle this situation properly. Further experimental validation of inferred sweeps (e.g. [1,8]) is hence necessary to suggest them as adaptive. In GWAS, kinship correction is now a standard approach to account for population structure that properly controls the false discovery rate. Unfortunately, correcting for genomic kinship often decreases the power to detect individual adaptive loci, which is likely the reason that no genome-wide significant associations to climate conditions were found in earlier GWAS analyses [1,8]. Nevertheless, a number of candidate adaptive loci could despite this be identified using extensive experimental validation [1,2,8], showing how valuable these populations are as a resource for finding the genomic footprint of climate adaptation.

Genome-wide association (GWA) datasets based on natural collections of *A. thaliana* accessions, such as the RegMap collection, are often genetically stratified. This is primarily due to the close relationships between accessions sampled at nearby locations. Furthermore, as the climate measurements used as phenotypes for the accessions are values representative for the sampling locations of the individual accessions, these measurements will be confounded with the general genetic relationship [9]. Unless properly controlled for, this confounding might lead to excessive false-positive signals in the association analysis; this as the differences in allele-frequencies between loci in locations that differ in climate, and at the same time are geographically distant, will create an association between the genotype and the trait. However, this association could also be due to other forces than selection. In traditional GWA analyses, mixed-model based approaches are commonly used to control for population-stratification. The downside of this approach is that it, in practice, will remove many true genetic signals coming from local adaptation due to the inherent confounding between local genotype and adaptive phenotype. Instead, the primary signals from such analyses will be due to effects of alleles that exist in, and have similar effects across, the entire studied population. In general, studies into the contributions of genetic variance-heterogeneity to the phenotypic variability in complex traits is a novel and useful approach with great potential [10]. Here, we have developed and used a new approach that combines a linear mixed model and a variance-heterogeneity test, which addresses these initial concerns and shown that it is possible to infer statistically robust results of genetically regulated phenotypic variability in GWA data from natural populations.

This study describes the results from a reanalysis of data from the RegMap collection to find loci contributing to climate adaptation through an alternative mechanism: genetic control of plasticity. Such loci are unlikely to be detected with standard GWAS or selective-sweep analyses as they have a different genomic signature of selection and distribution across climate envelopes. The reason for this difference is that plastic alleles are less likely to be driven to fixation by directional selection, but rather that multiple alleles remain in the population under extended periods of time by balancing selection [11]. To facilitate the detection of such loci, we extend and utilize an approach [12,13] that instead of mapping loci by differences in allele-frequencies between local environments, which is highly confounded by population structure, infer adaptive loci using a heterogeneity-of-variance test. This identifies loci where the minor allele is associated with a broader range of climate conditions than the major allele [12]. As such widely distributed alleles will be present across the

entire population, they are less confounded with population structure and detectable in our GWAS analysis that utilizes kinship correction to account for population stratification.

RESULTS

Genome-wide association analysis to detect loci with plastic response to climate

A genome-wide association analysis was performed for thirteen climate variables across ~215,000 SNPs in 948 *A. thaliana* accessions from the RegMap collection, representing the native range of the species [1,9]. In total, sixteen genome-wide significant loci were associated with eight climate variables (Table 1), none of which could be found using standard methods for GWAS analyses [1,8,14-16]. The effects were in general quite large, from 0.3 to 0.5 residual standard deviations (Table 1), meaning that the minor allele is associated with a climate that is between 21-35% more variable than that of the major allele. The detailed results from the association analysis for each of these climate variables are reported in Text S1 (Supplementary Figure 1-13). As expected, there was low confounding between the alleles associated with a broader range of climate conditions and population structure. This is illustrated by the plots showing the distributions of these alleles across the population strata in relation to their geographic origin and the climate envelopes in Text S1 (Supplementary Figure 14-35).

Identification of candidate mutations using re-sequencing data from the 1001-genomes project

Utilizing the publicly available whole-genome re-sequencing data from the 1001-genomes project [2-7] (http://1001genomes.org), we screened the loci with significant associations to the climate variables for candidate functional polymorphisms. Missense, nonsense or frameshift mutations in high linkage disequilibrium (LD; $r^2 > 0.8$) with the leading SNPs were identified in five functional candidate genes associated with eight climate variables (for details on these see Table 1) and 11 less characterized genes (Table S1). 76 additional linked loci or genes without candidate mutations in their coding regions are provided in Table S2.

Several loci are associated with multiple climate variables

Interestingly, three out of the eight loci with missense mutations affected more than one climate variable, even though these were only marginally correlated. One such potentially pleiotropic adaptive effect for day length and relative humidity in the spring was associated with a locus containing the genes *VEL1* and *XTH19* (Table 1). The major allele at this locus was predominant in short-day regions, whereas the alternative allele was more plastic in relation to day-length. *XTH19*

has been implied as a regulator of shade avoidance [17], but information about its potential involvement in regulation of photoperiodic length is lacking. *VEL1,* is a Plant Homeo Domain (PHD) finger protein. PHD finger proteins are known to affect vernalization and flowering of *A. thaliana*, e.g. by silencing the key flowering locus *FLC* during vernalization, and is involved in photoperiod-mediated epigenetic regulation of MAF5 [18-20]. The finding that *VEL1* is associated with day length and relative humidity is thus consistent with the role of previous reports on PHD finger proteins. It also makes this protein an interesting target for future studies into the genetics underlying simultaneous adaptation to day-length and humidity.

Another potentially pleiotropic adaptive effect was identified for two more highly correlated traits, minimum temperature and number of consecutive cold days (Pearson's $r^2 = 0.76$). In total, 17 missense mutations were found at this locus. The top candidate gene containing a missense mutation is galactinol synthase 1 (*GolS1*). This gene has been reported to be involved in extreme temperature-induced synthesis [21,22], making it an interesting target for further studies regarding the genetics of temperature adaptation.

Chromomethylase 2 *(*CMT2*) is associated with temperature seasonality in the RegMap collection*

A strong association to temperature seasonality, i.e. the ratio between the standard deviation and the mean of temperature records over a year, was identified near Chromomethylase 2 (*CMT2*; Table 1; Figure 1). Stable areas are generally found near large bodies of water (e.g. London near the Atlantic 11 ± 5◦C; mean ± SD) and variable areas inland (e.g. Novosibirsk in Siberia 1 ± 14◦C). A premature *CMT2* stop codon located on chromosome 4 at 10,414,556 bp (the 31st base pair of the first exon) segregated in the RegMap collection, with minor allele frequency of 0.05. This $CMT2_{STOP}$ allele had a genome-wide significant association with temperature seasonality ($P = 1.1 \times 10^{-7}$) and was in strong LD ($r^2 = 0.82$) with the leading SNP (Figure 1B). The geographic distributions of the wild-type ($CMT2_{WT}$) and the alternative ($CMT2_{STOP}$) alleles in the RegMap collection shows that the $CMT2_{WT}$ allele dominates in all major sub-populations sampled from areas with low or intermediate temperature seasonality. The plastic $CMT2_{STOP}$ allele is present, albeit at lower frequency, across all sub-populations in low- and intermediate temperature seasonality areas, and is more common in areas with high temperature seasonality (Figure 2A; Figure 3; Text S1, Supplementary Figure 36). Such global distribution across the major population strata indicates that the allele has been around in the Eurasian population sufficiently long to spread across most of the native range and that the

allele is not deleterious but rather maintained through balancing selection [11], perhaps by mediating an improved tolerance to variable temperatures.

Broader geographic distribution of the CMT2$_{STOP}$ *allele in the 1001-genomes collection*

To confirm that the *CMT2$_{STOP}$* association was not due to sampling bias in the RegMap collection, we also scored the *CMT2* genotype and collected the geographical origins from 665 accessions that were part of the 1001-genomes project (http://1001genomes.org) [2,3,5-7]. In this more geographically diverse set (Figure 2A), *CMT2$_{STOP}$* was more common (MAF = 0.10) and had a similar allele distribution across Eurasia as in RegMap (Text S1, Supplementary Figure 36-37). Two additional mutations were identified on unique haplotypes (r^2 = 0.00) - one nonsense *CMT2$_{STOP2}$* at 10,416,213 bp (MAF = 0.02) and a frameshift mutation at 10,414,640 bp (two accessions). Both *CMT2$_{STOP}$* and *CMT2$_{STOP2}$* had genotype-phenotype maps implying a plastic response to variable temperature (Figure 2B) and the existence of multiple mutations disrupting *CMT2* further suggest lack of *CMT2* function as a potentially evolutionary beneficial event [23].

Accessions with the CMT2$_{STOP}$ *allele has an altered genome-wide CHH-methylation pattern*

CMT2 is a plant DNA methyltransferase that methylates mainly cytosines in CHH (H = any base but G) contexts, predominantly at transposable elements (TEs) [24,25]. We tested the effect of *CMT2$_{STOP}$* on genome-wide DNA methylation using 135 *CMT2$_{WT}$* and 16 *CMT2$_{STOP}$* accessions, for which high-quality MethylC-sequencing data was publicly available [7]. In earlier studies [24,25], it has been shown that *CMT2*-mediated CHH methylation primarily affects TE-body methylation. In *cmt2* knock-outs in a Col-0 genetic background, this results in a near lack of CHH methylation at such sites. Here, we compared the levels of CHH-methylation across TE's between *CMT2$_{STOP}$* and *CMT2$_{WT}$* accessions. Our analyses revealed that the accessions carrying the *CMT2$_{STOP}$* allele on average had a small (1%) average decrease in CHH-methylation across the TE-body compared to the *CMT2$_{WT}$* accessions. A more detailed analysis showed that this difference was primarily due to two of 16 *CMT2$_{STOP}$* accessions, Kz-9 and Neo-6, showing a TE-body CHH methylation pattern resembling that of the *cmt2* knockouts in the data of [24]. Interestingly, none of the 135 *CMT2$_{WT}$* accessions displayed such a decrease in TE-body CHH methylation, and hence there is a significant increase in the frequency of the *cmt2* knock-out TE-body CHH methylation pattern among the natural *CMT2$_{STOP}$* accessions (P = 0.01; Fisher's exact test). Our analyses show that the methylation-pattern is more heterogeneous among the natural accessions than within the Col-0 accession, both for the *CMT2$_{STOP}$* and *CMT2$_{WT}$* accessions (both P = 0.01 ; Brown-Forsythe

heterogeneity of variance test; Figure 4). There is thus a significant association between the *CMT2$_{STOP}$* polymorphism and decreased genome-wide TE-body CHH-methylation levels, and we show that this is apparently due to an increased frequency of the *cmt2*-mutant methylation phenotype. Further, the results also show a variable contribution of *CMT2*-independent CHH methylation pathways in the natural accessions. The reason why not all *CMT2$_{STOP}$* accessions behave like null alleles is unclear, but the variability amongst in the level of CHH-methylation across the natural accessions suggest that it is possible that *CMT2*-independent pathways, such as the RNA-dependent DNA-methylation pathway, compensate for the lack of *CMT2* due to segregating polymorphisms also at these loci. Alternatively, *CMT2$_{STOP}$* alleles may not be null, maybe due to stop codon read-through, which is more common than previously thought [26]. Although our analyses of genomewide methylation data have established that *CMT2$_{STOP}$* allele has a quantitative effect on CHH methylation, further studies are needed to fully explore the link between the *CMT2$_{STOP}$* allele, other pathways affecting genome-wide DNA-methylation and their joint contributions to the inferred association to temperature seasonality.

Cmt2 *mutant plants have an improved heat-stress tolerance*

To functionally explore whether *CMT2* is a likely contributor to the temperature-stress response, we have subjected *cmt2* mutants to two types of heat-stress. First, we tested the reaction of Col-0 and the *cmt2-5* null mutant (Supplementary Fig. 45) to severe heat-stress (24h at 37°C). This treatment was used because it can release transcriptional silencing of some TEs [27] and could thus be a good starting point to evaluate potential stress effects on *cmt2*. Under these conditions, the *cmt2* mutant had significantly higher survival-rate (1.6-fold; $P = 9.1 \times 10^{-3}$; Figure 5A) than Col-0. To evaluate whether a similar response could also be observed under less severe, non-lethal stress, we subjected the same genotypes to heat-stress of shorter duration (6h at 37°C) and measured root growth after stress as a measure of the ability of plants to recover. Also under these conditions, the *cmt2* mutant was found to be more tolerant to heat-stress, as its growth was less affected after being stressed (Figure 5B; 1.9-fold higher in *cmt2*; P =0.026, one-sided t-test). This striking improvement in tolerance to heat-stress of *cmt2* plants suggests *CMT2*-dependent CHH methylation as an important alleviator of stress responses in *A. thaliana* and a candidate mechanism for temperature adaptation.

The CMT2$_{STOP}$ *allele is associated with increased leaf serration and higher disease presence after bacterial inoculation*

To also explore the potential effects of the $CMT2_{STOP}$ allele on other phenotypes measured in collections of natural accessions, we tested for associations between this *CMT2* polymorphism and the 107 phenotypes measured as part of a previous study [28]. Three phenotypes were found to be significantly associated with the genotype at this locus (Supplementary Figure 39).

Associations were found to two phenotypes related to disease presence following inoculation with *Pseudomonas viridiflava* (strains PNA3.3a and ME3.1b; $P = 4.8 \times 10^{-3}$ and $P = 1.3 \times 10^{-4}$, respectively). Scoring of disease was done by eye four days after inoculation in 6 replicates / strain × accession using a scale from 0 (no visible symptom) to 10 (leaves collapse and turn yellow) with an increment of 1[28]. The connection between an increased susceptibility (0.6 and 0.7 units for PNA3.3a and ME3.1b, respectively) to disease and an increased tolerance to temperature seasonality is not obvious. However, recent work by [29] has shown that widespread dynamic CHH-methylation is important for the response to *Pseudomonas syringae* infection. In light of this finding, it is therefore not unlikely that these phenotypes are functionally related via an altered *CMT2*-mediated CHH-methylation in response to abiotic and biotic stress.

An association was also found for the level of leaf serration (increase by 0.23 units for the $CMT2_{STOP}$ allele; $P = 3.3 \times 10^{-3}$), determined after growth for 8 weeks at 10°C (level from 0: entire lamina, to 1.5: sharp/ jagged serration), across 4 plants / accession[28]. Measures of leaf serration were also available at 16 and 22°C, and interestingly there was a significant *CMT2* genotype × temperature interaction ($P = 0.048$). The $CMT2_{STOP}$ accessions have the same level of serration across the three measured temperatures, whereas the level of serration decreases with temperature for the $CMT2_{WT}$ accessions (Text S1, Supplementary Fig. 38). Although we are not aware of any earlier results connecting leaf serration to the *CMT2* locus or the level of CHH-methylation in the plant, this result further indicate that the effects of the $CMT2_{STOP}$ and the $CMT2_{WT}$ alleles depends on temperature.

DISCUSSION

A major challenge in attempts to identify individual loci involved in climate adaptation is the strong confounding between geographic location, climate and population structure in the natural *A. thaliana* population. Earlier genome-wide association analyses in large collections of natural accessions experienced a lack of statistical power when correcting for population-structure [1,8]. We used an alternative GWAS approach [12] to test for a variance-heterogeneity, instead of a mean

difference, between genotypes. This analysis identifies loci where the minor allele is more plastic (i.e. exist across a broader climatic range) than the major allele. As it has low power to detect cases where the minor allele is associated with a lower variance (here with local environments), it will not map private alleles in local environments in a genome-wide analysis[12,30]. In contrast, a standard GWAS map loci where the allele-frequencies follow the climatic cline. Although plastic alleles might be less frequent in the genome, they are easier to detect in this data due to their lower confounding with population-structure. This overall increase in power is also apparent when comparing the signals that reach a lower, sub-GWAS significance level (Text S1, Supplementary Figure 40-44).

Several novel genome-wide significant associations were found to the tested climate variables, and a locus containing *VEL1* was associated to both day length and relative humidity in the spring. *A thaliana* is a facultative photoperiodic flowering plant and hence non-inductive photoperiods will delay, but not abolish, flowering. A genetic control of this phenotypic plasticity is thus potentially an adaptive mechanism. *VEL1* regulates the epigenetic silencing of genes in the *FLC*-pathway in response to vernalization [19] and photoperiod length [20] resulting in an acceleration of flowering under non-inductive photoperiods. Our results suggest that genetically plastic regulation of flowering, via the high-variance *VEL1* allele, might be beneficial under short-day conditions where both accelerated and delayed flowering is allowed. In long-daytime areas, accelerated flowering is potentially detrimental hence the wild-type allele has the highest adaptive value. It can be speculated whether this is connected to the fact that day-length follows a latitudinal cline, where early flowering might be detrimental in northern areas where accelerated flowering, when the day-length is short, could lead to excessive exposure to cold temperatures in the early spring and hence a lower fitness.

A particularly interesting finding in our vGWAS was the strong association between the *CMT2*-locus and temperature seasonality. Here the allele associated with higher temperature seasonality (i.e the plastic allele) had an altered genome-wide CHH methylation pattern where some accessions displayed a TE-body CHH methylation pattern similar to that of *cmt2* mutant plants. Interestingly, we also found that *cmt2* mutants were more tolerant to both mild and severe heat-stress, a finding that both strongly implicates *CMT2* as an adaptive locus and clearly illustrates the potential of our method as a useful approach to identify novel associations of functional importance.

It is not clear via which mechanism *CMT2*-dependent CHH methylation might affect plant heat tolerance. Although our results show that the $CMT2_{STOP}$ allele is present across regions with both low and high temperature seasonality, it remains to be shown whether this is due to this allele being generally more adaptable across all environments, or whether the $CMT2_{WT}$ allele is beneficial in environments with stable temperature and the $CMT2_{STOP}$ in high temperature seasonality areas. Regardless, we consider it most likely that the effect will be mediated by TEs in the immediate neighborhood of protein-coding genes. Heterochromatic states at TEs can affect activity of nearby genes and thus potentially plant fitness [31]. Consistent with a repressive role of *CMT2* on heat stress responses, *CMT2* expression is reduced by several abiotic stresses including heat [32]. Because global depletion of methylation has been shown to enhance resistance to biotic stress [29], it is possible that DNA-methylation has a broader function in shaping stress responses than currently thought.

Our results show that $CMT2_{STOP}$ accessions have more heterogeneous CHH methylation patterns than $CMT2_{WT}$ accessions. The $CMT2_{STOP}$ polymorphism is predicted to lead to a non-functional *CMT2* protein, and hence a genome-wide CHH-methylation profile resembling that of a complete *cmt2* mutant [24]. Although some of the accessions carrying the $CMT2_{STOP}$ allele displayed this pattern with a lower CHH-methylation inside TE-bodies, most of these accessions did not have any major loss of genome-wide CHH methylation. Such heterogeneity might indicate the presence of compensatory mechanisms and hence that the effects of altered *CMT2* function could be dependent on the genetic-background. This is an interesting finding that deserves further investigation, although such work is beyond the scope of the current study. Our interpretation of the available results is that our findings reflect the genetic heterogeneity among the natural accessions studied. In light of the recent report by [25], who showed a role also of *CMT3* in TE-body CHH methylation, it is not unlikely that the regulation of CHH methylation may result from the action and interaction of several genes.

We identified several alleles associated with a broader range of climates across the native range of *A. thaliana*, suggesting that a genetically mediated plastic response might of important for climate adaptation. Using publicly available data from several earlier studies, we were able to show that an allele at the *CMT2* locus displays an altered genome-wide CHH-methylation pattern was strongly associated with temperature seasonality. Using additional experiments, we also found that *cmt2* mutant plants tolerated heat-stress better than wild-type plants. Together, these findings suggest this

genetically determined epigenetic variability as a likely mechanism contributing to a plastic response to the environment that has been of adaptive advantage in natural environments.

MATERIALS AND METHODS

Climate data and genotyped Arabidopsis thaliana *accessions*

Climate phenotypes and genotype data for a subset of the *A. thaliana* RegMap collection were previously analyzed by [1]. We downloaded data on 13 climate variables and genotypes of 214,553 single nucleotide polymorphisms (SNPs) for 948 accessions from: http://bergelson.uchicago.edu/regmap-data/climate-genome-scan. The climate variables used in the analyses were: aridity, number of consecutive cold days (below 4 degrees Celsius), number of consecutive frost-free days, day-length in the spring, growing-season length, maximum temperature in the warmest month, minimum temperature in the coldest month, temperature-seasonality, photosynthetically active radiation, precipitation in the wettest month, precipitation in the driest month, precipitation-seasonality, and relative humidity in the spring. More information on these variables is provided by [1]. No squared pairwise Pearson's correlation coefficients between the phenotypes were greater than 0.8 (Figure S7 of [1]).

We calculated the temperature seasonality for at sampling locations of a selection of 1001-genomes (http://1001genomes.org) accessions. Raw climate data was downloaded from http://www.worldclim.org/, re-formatted and thereafter processed by the raster package in R. The R code

for generating this data is provided in Text S1. The genotype for the *CMT2*$_{STOP}$ polymorphism was obtained by extracting the corresponding SNP data for the 1001-genomes accessions.

Statistical modeling in genome-wide scans for adaptability

The climate data at the geographical origins of the *A. thaliana* accessions were treated as phenotypic responses. Each climate phenotype vector y for all the accessions were normalized via an inverse-Gaussian transformation. The squared normalized measurement $z_i = y_i^2$ of accession i is modeled by the following linear mixed model to test for an association with climate adaptability (i.e. a greater plasticity to the range of the environmental condition):

$$z_i = \mu + \beta x_i + g_i + e_i$$

where μ is an intercept, x_i the SNP genotype for accession i, β the genetic SNP effect, $\mathbf{g} \sim MVN(\mathbf{0}, \mathbf{G}^* \sigma_g^2)$ the polygenic effects and $\mathbf{e} \sim MVN(\mathbf{0}, \mathbf{I}\sigma_e^2)$ the residuals. x_i is coded 0 and 2 for the two homozygotes (inbred lines). The genomic kinship matrix $\mathbf{G}^*$ is constructed via the whole-genome generalized ridge regression method HEM (heteroscedastic effects model) [13] as $\mathbf{G}^* = \mathbf{ZWZ}'$,where $\mathbf{Z}$ is a number of individuals by number of SNPs matrix of genotypes standardized by the allele frequencies. $\mathbf{W}$ is a diagonal matrix with element $w_{jj} = \hat{b}_j/(1 - h_{jj})$ for the j-th SNP, where $\hat{b}_j$ is the SNP-BLUP (SNP Best Linear Unbiased Prediction) effect estimate for the j-th SNP from a whole-genome ridge regression, and h_{jj} is the hat-value for the j-th SNP. Quantities in $\mathbf{W}$ can be directly calculated using the $\mathbf{bigRR}$ package [13] in R. An example R source code for performing the analysis is provided in the Text S1.

The advantage of using the HEM genomic kinship matrix $\mathbf{G}^*$, rather than an ordinary genomic kinship matrix $\mathbf{G} = \mathbf{ZZ}'$, is that HEM is a significant improvement of the ridge regression (SNP-BLUP) in terms of the estimation of genetic effects [13,33]. Due to this, the updated genomic kinship matrix $\mathbf{G}^*$ better represents the relatedness between accessions and also accounts for the genetic effects of the SNPs on the phenotype.

Testing and quality control for association with climate adaptability

The test statistic for the SNP effect β is constructed as the score statistic [34]:

$$T^2 = \frac{(\tilde{\mathbf{x}}'\mathbf{G}^{*-1}\tilde{\mathbf{z}})^2}{\tilde{\mathbf{x}}'\mathbf{G}^{*-1}\tilde{\mathbf{x}}}$$

implemented in the GenABEL package [35], where $\tilde{\mathbf{x}} = \mathbf{x} - E[\mathbf{x}]$ are the centered genotypic values and $\tilde{\mathbf{z}} = \mathbf{z} - E[\mathbf{z}]$ the centered phenotypic measurements. The T^2 statistic has an asymptotic χ^2 distribution with 1 degree of freedom. Subsequent genomic control (GC) [36] of the genome-wide association results was performed under the null hypothesis that no SNP has an effect on the climate phenotype. SNPs with minor allele frequency (MAF) less than 0.05 were excluded from the analysis. A 5% Bonferroni-corrected significance threshold was applied. As suggested by [30], the significant SNPs were also analyzed using a Gamma generalized linear model to exclude positive findings that might be due to low allele frequencies of the high-variance SNP.

Statistical testing for associations between the $CMT2_{STOP}$ *polymorphism and phenotypes measured in a collection of natural accessions*

The $CMT2_{STOP}$ genotype was extracted from the publicly available genome-wide genotype data with 107 phenotype measured from [28]. The association between the $CMT2_{STOP}$ genotype and each phenotype was tested by fitting a normal linear mixed model to account for population stratification, where the genomic kinship matrix was calculated by the ibs(, weight = 'freq') procedure in the GenABEL package [35], and the linear mixed model was fitted using the hglm package [37].

Functional analysis of polymorphisms in loci with significant genome-wide associations to climate

All the loci that showed genome-wide significance in the association study was further characterized using the genome sequences of 728 accessions sequenced as part of the 1001-genomes project (http://1001genomes.org). Mutations within a ±100Kb interval of each leading SNP and that are in LD with the leading SNP ($r^2 > 0.8$) were reported (Table S1). The consequences of the identified polymorphisms were predicted using the Ensembl variant effect predictor [38] and their putative effects on the resulting protein estimated using the PASE (Prediction of Amino acid Substitution Effects) tool [39].

Evaluation of TE-body methylation of $CMT2_{STOP}$ *and* $CMT2_{WT}$ *natural accessions*

In a previous study, the methylation levels were scored at 43,182,344 sites across the genome using MethylC-sequencing in 152 natural *A. thaliana* accessions (data available at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE43857) [7]. 135 of these accessions carried the $CMT2_{WT}$ and 17 the $CMT2_{STOP}$ alleles. Upon further inspection, the accession Rd-0 was excluded as it did not have sufficient sequence coverage to be used in the analyses. For each accession, across all TEs, moving averages of the CHH methylation level were calculated using a 100 bp sliding window from the borders of the TEs. The same analysis was also performed for four wild-type and four cmt2 knock-out accessions (data available at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41302) [24]. The results showing the TE-body CHH methylation patterns are visualized in Figure 4.

Heat-stress treatments on cmt2 knockouts and natural $CMT2_{STOP}$ *accessions*

A *CMT2* T-DNA insertion line (SAIL_906_G03, *cmt2-5* [24,40]) was ordered from NASC. Seeds of Col-0 wild-type and *cmt2-5* was then used for heat stress experiments based on a previously

described protocol [27]. This treatment was used because it was shown to interfere with epigenetic gene silencing as evident from transcription of some TE [27]. Seeds were plated on ½ MS medium (0.8% agar, 1% sucrose), stratified for two days at 4°C in the dark and transferred to a growth chamber with 16 h light (110 µmol m^{-2} s^{-1}, 22°C) and 8 h dark (20°C) periods. Ten-day-old seedlings were transferred to 4°C for one hour and subsequently placed for 6 h or 24 h at 37.5°C in the dark. Plant survival was scored two days after 24 h of heat stress with complete bleaching of shoot apices as lethality criterion (Text S1, Supplementary Fig. 46, Table S3). Experiments were repeated six times, each with ~30 plants per genotype. Root length was measured immediately before the 6 h heat stress and two days after heat stress (Table S4).

A log-linear regression was conducted to test for the difference in survival rate between Col-0 and *cmt2-5* knockout, i.e.

$$\log\left(E\left[\frac{s_i}{t_i}\right]\right) = \beta_0 + E_i + a_i$$

where s_i is the number of surviving plants of accession i, t_i the corresponding total number of plants, E_i the experiment effect, a_i the accession effect, and β_0 an intercept. The model fitting procedure was implemented using the `glm()` procedure in R, with option `family = gaussian(link = log)`, s_i as response, t_i as offset, and β_0, E_i, a_i as fixed effects.

ACKNOWLEDGEMENTS

We thank Leif Andersson, Jennifer Lachowiec and Yanjun Zan for helpful input. Also, providers of pre-publication sequence data within the 1001-genomes project are acknowledged for their efforts in creating this community resource, including Monsanto Company, the Weigel laboratory at the Max Planck Institute for Developmental Biology, the IGS of the Center for Biotechnology of the University of Bielefeld, the DOE Joint Genome Institute (JGI), the Joint BioEnergy Institute, the Nordborg laboratory of the Gregor Mendel Institute of Molecular Plant Biology, the Bergelson lab of the University of Chicago and the Ecker lab of the Salk Institute for Biological Studies, La Jolla, CA.

REFERENCES

1. Hancock AM, Brachi B, Faure N, Horton MW, Jarymowycz LB, et al. (2011) Adaptation to Climate Across the Arabidopsis thaliana Genome. Science 334: 83–86.

2. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, et al. (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet 43: 956–963. doi: 10.1038/ng.911.

3. Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, et al. (2008) Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome Research 18: 2024–2033. doi:10.1101/gr.080200.108.

4. Schneeberger K, Hagmann J, Ossowski S, Warthmann N, Gesing S, et al. (2009) Simultaneous alignment of short reads against multiple genomes. Genome Biol 10: R98. doi:10.1186/gb-2009-10-9-r98.

5. Schneeberger K, Ossowski S, Ott F, Klein JD, Wang X, et al. (2011) Reference-guided assembly of four diverse Arabidopsis thaliana genomes. Proceedings of the National Academy of Sciences 108: 10249–10254. doi:10.1073/pnas.1107739108.

6. Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, et al. (2013) Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. Nat Genet 45: 884–890. doi:10.1038/ng.2678.

7. Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, et al. (2013) Patterns of population epigenomic diversity. Nature 495: 193–198. doi:10.1038/nature11968.

8. Fournier-Level A, Korte A, Cooper MD, Nordborg M, Schmitt J, et al. (2011) A Map of Local Adaptation in Arabidopsis thaliana. Science 334: 86–89.

9. Horton MW, Hancock AM, Huang YS, Toomajian C, Atwell S, et al. (2012) Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. Nat Genet 44: 212–216. doi:10.1038/ng.1042.

10. Geiler-Samerotte K, Bauer C, Li S, Ziv N, Gresham D, et al. (2013) The details in the distributions: why and how to study phenotypic variability. Current Opinion in Biotechnology: 1–8. doi:10.1016/j.copbio.2013.03.010.

11. Pettersson ME, Nelson RM, Carlborg Ö (2012) Selection on variance-controlling genes: adaptability or stability. Evolution 66: 3945–3949. doi:10.1111/j.1558-5646.2012.01753.x.

12. Shen X, Pettersson M, Rönnegård L, Carlborg Ö (2012) Inheritance beyond plain heritability: variance-controlling genes in Arabidopsis thaliana. PLoS Genet 8: e1002839. doi:10.1371/journal.pgen.1002839.

13. Shen X, Alam M, Fikse F, Rönnegård L (2013) A novel generalized ridge regression method for quantitative genetics. Genetics 193: 1255–1268. doi:10.1534/genetics.112.146720.

14. Baxter I, Brazelton JN, Yu D, Huang YS, Lahner B, et al. (2010) A coastal cline in sodium accumulation in Arabidopsis thaliana is driven by natural variation of the sodium transporter AtHKT1;1. PLoS Genet 6: e1001193. doi:10.1371/journal.pgen.1001193.

15. Trontin C, Tisné S, Bach L, Loudet O (2011) What does Arabidopsis natural variation teach us (and does not teach us) about adaptation in plants? Curr Opin Plant Biol 14: 225–231. doi:10.1016/j.pbi.2011.03.024.

16. Weigel D (2012) Natural variation in Arabidopsis: from molecular genetics to ecological genomics. Plant Physiology 158: 2–22. doi:10.1104/pp.111.189845.

17. Sasidharan R, Chinnappa CC, Staal M, Elzenga JTM, Yokoyama R, et al. (2010) Light quality-mediated petiole elongation in Arabidopsis during shade avoidance involves cell wall modification by xyloglucan endotransglucosylase/hydrolases. Plant Physiology 154: 978–990. doi:10.1104/pp.110.162057.

18. Sung S, Schmitz RJ, Amasino RM (2006) A PHD finger protein involved in both the vernalization and photoperiod pathways in Arabidopsis. Genes & Development 20: 3244–3248. doi:10.1101/gad.1493306.

19. De Lucia F, Crevillen P, Jones AME, Greb T, Dean C (2008) A PHD-polycomb repressive complex 2 triggers the epigenetic silencing of FLC during vernalization. Proceedings of the National Academy of Sciences 105: 16831–16836. doi:10.1073/pnas.0808687105.

20. Kim D-H, Sung S (2010) The Plant Homeo Domain finger protein, VIN3-LIKE 2, is necessary for photoperiod-mediated epigenetic regulation of the floral repressor, MAF5. Proceedings of the National Academy of Sciences 107: 17029–17034. doi:10.1073/pnas.1010834107.

21. Panikulangara TJ, Eggers-Schumacher G, Wunderlich M, Stransky H, Schöffl F (2004) Galactinol synthase1. A novel heat shock factor target gene responsible for heat-induced synthesis of raffinose family oligosaccharides in Arabidopsis. Plant Physiology 136: 3148–3158. doi:10.1104/pp.104.042606.

22. Taji T, Ohsumi C, Iuchi S, Seki M, Kasuga M, et al. (2002) Important roles of drought- and cold-inducible genes for galactinol synthase in stress tolerance in Arabidopsis thaliana. Plant J 29: 417–426.

23. Barrick JE, Lenski RE (2013) Genome dynamics during experimental evolution. Nat Rev Genet 14: 827–839. doi:10.1038/nrg3564.

24. Zemach A, Kim MY, Hsieh P-H, Coleman-Derr D, Eshed-Williams L, et al. (2013) The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. Cell 153: 193–205. doi:10.1016/j.cell.2013.02.033.

25. Stroud H, Do T, Du J, Zhong X, Feng S, et al. (2013) Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. Nat Struct Mol Biol 21: 64–72. doi:10.1038/nsmb.2735.

26. Joshua G Dunn CKFNGBERGJSW (2014) Correction: Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. eLife 3. doi:10.7554/eLife.03178.

27. Ito H, Gaubert H, Bucher E, Mirouze M, Vaillant I, et al. (2011) An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. Nature 472: 115–119. doi:10.1038/nature09861.

28. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, et al. (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465: 627–631. doi:10.1038/nature08800.

29. Dowen RH, Pelizzola M, Schmitz RJ, Lister R, Dowen JM, et al. (2012) Widespread dynamic DNA methylation in response to biotic stress. Proceedings of the National Academy of Sciences 109: E2183–E2191. doi:10.1073/pnas.1209329109.

30. Shen X, Carlborg Ö (2013) Beware of risk for increased false positive rates in genome-wide association studies for phenotypic variability. Front Genet 4: 93. doi:10.3389/fgene.2013.00093.

31. Köhler C, Wolff P, Spillane C (2012) Epigenetic mechanisms underlying genomic imprinting in plants. Annu Rev Plant Biol 63: 331–352. doi:10.1146/annurev-arplant-042811-105514.

32. Kilian J, Whitehead D, Horak J, Wanke D, Weinl S, et al. (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. The Plant Journal 50: 347–363. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1365-313X.2007.03052.x/full.

33. Shen X, Li Y, Rönnegård L, Uden P, Carlborg Ö (2014) Application of a genomic model for high-dimensional chemometric analysis. Journal of Chemometrics: n/a–n/a.

34. Chen W-M, Abecasis GR (2007) Family-based association tests for genomewide association scans. Am J Hum Genet 81: 913–926. doi:10.1086/521580.

35. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007) GenABEL: an R package for genome-wide association analysis. Bioinformatics 23: 1294–1296.

36. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997–1004.

37. Rönnegård L, Shen X, Alam M (2010) hglm: A package for fitting hierarchical generalized linear models. The R Journal 2: 20–28.

38. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, et al. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26: 2069–2070. doi:10.1093/bioinformatics/btq330.

39. Li X, Kierczak M, Shen X, Ahsan M, Carlborg Ö, et al. (2013) PASE: a novel method for functional prediction of amino acid substitutions based on physicochemical properties. Front Genet 4: 21. doi:10.3389/fgene.2013.00021.

40. Alonso JM (2003) Genome-Wide Insertional Mutagenesis of Arabidopsis thaliana. Science 301: 653–657. doi:10.1126/science.1086391.

FIGURES:

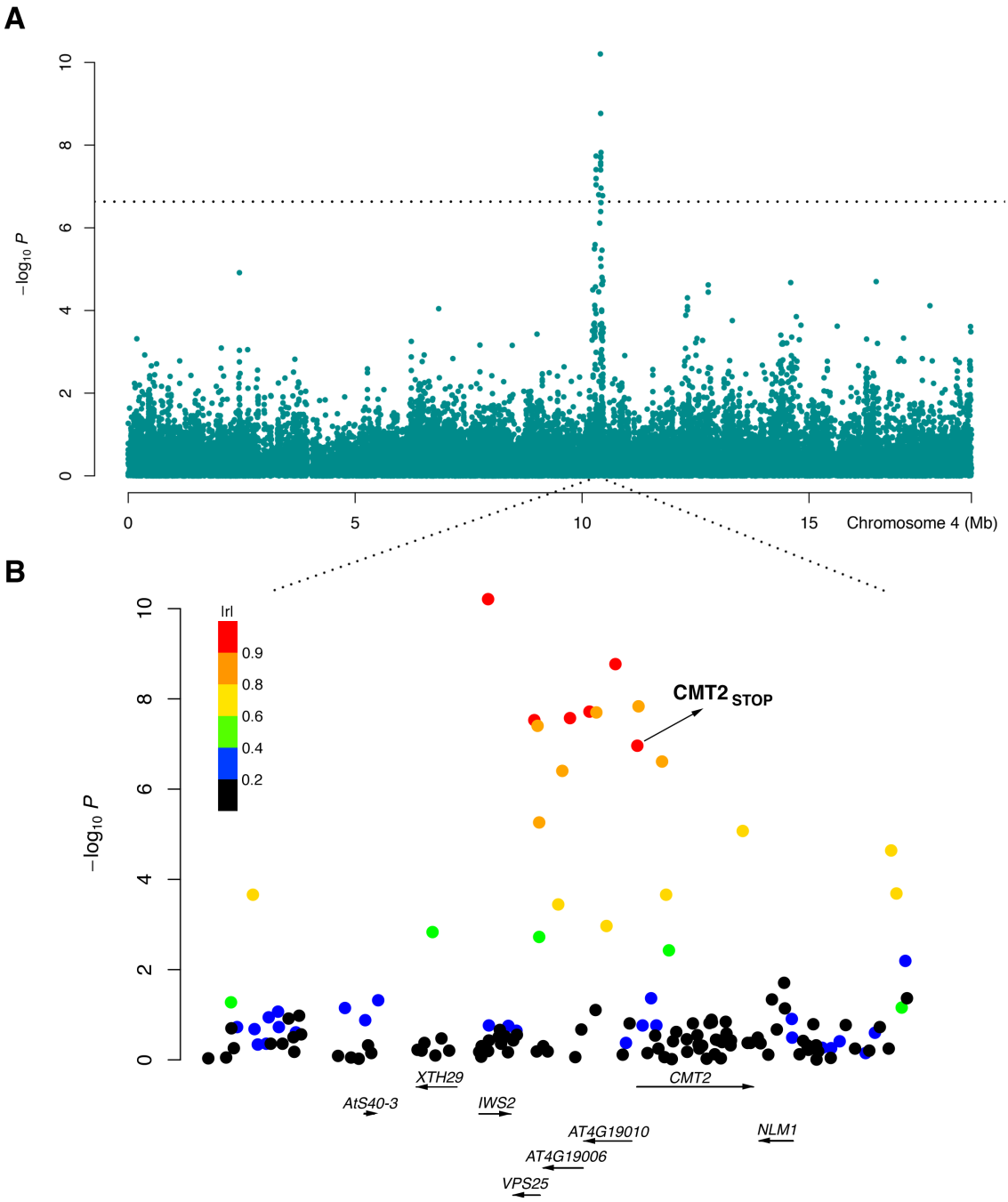


Figure 1. An LD block associated with temperature seasonality contains *CMT2*. A genome-wide significant variance-heterogeneity association signal was identified for temperature seasonality in the RegMap collection of natural *Arabidopsis thaliana* accessions [1]. The peak on chromosome 4 around 10 Mb (A) mapped to a haplotype block (B) containing a nonsense mutation (*CMT2*$_{STOP}$) early in the first exon of the Chromomethylase 2 (*CMT2*) gene. Color coding based on |r| (the absolute value of the correlation coefficient) as a measure of LD between each SNP in the region and the leading SNP in the association analysis.

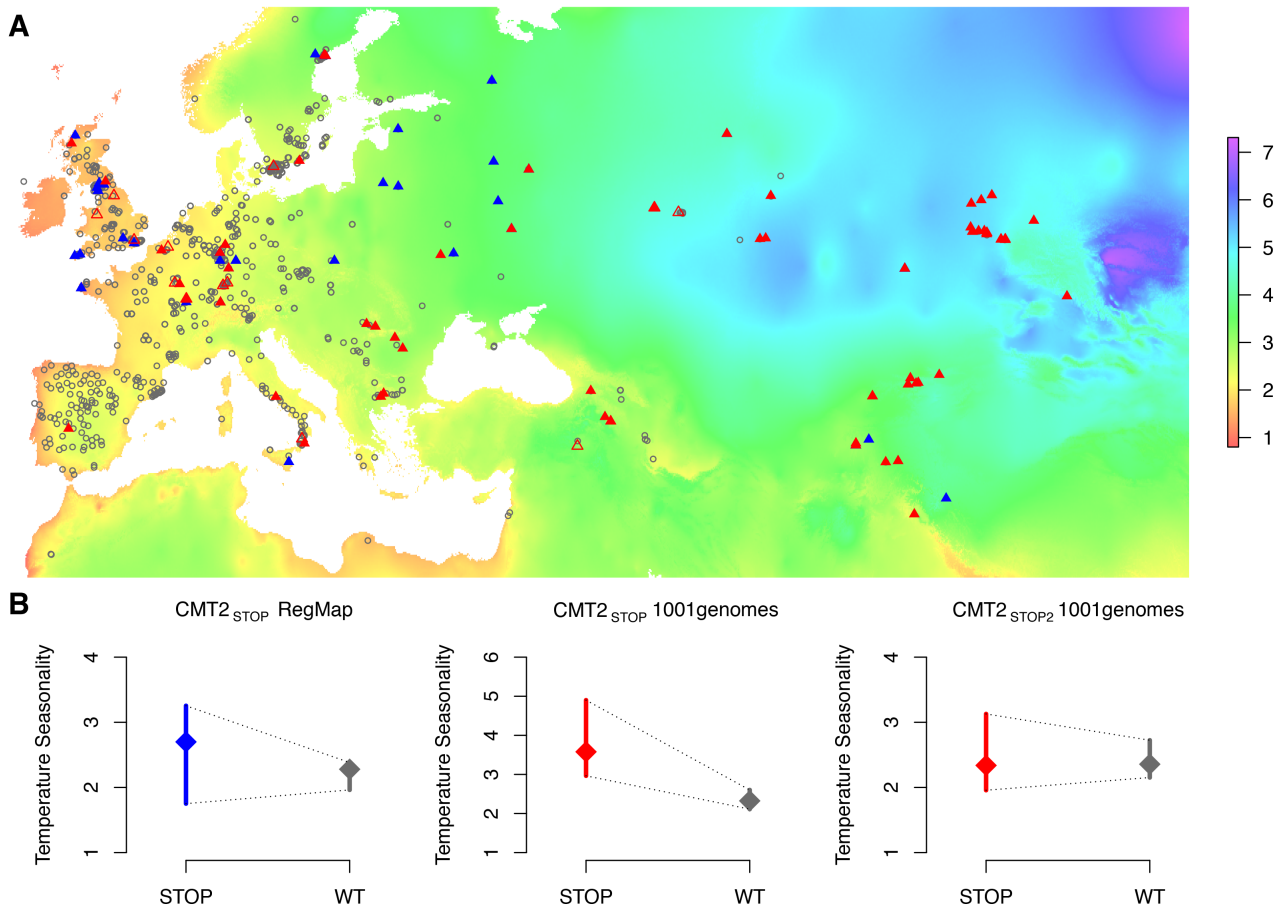


Figure 2. Geographic distribution of, and heterogenous variance for, three *CMT2* alleles in two collections of *A. thaliana* accessions. The geographic distributions (A) of the wild-type ($CMT2_{WT}$; gray open circles) and two nonsense alleles ($CMT2_{STOP}$/$CMT2_{STOP2}$; filled/open triangles) in the *CMT2* gene that illustrates a clustering of $CMT2_{WT}$ alleles in less variable regions and a greater dispersion of the nonsense alleles across different climates both in the RegMap [1] (blue) and the 1001-genomes [2](red) *A. thaliana* collections. The resulting variance-heterogeneity in temperature seasonality between genotypes is highly significant, as illustrated by the quantile plots in (B) where the median is indicated by a diamond and a bar representing the 25% to 75% quantile range. The color scale indicate the level of temperature seasonality across the map. The colorkey in (A) represent the temperature seasonality values, given as the standard-deviation in % of the mean temperature (K).

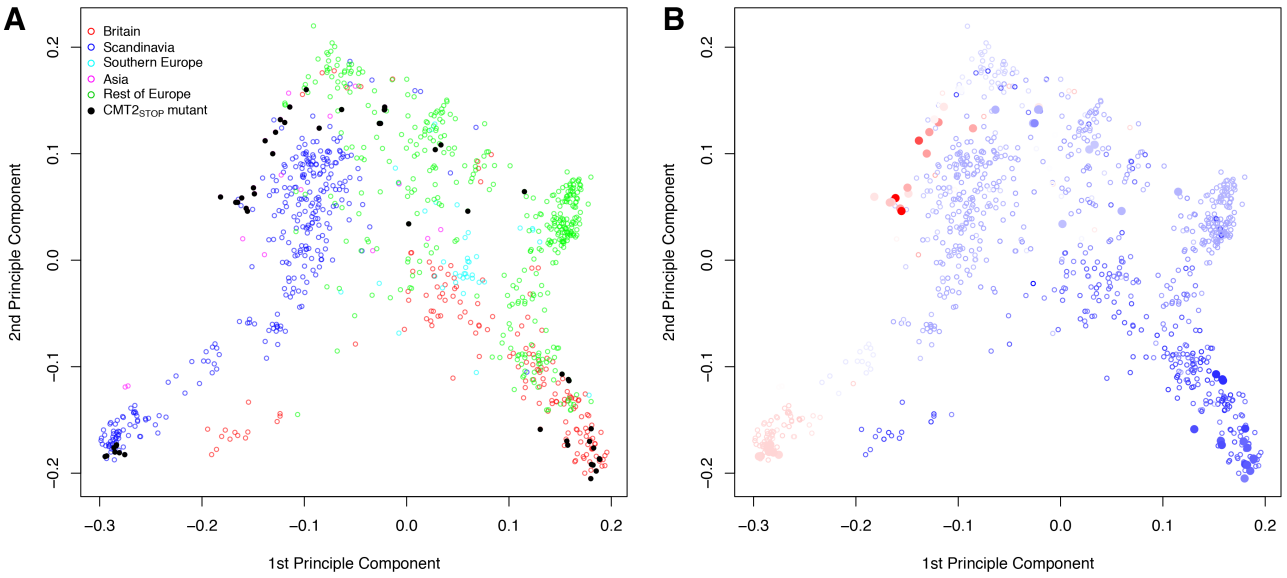


Figure 3. Principle components of the genomic kinship in the RegMap collection for the accessions carrying the alternative alleles at the Chromomethylase 2 locus (*CMT2STOP* and *CMT2WT* as filled and empty circles, respectively). Coloring is based on (A) geographical regions (defined as in Supplementary Fig. 37) and (B) temperature seasonality, ranging from dark blue (least variable) to red (most variable).

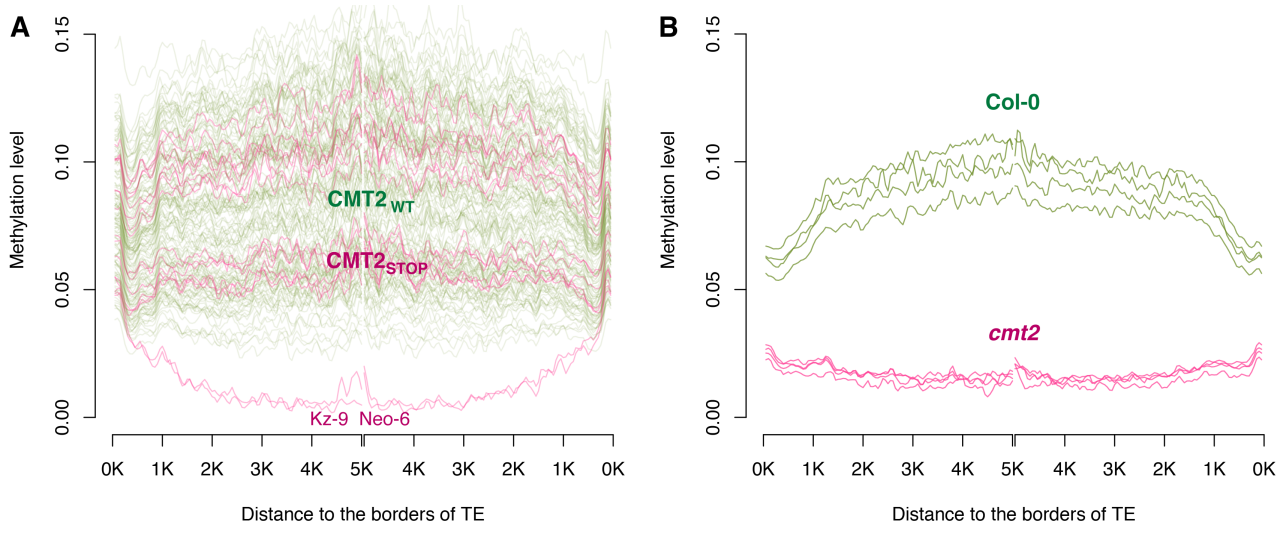


Figure 4. Comparison of CHH methylation patterns inside TE-bodies, (A) between *CMT2WT* and *CMT2STOP* accessions using the data from [7], and (B) between four replicate Col-0 wild-type and *cmt2* knock-outs from [24]. For each accession, the curve is to illustrate the moving average methylation level in a sliding 100 bp window. On the x-axis, the two different strands of DNA are aligned in the middle, truncated at 5 kb from the edge of the TEs.

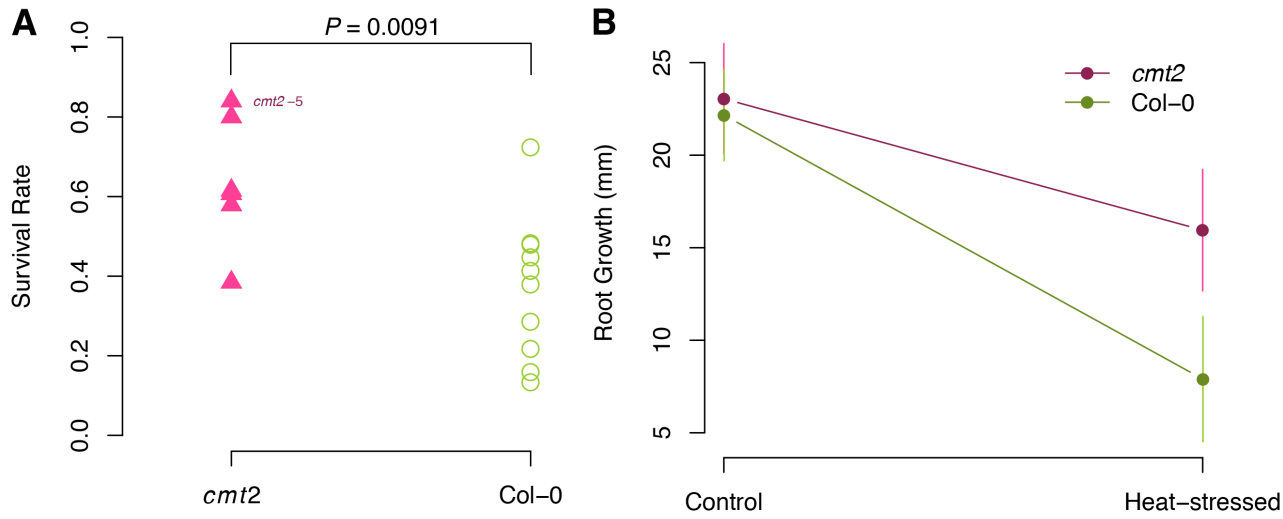


Figure 5. *cmt2* mutant plants display an increased tolerance to heat-stress. A. The survival rate is significantly higher for *cmt2-5* mutant than for Col-0 plants under severe heat-stress (24 h at 37.5°C). *P*-values in A were obtained using a log-linear regression. B. The *cmt2-5* mutant was also more tolerant to less severe heat-stress heat-stress (6 h at 37.5°C) than Col-0, here illustrated by its significantly faster growth of the root ($P = 0.026$; one-sided t-test) during the first 48h following heat stress.

TABLE:

Table 1. Loci with genome-significant, non-additive effects on climate adaptation and a functional analysis of nearby genes ($r^2 > 0.8$) containing missense or nonsense mutations.

Trait	Leading SNP					Selected Candidate Genes			Mutant Analysis[1]	
	Chrom	Pos (bp)	MAF[2]	P-value[3]	Effect[4]	Gene name	Locus	# Mut[5]	PASE	MSA
Temperature seasonality										
	2[a]	12 169 701	0.08	2.0E-08	0.53±0.08	*BGAL8*	*AT2G28470*	4	0.36[6]	0.72[6]
	4	10 406 018	0.05	6.1E-11	0.51±0.06	*IWS2*	*AT4G19000*	1	0.21	0.35
						CMT2	*AT4G19020*	7	STOP	STOP
Maximum temperature in the warmest month										
	1	6 936 457	0.05	1.9E-07	0.34±0.07		*AT1G19990*	1	0.64	0.2
Minimum temperature in the coldest month										
	2[b]	18 620 697	0.08	3.5E-08	0.33±0.05					
	2[c]	19 397 389	0.05	5.0E-08	0.38±0.06					
	5	14 067 526	0.07	4.2E-08	0.33±0.05		*AT5G35930*	1	0.30	0.05
	5[d]	18 397 418	0.11	1.2E-07	0.28±0.05					
Number of consecutive cold days										
	2[b]	18 620 697	0.08	1.7E-07	0.33±0.05					
	2[c]	19 397 389	0.05	7.2E-08	0.39±0.06					
	5	7 492 277	0.08	4.3E-09	0.38±0.06		*AT5G22560*	4	0.63	0.11
	5[d]	18 397 418	0.11	1.9E-07	0.29±0.05					
Day length in spring										
	2[a]	12 169 701	0.08	2.0E-07	0.46±0.08	*BGAL8*	*AT2G28470*	4	0.36	0.72
	3	12 642 006	0.07	9.4E-08	0.29±0.05					
	4[e]	14 788 320	0.08	2.2E-08	0.39±0.06	*VEL1*	*AT4G30200*	2	0.26	0.06
Relative humidity in spring										
	3	1 816 353	0.07	1.2E-08	0.39±0.06					
	4[e]	14 834 441	0.06	6.4E-08	0.49±0.08	*VEL1*	*AT4G30200*	2	0.26	0.06
					0.43±0.07	*XTH19*	*AT4G30290*	1	0.14	0.49
	5	8 380 640	0.07	6.3E-08						
Length of the growing season										
	3	576 148	0.08	1.4E-07	0.27±0.04					
Number of consecutive frost-free days										
	1	953 031	0.24	8.0E-08	0.25±0.04	*SOM*	*AT1G03790*	3	0.25	0.55
	1	6 463 065	0.08	1.9E-07	0.33±0.06					
	2	9 904 076	0.22	2.2E-08	0.22±0.04					

[a,b,c,d,e]Loci affecting affect multiple traits; [1]The predicted functional effect score for the strongest mis-sense mutation in the gene based on amino-acid physiochemical properties (PASE) and evolutionary conservation (MSA) [39]; [2]MAF: Minor Allele Frequency; [3]P-value: significance after genomic-control from a linear regression analysis of squared z-scores accounting for population stratification. [4]Effect: Standardized genetic effect on adaptability (Chi-square distributed) of squared z-scores accounting for population stratification; [5]#Mut: number mis- and non-sense mutations in the gene in the 1001-genomes dataset [2]. [6]Locus contains two missense mutations with equally strong predicted effects.

Supplementary Information for

Natural *CMT2* variation is associated with genome-wide methylation changes and temperature seasonality

Xia Shen, Jennifer De Jonge, Simon Forsberg, Mats Pettersson, Zheya Sheng, Lars Hennig and Örjan Carlborg*

*To whom correspondence should be addressed. E-mail: orjan.carlborg@slu.se

This PDF file includes:

Supplementary Text
Supplementary Fig. 1-46
Supplementary Table S1-S4

Additional Results

CMT2 is a potential target for the nonsense-mediated RNA decay (NMD) pathway

The $CMT2_{STOP}$ allele is predicted to produce an mRNA with a pre-mature translation termination codon, which makes it a likely target for the nonsense-mediated mRNA decay (NMD) pathway (Nicholson and Mühlemann 2010). Thus, accessions carrying $CMT2_{STOP}$ may have lower transcript abundance than those with wild-type *CMT2*. To test whether this was the case, we evaluated the levels of *CMT2* mRNA using two publicly available datasets. First, we analyzed the 19 genomes project data (Gan et al. 2011) that contained full genome sequences and transcriptome for 19 *A. thaliana* accessions, two of which (Ct-1 and Kn-0) are part of the RegMap panel and carry the $CMT2_{STOP}$ mutation. The difference in mRNA abundance was highly significant between *CMT2* mutant and wild-type accessions ($P = 6.9 \times 10^{-5}$), with a higher expression in the wild type. A similar analysis was done using a larger data set from[7], with CMT2 genotypes called from whole-genome re-sequencing data and using RNA-seq data from leaf tissue in 14 $CMT2_{STOP}$ and 92 $CMT2_{WT}$ accessions. As in the other dataset, the average mRNA abundance was higher for the CMT2 wild-type accessions, but here the difference was not significant ($P = 0.14$). However, the mRNA levels were significantly higher among the four $CMT2_{STOP}$ accessions that displayed a methylation pattern resembling that of the $CMT2_{WT}$ in the analysis above (t-test; $P = 0.01$). When those lines were omitted, the average mRNA level was significantly higher in the $CMT2_{WT}$ accessions than in the ten remaining $CMT2_{STOP}$ accessions (t-test; $P = 0.02$). Together, these results indicate that *CMT2* mRNA levels are often lower in $CMT2_{STOP}$ than in *CMT2* wild-type accessions. This is consistent with the notion that the premature stop codon does interfere with translation as expected.

Methods

Plant Material

A T-DNA insertion line (SAIL_906_G03, *cmt2*-5) (Figure S45) was ordered from NASC and seeds were produced from confirmed homozygous insertion plants. *CMT2* (AT4G19020) transcripts spanning the insertion site were assayed by RT-PCR with the following primers: forward_CMT2 CTCATTACCCCCAGAGGAGAG, reverse_CMT2 AAAGTTTGGGTCGAGGAAGAG. As a control a *PP2A* gene (AT1G13320) was used with primers forward_PP2A ATTCCGATAGTCGACCAAGC, reverse_PP2A: AACATCAACATCTGGGTCTTCA. RNA was extracted from five seedlings using the Trizol method as described previously (Steinbach and Hennig 2014). RNA was DNA treated with DNase I (Thermo Scientific) according to the manufacturer's instructions. 1 μg of RNA was used to synthesize cDNA with the RervertAid First Strand cDNA synthesis Kit (Thermo scientific) according to manufacturer's instructions. No CMT2 transcripts spanning the insertion site were detected strongly suggesting that *cmt2*-5 is a null allele.

Source Code

Example R source code for calculating HEM genomic kinship matrix

Here, we use the example data in the **bigRR** package: http://cran.r-project.org/web/packages/bigRR/ to illustrate how an ordinary identity-by-state (IBS) kinship matrix can be update to a HEM genomic kinship matrix. The full theoretical details on this procedure are provided in[13].

```
# load the bigRR package
require(bigRR)

# load the example data
data(Arabidopsis)
X <- matrix(1, length(y), 1)
Z <- scale(Z)

# fitting SNP-BLUP, i.e. a ridge regression on all the markers across the genome
SNP.BLUP <- bigRR(y = y, X = X, Z = Z, family = binomial(link = 'logit'))

# calculate HEM (heteroscedastic effects model) genomic kinship matrix
w <- as.numeric(SNP.BLUP$u^2/(1 - SNP.BLUP$leverage))
wZt <- sqrt(w)*t(Z)
G <- crossprod(wZt)
```

Temperature seasonality phenotype preparation in R

We present the source code for phenotyping of temperature seasonality in Euro-Asia. The data downloaded from http://www.worldclim.org/ were processed using the following code to obtain an object readable by the **raster** package: http://cran.r-project.org/web/packages/raster/.

```
# load the original data files
bil_files <- grep(".bil", dir("tmean_30s_bil Folder/"), value = T)
bil_file_order <- as.numeric(sub(pattern = ".+_([0-9]+).bil",
                  x = bil_files, replacement = "\\1"))
bil_files <- bil_files[order(bil_file_order)]

# create rasters
WorldClim_stack <- stack()
for (bil_file in bil_files){
r <- raster(paste("tmean_30s_bil Folder/",bil_file, sep = ""))
WorldClim_stack <- addLayer(WorldClim_stack, r)
```

```
}

# temperature seasonality calculation
r_mean <- calc(WorldClim_stack, mean)
save(r_mean, file = "WorldClim_mean.Rdata")
writeRaster(r_mean, file = "WorldClim_mean.raster")
r_sd <- calc(WorldClim_stack, sd)
save(r_sd, file = "WorldClim_sd.Rdata")
writeRaster(r_sd, filename = "WorldClim_sd.raster")
r_mean_corr <- r_mean/10 + 273.15
save(r_mean_corr, file = "WorldClim_mean_corr.Rdata")
r_coeff_var <- 100*(r_sd/10)/r_mean_corr

# output to a raster object
writeRaster(r_coeff_var, file = "WorldClim_coeff_var.raster")

# phenotyping at given coordinates
# (LONGITUDE and LATITUDE already loaded)
require(raster)
world_temp_seas <- raster('WorldClim_coeff_var')
temp_seas <- raster::extract(world_bio5, cbind(LONGITUDE, LATITUDE))
```

Comments

Statistical properties of the vGWAS in relation to population stratification

Inherent properties of the variance heterogeneity test decreases risk of identifying locally adapted alleles
An important property of the variance heterogeneity GWAS analysis is that it is inherently more powerful in detecting loci where the minor allele is associated with a higher variance than the major allele. In practice, there is no power to detect low-variance minor alleles in a GWAS setting[12,30]. Hence, the method facilitates detection of alternative (*i.e.* minor) alleles associated with a broader range of the climate variables than the reference (*i.e.* major) alleles. The method is powerful in finding associations to minor alleles associated with a broader range of climate-variables than the reference. Such alleles will, by definition, be present across a large proportion of the global population and due to this be considerably less affected by population structure. In **Figure 3** or **Figure S15** (top panel), we illustrate this property for the inferred loci using the *CMT2* locus as an example. The MDS-plot visualizes the distribution of the $CMT2_{STOP}$ allele across the population structure present in the RegMap collection using the pairwise genome-wide relationship between the accessions based on the first two principle-components of the kinship matrix. The link between the geographic origin of the

accessions and kinship is visualized by coloring the dots for each accession based on geographic origin. As expected, accessions from nearby regions (*e.g.* UK, Scandinavia and mainland Europe) are more related. The $CMT2_{STOP}$ allele is, however, not heavily confounded with population-structure and is present in most major sub-groups of the population (albeit with a higher frequency in Asia - see **Figure S36**).

Mixed models based vGWAS analyses to account for population structure via modeling of genome-wide kinship We statistically deal with the strong correlations that exist between climate & population structure (*e.g.* along east-west/north-south clines) using a mixed-model based approach accounting for genomic kinship combined with genomic control. This approach has earlier been shown to control type I errors (*i.e.* genome-wide *P*-value inflation) in structured populations. The major challenge in analyzing this population is thus not the false-positive rate, as also standard GWAS analyses can be implemented in the same mixed-models framework, but rather to avoid unacceptably high type II errors (*i.e.* low power) for traits confounded with population-structure. Traditional GWAS analyses model alleles to have a linear relationship with climate, which in practice means that they mostly coincide with the population-structure along geographic clines. Hence, analyses will either be prone to identify false associations (when population-structure is not accounted for), or be under-powered (when accounting for population-structure). Although this is not explicitly discussed in the earlier reports based on this data, this is the primary reason for their lack of genome-wide significant associations to individual adaptive loci. As illustrated in **Figure 3**, the variance-heterogeneity test identifies loci present across population strata, where the signal therefore remain even after accounting for population structure via the mixed-model approach. The independence between the effect of the inferred locus and population structure can be evaluated statistically by fitting a linear mixed model where the genotype is regressed on the genomic kinship, where the heritability differs from 0 when confounding is present. For *CMT2* this estimate is zero, showing that the *CMT2* genotype is not confounded with population structure in this data.

On the power of vGWAS and GWAS analyses in highly structured populations There are several reasons for why a low overlap is expected between the results from traditional GWAS/selective-sweep analyses (as performed earlier) and the variance-heterogeneity GWAS (vGWAS) used here. First, in the absence of population stratification, the GWAS is more powerful than the vGWAS. In the presence of population stratification, however, loci affecting the mean phenotype will often be highly confounded with population-structure as they are a main genetic mechanism leading to local adaptation. In order to infer such loci when controlling for population structure, the same alleles need to have been under selection in multiple, unrelated populations, which is apparently a rare event as no such loci could be detected in the earlier studies of climate adaptation. The population genetics forces acting on variance-controlling loci are still poorly explored. Studies have, however, shown that high- and low-variance alleles are likely to co-exist in the population over extended periods of time at a frequency balanced depending on fluctuations in the surrounding environment that the population adapts to[11]. Due to this, both alleles are more likely to be present across different population strata than mean affecting alleles and therefore be less confounded with population structure. In **Figure S36**, we exemplify this by visualizing the allele-frequency of the $CMT2_{STOP}$ allele across the major sub-populations in the RegMap

and 1001-genomes data. Although the uneven sampling across the regions in the 1001-genomes makes allele-frequency estimates uncertain, the overall picture shows that minor allele is present across all sub-populations at a lower frequency and that it has increased in frequency in Asia.

Second, a traditional GWAS searches for difference in means between genotypes, whereas the vGWAS searches for differences in variances between genotypes. As these are two different statistical properties of the phenotypic distribution, the basic assumption is that they are both statistically and biologically unrelated and consequently the loci identified by the two methods are not expected to overlap. Although some degree of overlap might be expected, *e.g.* in situations where the variance scales with the mean, the high-significance required to reach genome-wide significance in the testing, in practice only loci with strong, pure effects on one of the statistical moments seem to be able to reach such significance levels[12]. Formal comparisons between the results in the association studies will thus be misleading to the readers, as these will indicate that an overlap is to be expected. Results from evaluations of the overlap for sub-GWAS signals to explore the potential overlap of loci with weaker effects on both the mean and the variance shows some overlap (**Figure S40-S44**). It should be noted, however, that comparisons of overlap at individual loci is not appropriate at these significance levels due to the lack of proper control of the type I error rate. The overall conclusions from these comparisons is i) that the power is generally very low for the GWAS after control for population stratification and ii) that even the sub-GWAS overlap is low for the two methods, but the overlap that exists is consistent with the correlation between the climate variables.

Third, the earlier studies have also inferred loci using traditional selective-sweep mapping. These analyses are designed to infer hard selective-sweeps where (potentially) adaptive alleles are assumed to have increased in frequency due to directional selection. As discussed above, the population genomic dynamics of plastic alleles does not follow the same pattern as for alleles affecting the mean[11], leading to a co-existence of the alleles over prolonged periods of time. This means that they will not be surrounded by a traditional genomic footprint of directional selection that can be detected in a selective-sweep analysis and one would not expect any overlap between the loci inferred in the selective sweep and vGWAS based analyses.

References

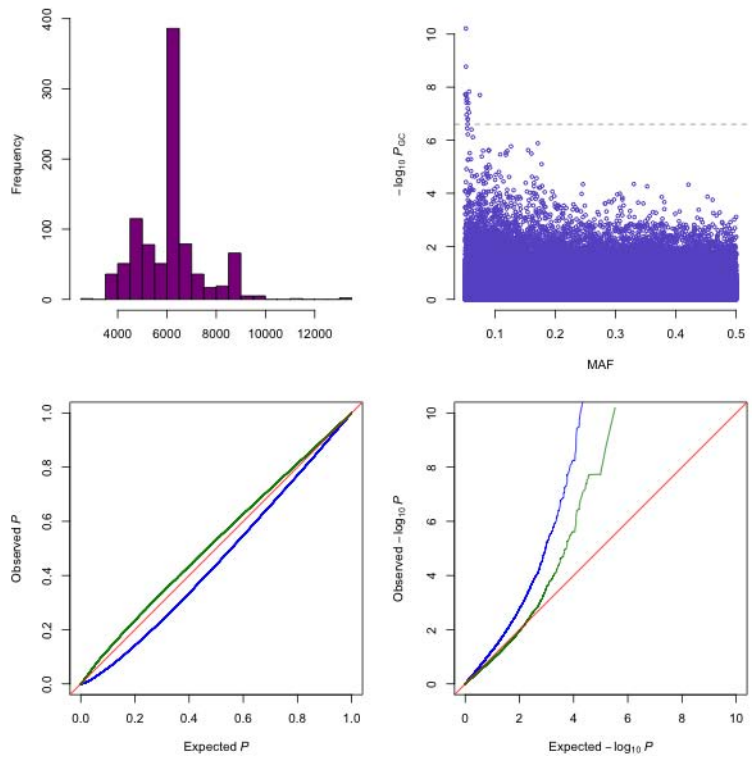
X. Gan, O. Stegle, J. Behr, J. G. Steffen, P. Drewe, K. L. Hildebrand, R. Lyngsoe, S. J. Schultheiss, E. J. Osborne, V. T. Sreedharan, A. Kahles, R. Bohnert, G. Jean, P. Derwent, P. Kersey, E. J. Belfield, N. P. Harberd, E. Kemen, C. Toomajian, P. X. Kover, R. M. Clark, G. Rätsch, and R. Mott. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature*, 477(7365): 419–423, Sept. 2011. doi: 10.1038/nature10414. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=21874022&retmode=ref&cmd=prlinks.

P. Nicholson and O. Mühlemann. Cutting the nonsense: the degradation of PTC-containing mRNAs. *Biochem Soc Trans*, 38(6):1615–1620, 2010.

Y. Steinbach and L. Hennig. Arabidopsis MSI1 functions in photoperiodic flowering time control. *Front Plant Sci*, 5(77), 2014.

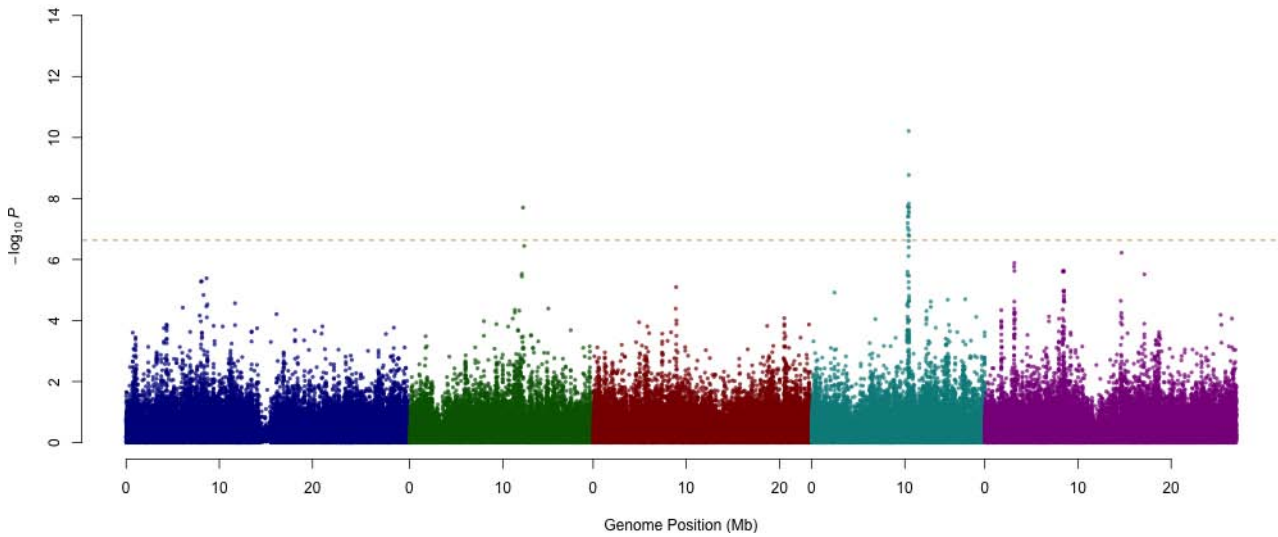
a: Phenotypic and *p*-value distributions.

Top-left: phenotypic distribution; Top-right: -$\log_{10}p$-values after genomic control (GC) against minor allele frequencies (MAF); Bottom panels: Quantile-quantile plots of *p*-values and -$\log_{10}p$-values before (blue) and after (green) GC.



b: Genome-wide association mapping for climate adaptability.

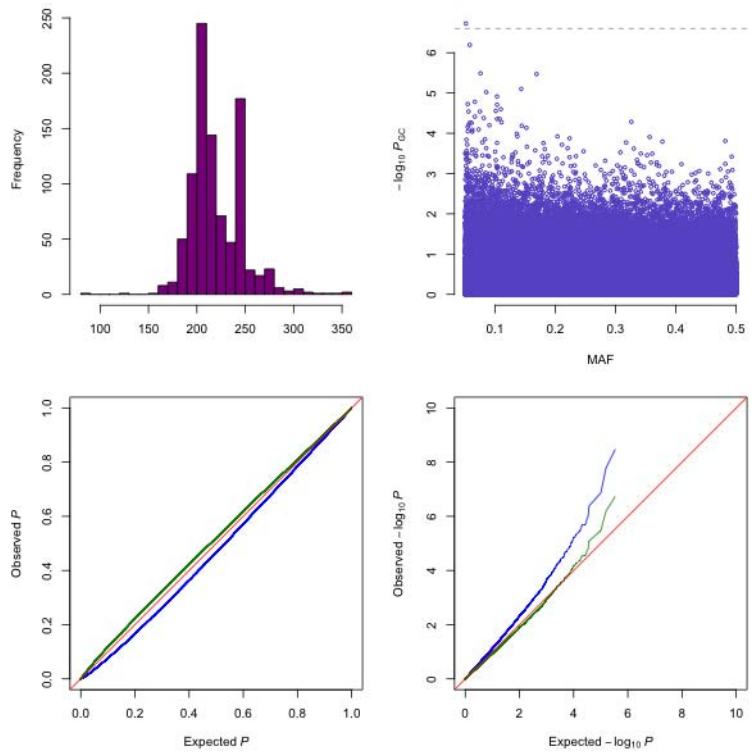
The plotted -$\log_{10}p$-values are genomic controlled. Markers with minor allele frequencies less than 5% are removed. Chromosomes are distinguished by colors. The Bonferroni-corrected significance threshold is marked by the horizontal line.



Supplementary Figure 1 - Summary of results for temperature seasonality.

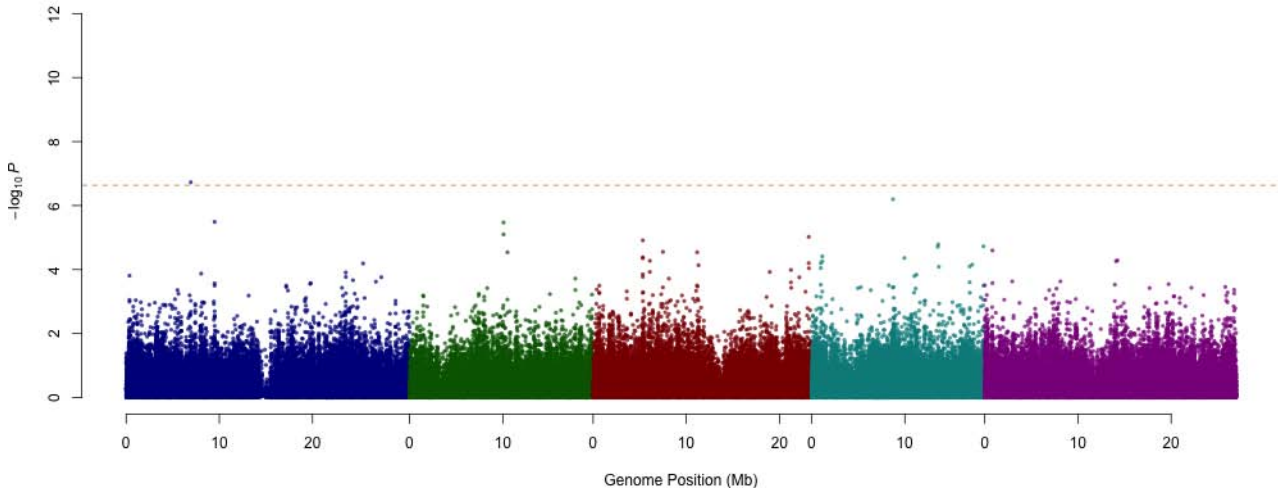
a: Phenotypic and *p*-value distributions.

Top-left: phenotypic distribution; Top-right: -$\log_{10}p$-values after genomic control (GC) against minor allele frequencies (MAF); Bottom panels: Quantile-quantile plots of *p*-values and -$\log_{10}p$-values before (blue) and after (green) GC.



b: Genome-wide association mapping for climate adaptability.

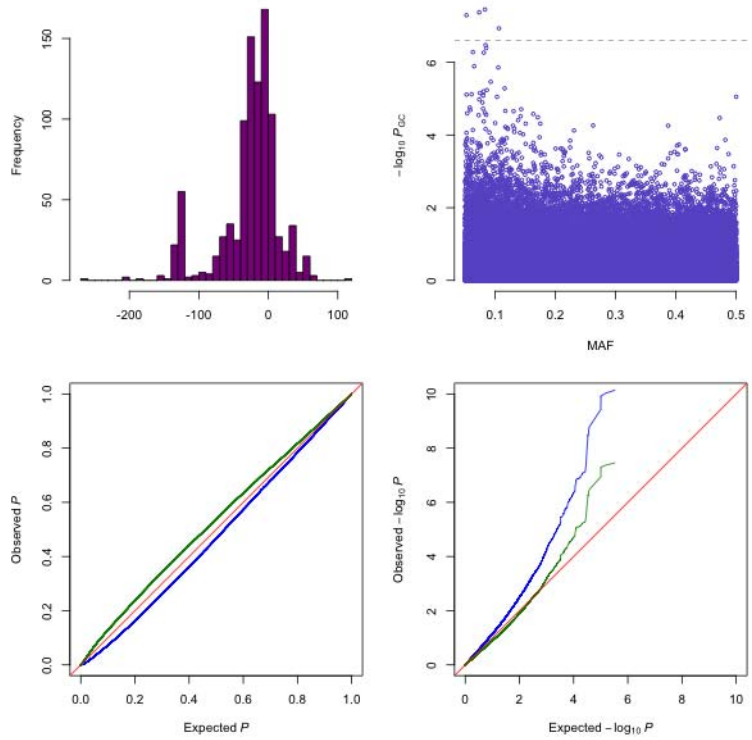
The plotted -$\log_{10}p$-values are genomic controlled. Markers with minor allele frequencies less than 5% are removed. Chromosomes are distinguished by colors. The Bonferroni-corrected significance threshold is marked by the horizontal line.



Supplementary Figure 2 - Summary of results for maximum temperature in the warmest month.

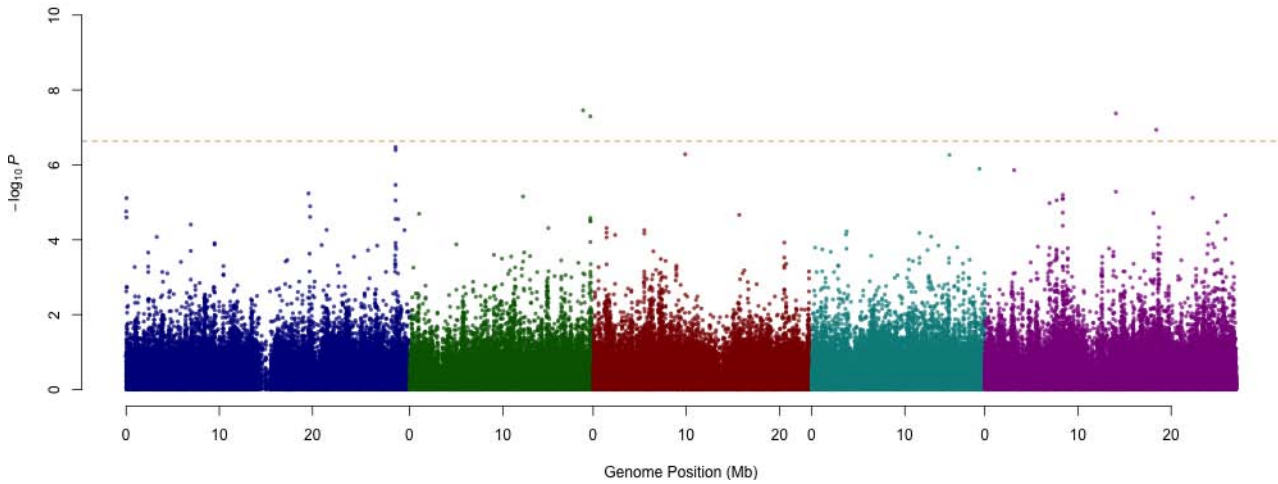
a: Phenotypic and *p*-value distributions.

Top-left: phenotypic distribution; Top-right: -$\log_{10}p$-values after genomic control (GC) against minor allele frequencies (MAF); Bottom panels: Quantile-quantile plots of *p*-values and -$\log_{10}p$-values before (blue) and after (green) GC.



b: Genome-wide association mapping for climate adaptability.

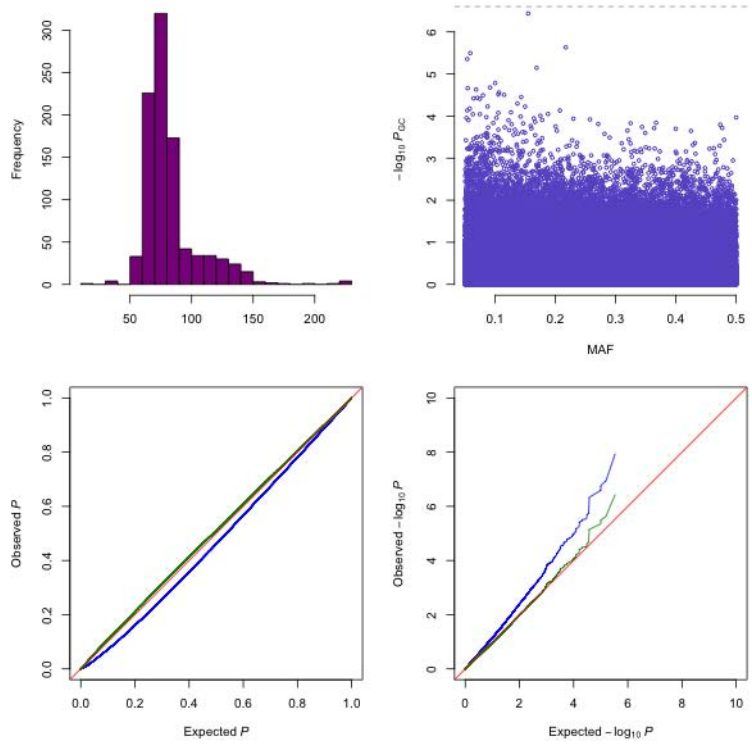
The plotted -$\log_{10}p$-values are genomic controlled. Markers with minor allele frequencies less than 5% are removed. Chromosomes are distinguished by colors. The Bonferroni-corrected significance threshold is marked by the horizontal line.



Supplementary Figure 3 - Summary of results for minimum temperature in the coldest month.

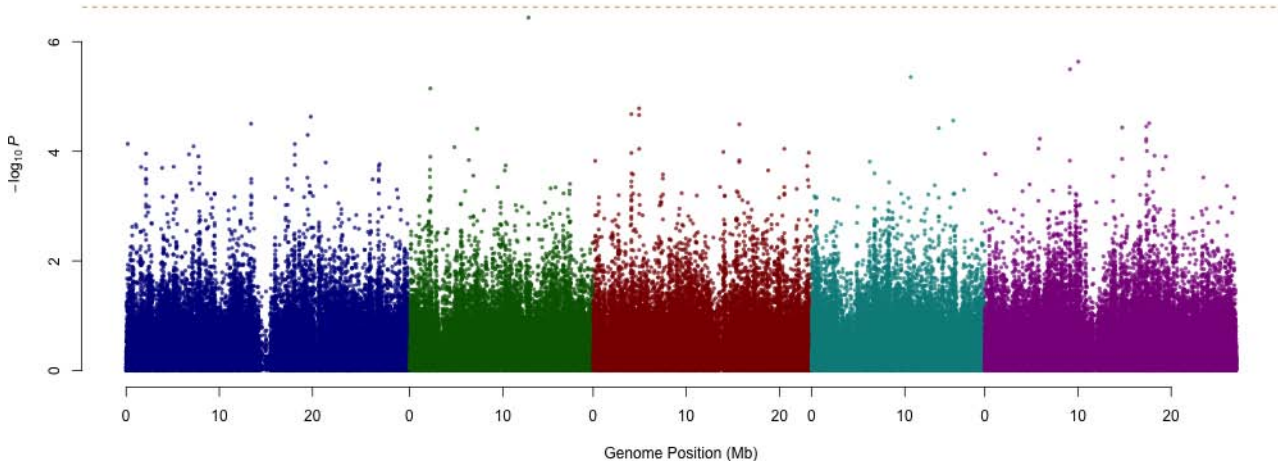
a: Phenotypic and *p*-value distributions.

Top-left: phenotypic distribution; Top-right: -$\log_{10}p$-values after genomic control (GC) against minor allele frequencies (MAF); Bottom panels: Quantile-quantile plots of *p*-values and -$\log_{10}p$-values before (blue) and after (green) GC.



b: Genome-wide association mapping for climate adaptability.

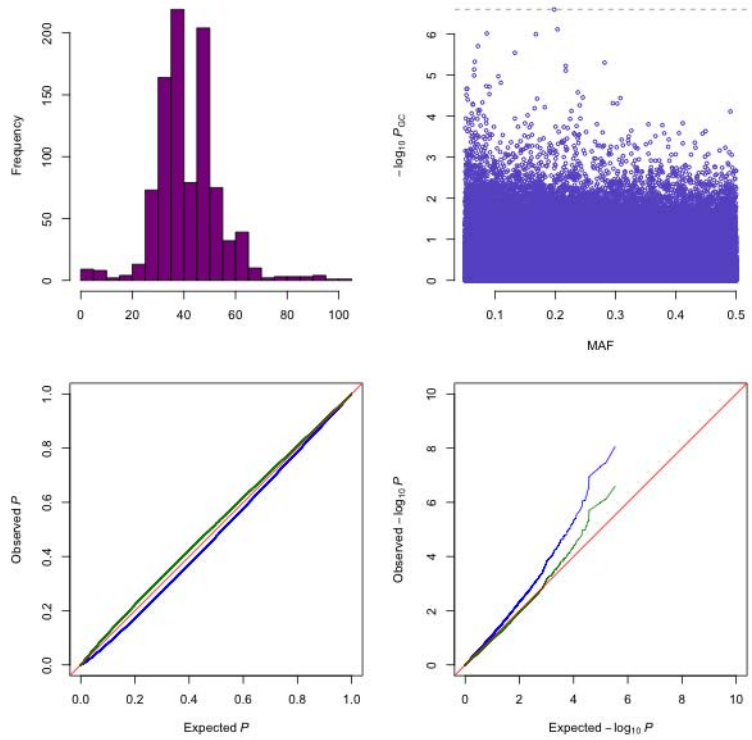
The plotted -$\log_{10}p$-values are genomic controlled. Markers with minor allele frequencies less than 5% are removed. Chromosomes are distinguished by colors. The Bonferroni-corrected significance threshold is marked by the horizontal line.



Supplementary Figure 4 - Summary of results for precipitation in the wettest month.

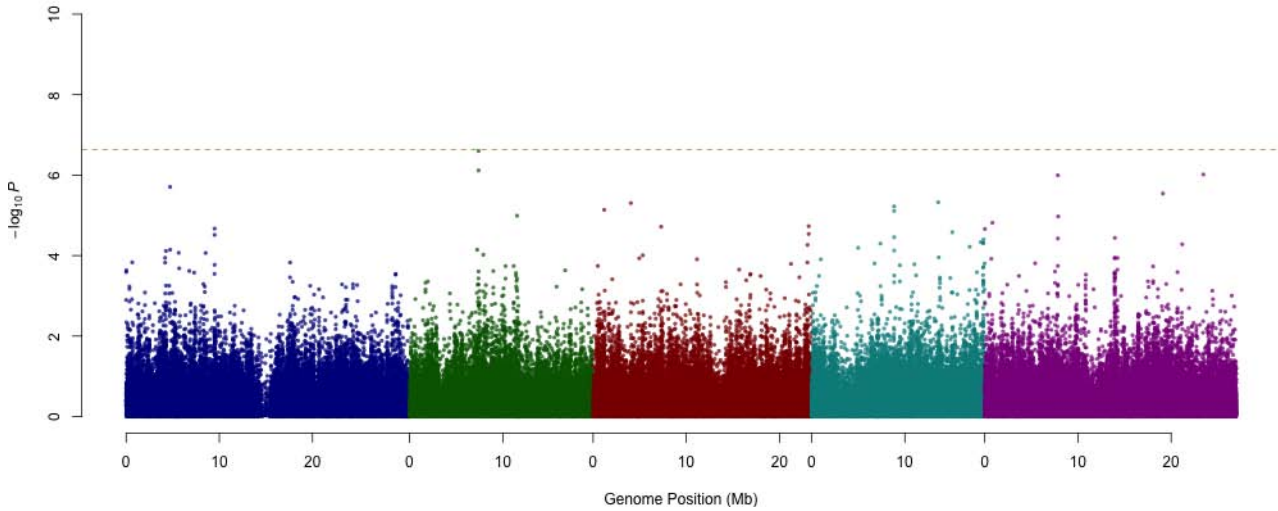
a: Phenotypic and *p*-value distributions.

Top-left: phenotypic distribution; Top-right: -$\log_{10}p$-values after genomic control (GC) against minor allele frequencies (MAF); Bottom panels: Quantile-quantile plots of *p*-values and -$\log_{10}p$-values before (blue) and after (green) GC.



b: Genome-wide association mapping for climate adaptability.

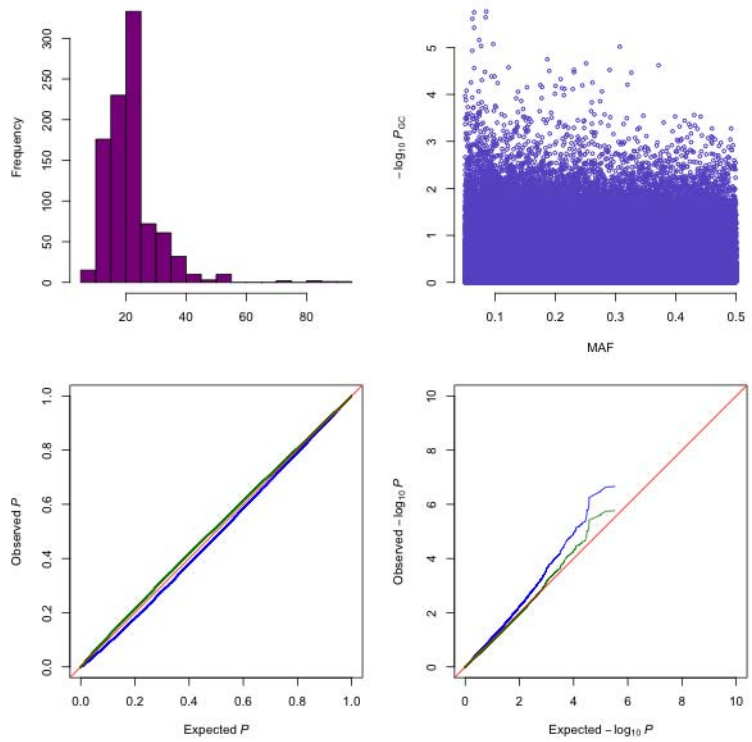
The plotted -$\log_{10}p$-values are genomic controlled. Markers with minor allele frequencies less than 5% are removed. Chromosomes are distinguished by colors. The Bonferroni-corrected significance threshold is marked by the horizontal line.



Supplementary Figure 5 - Summary of results for precipitation in the driest month.

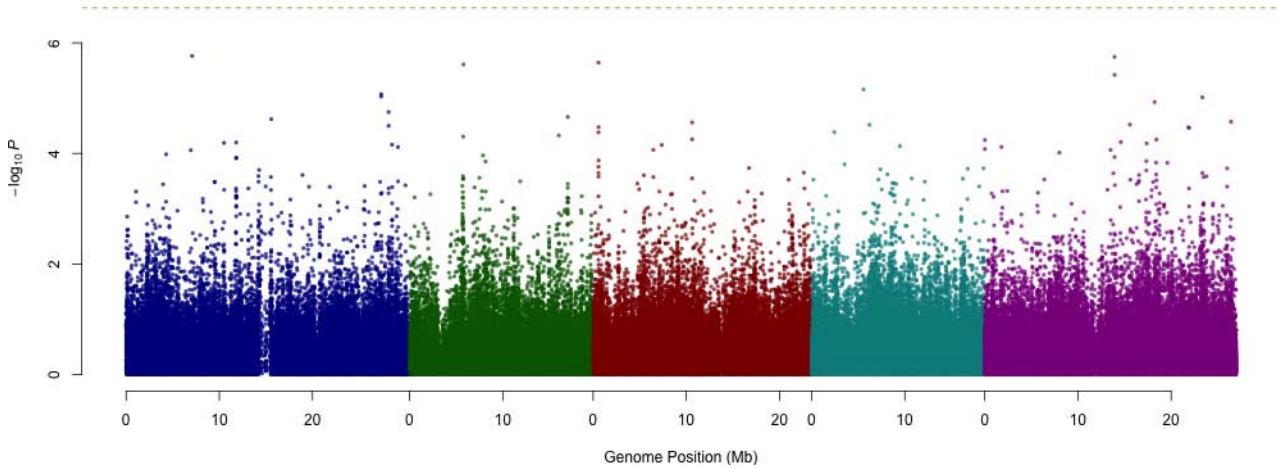
a: Phenotypic and *p*-value distributions.

Top-left: phenotypic distribution; Top-right: -$\log_{10}p$-values after genomic control (GC) against minor allele frequencies (MAF); Bottom panels: Quantile-quantile plots of *p*-values and -$\log_{10}p$-values before (blue) and after (green) GC.



b: Genome-wide association mapping for climate adaptability.

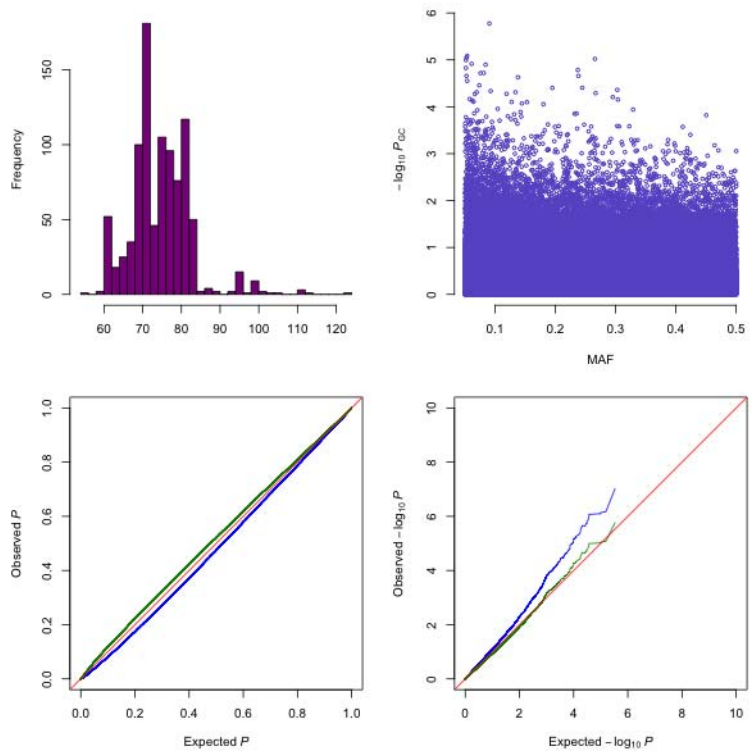
The plotted -$\log_{10}p$-values are genomic controlled. Markers with minor allele frequencies less than 5% are removed. Chromosomes are distinguished by colors. The Bonferroni-corrected significance threshold is marked by the horizontal line.



Supplementary Figure 6 - Summary of results for precipitation CV.

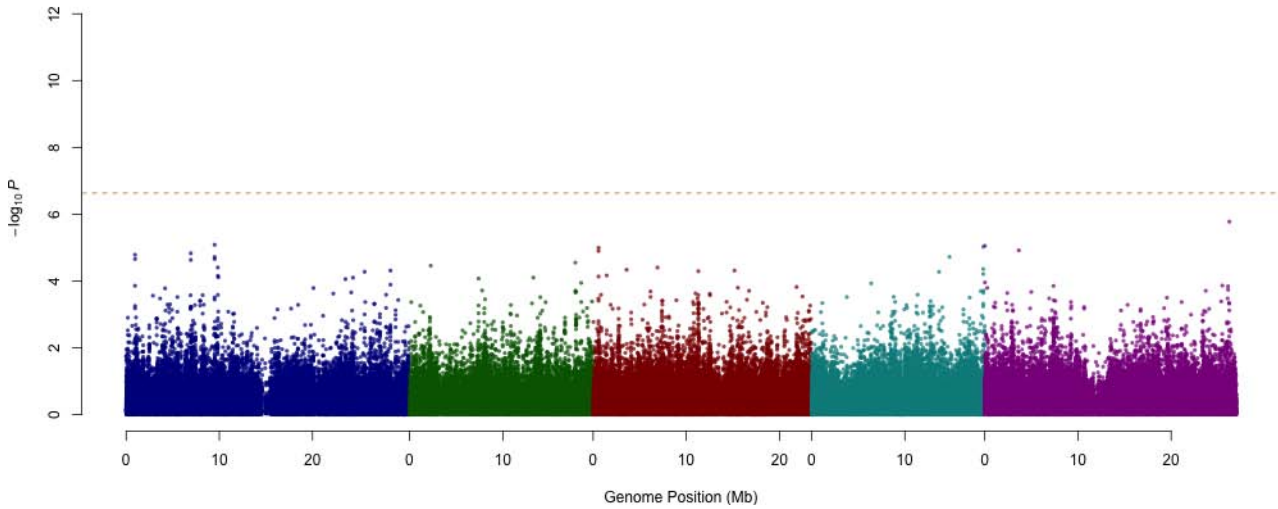
a: Phenotypic and *p*-value distributions.

Top-left: phenotypic distribution; Top-right: -$\log_{10}p$-values after genomic control (GC) against minor allele frequencies (MAF); Bottom panels: Quantile-quantile plots of *p*-values and -$\log_{10}p$-values before (blue) and after (green) GC.



b: Genome-wide association mapping for climate adaptability.

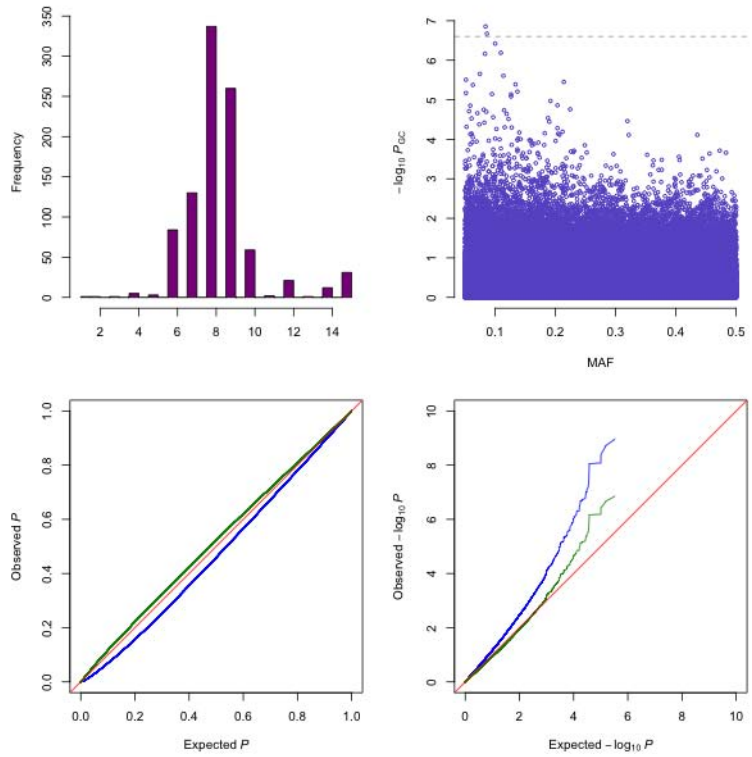
The plotted -$\log_{10}p$-values are genomic controlled. Markers with minor allele frequencies less than 5% are removed. Chromosomes are distinguished by colors. The Bonferroni-corrected significance threshold is marked by the horizontal line.



Supplementary Figure 7 - Summary of results for photosynthetically active radiation in spring.

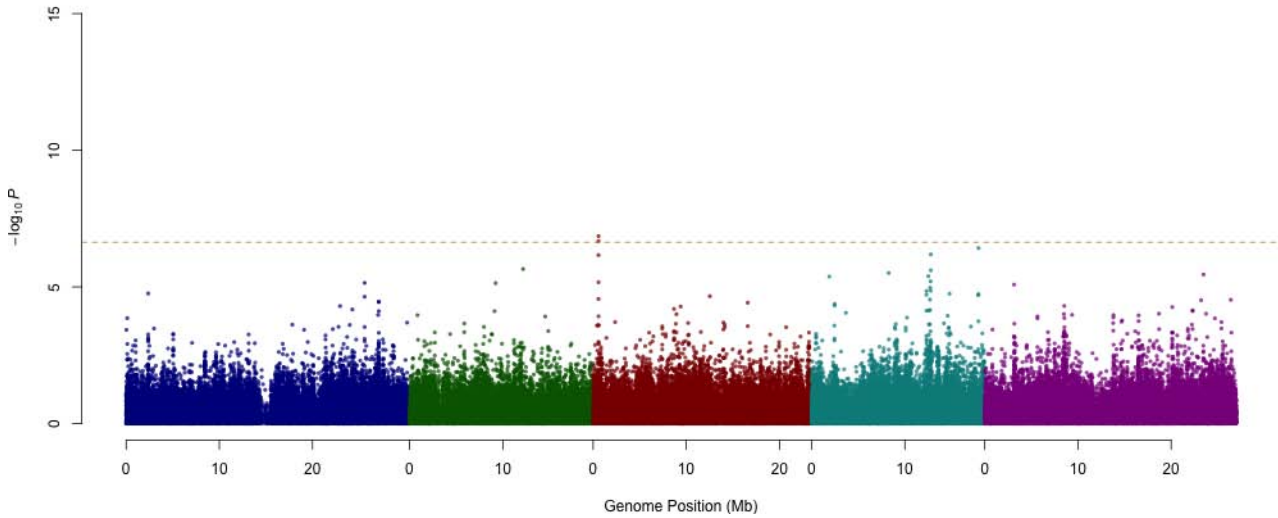
a: Phenotypic and *p*-value distributions.

Top-left: phenotypic distribution; Top-right: -$\log_{10}p$-values after genomic control (GC) against minor allele frequencies (MAF); Bottom panels: Quantile-quantile plots of *p*-values and -$\log_{10}p$-values before (blue) and after (green) GC.



b: Genome-wide association mapping for climate adaptability.

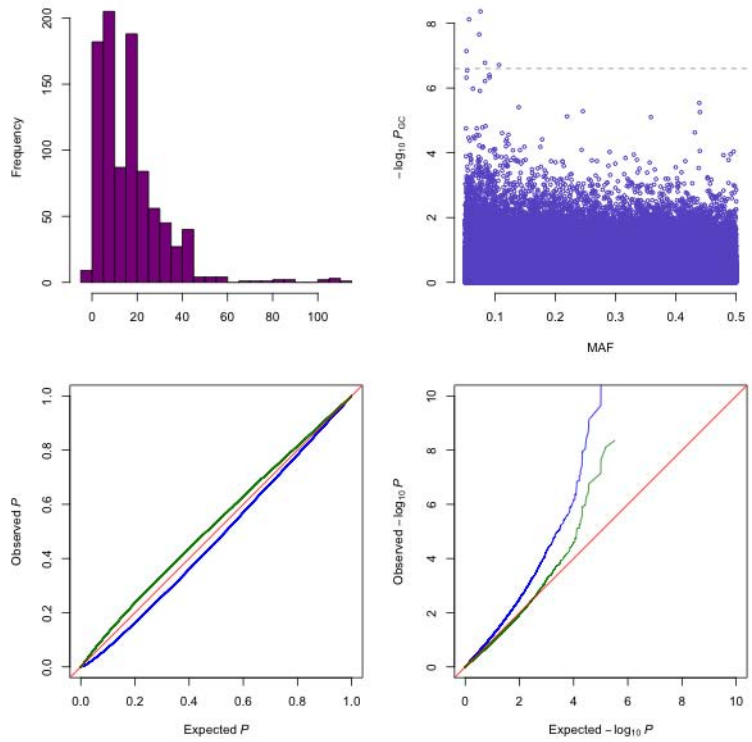
The plotted -$\log_{10}p$-values are genomic controlled. Markers with minor allele frequencies less than 5% are removed. Chromosomes are distinguished by colors. The Bonferroni-corrected significance threshold is marked by the horizontal line.



Supplementary Figure 8 - Summary of results for length of the growing season.

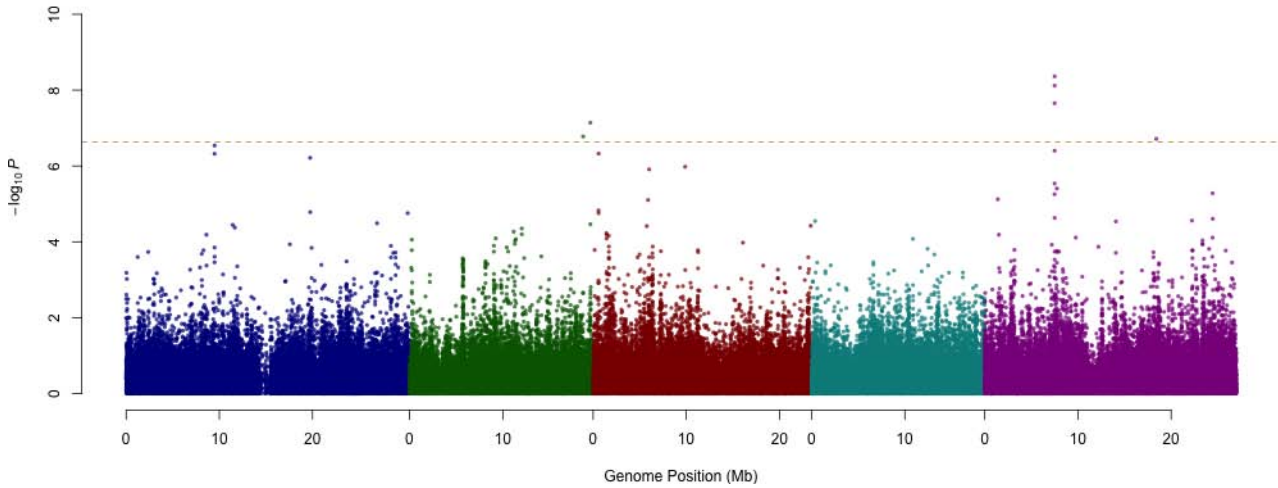
a: Phenotypic and *p*-value distributions.

Top-left: phenotypic distribution; Top-right: -$\log_{10}p$-values after genomic control (GC) against minor allele frequencies (MAF); Bottom panels: Quantile-quantile plots of *p*-values and -$\log_{10}p$-values before (blue) and after (green) GC.



b: Genome-wide association mapping for climate adaptability.

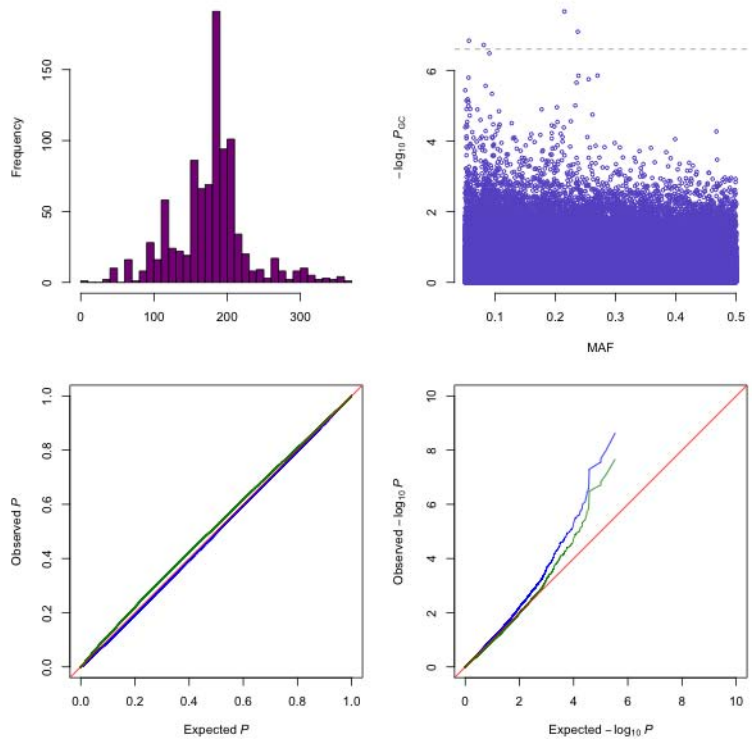
The plotted -$\log_{10}p$-values are genomic controlled. Markers with minor allele frequencies less than 5% are removed. Chromosomes are distinguished by colors. The Bonferroni-corrected significance threshold is marked by the horizontal line.



Supplementary Figure 9 - Summary of results for number of consecutive cold days.

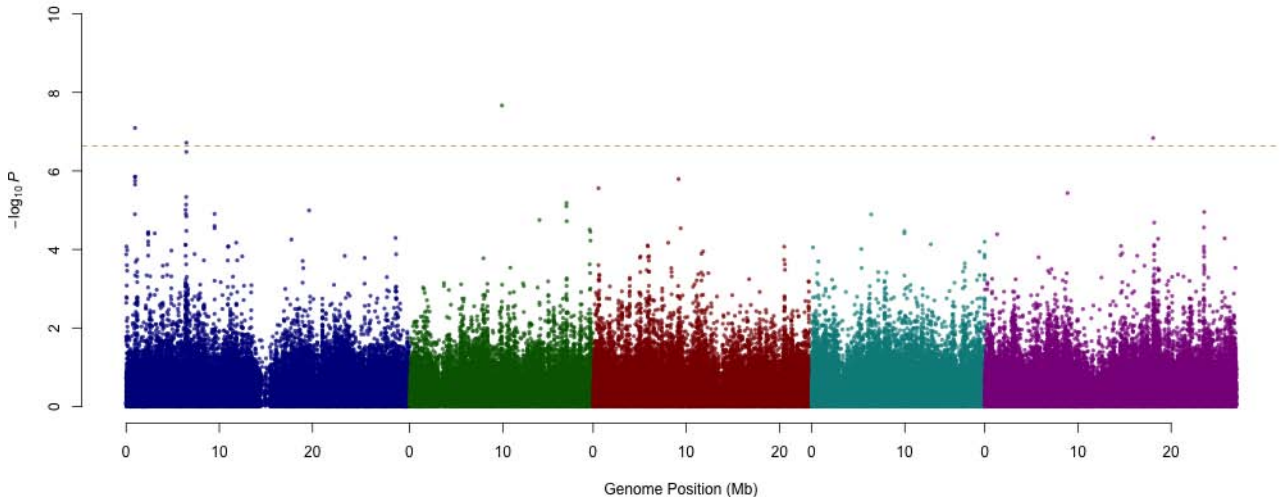
a: Phenotypic and *p*-value distributions.

Top-left: phenotypic distribution; Top-right: -$\log_{10}p$-values after genomic control (GC) against minor allele frequencies (MAF); Bottom panels: Quantile-quantile plots of *p*-values and -$\log_{10}p$-values before (blue) and after (green) GC.



b: Genome-wide association mapping for climate adaptability.

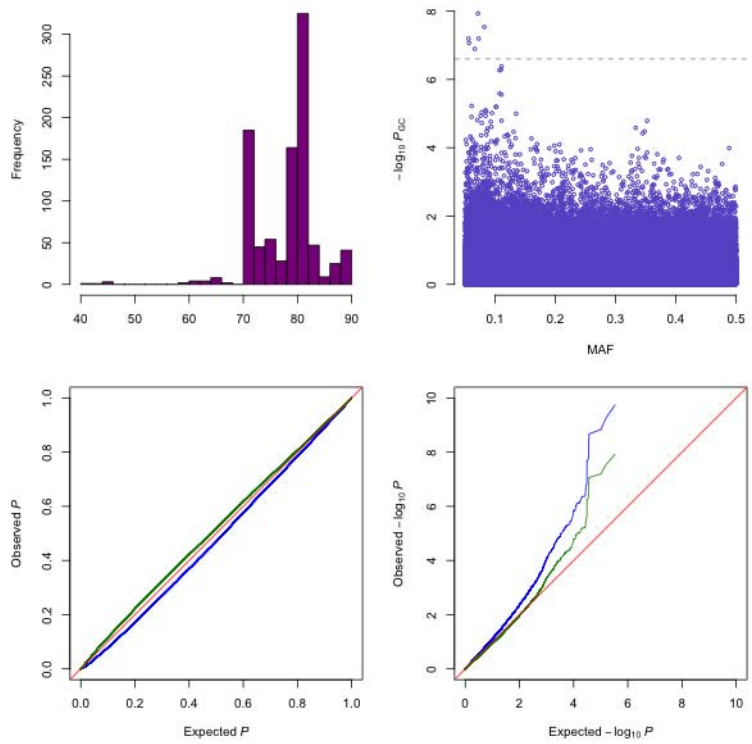
The plotted -$\log_{10}p$-values are genomic controlled. Markers with minor allele frequencies less than 5% are removed. Chromosomes are distinguished by colors. The Bonferroni-corrected significance threshold is marked by the horizontal line.



Supplementary Figure 10 - Summary of results for number of consecutive frost-free days.

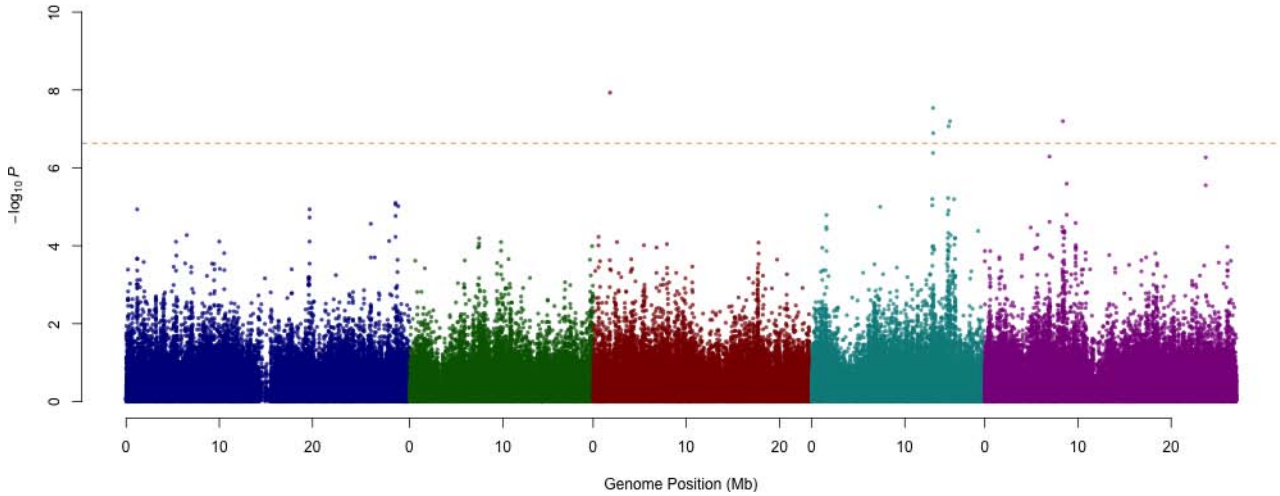
a: Phenotypic and *p*-value distributions.

Top-left: phenotypic distribution; Top-right: -$\log_{10}p$-values after genomic control (GC) against minor allele frequencies (MAF); Bottom panels: Quantile-quantile plots of *p*-values and -$\log_{10}p$-values before (blue) and after (green) GC.



b: Genome-wide association mapping for climate adaptability.

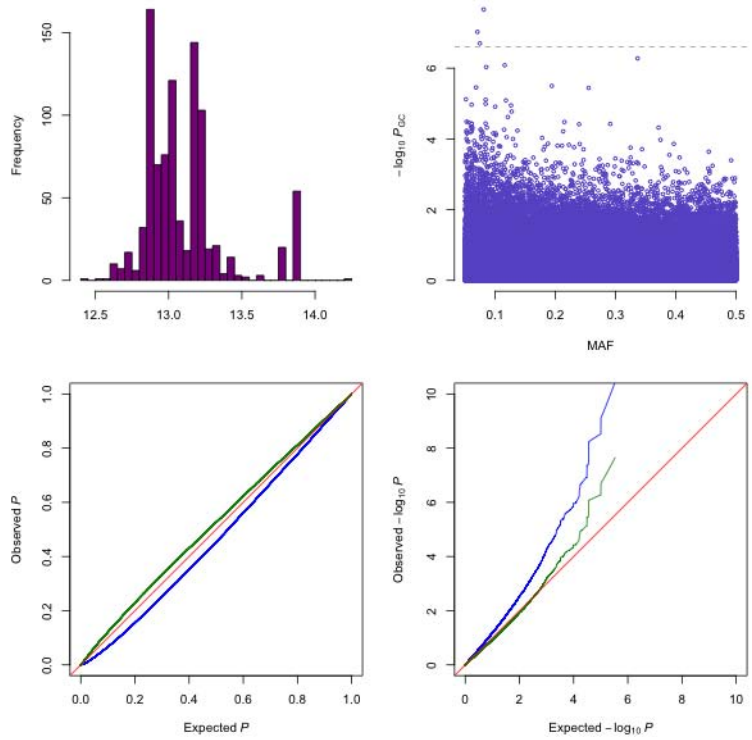
The plotted -$\log_{10}p$-values are genomic controlled. Markers with minor allele frequencies less than 5% are removed. Chromosomes are distinguished by colors. The Bonferroni-corrected significance threshold is marked by the horizontal line.



Supplementary Figure 11 - Summary of results for relative humidity in spring.

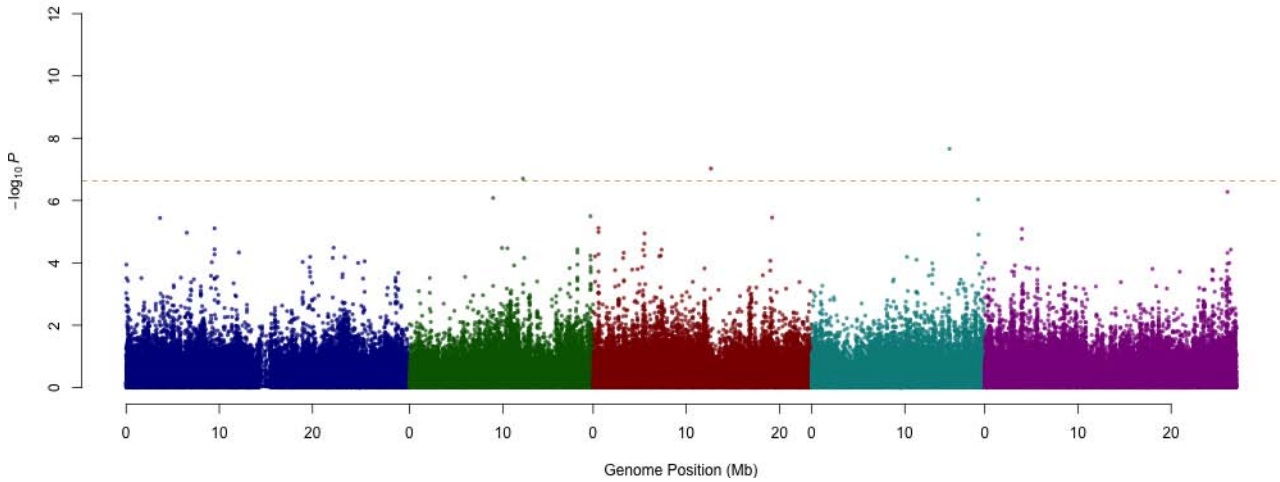
a: Phenotypic and *p*-value distributions.

Top-left: phenotypic distribution; Top-right: -$\log_{10}p$-values after genomic control (GC) against minor allele frequencies (MAF); Bottom panels: Quantile-quantile plots of *p*-values and -$\log_{10}p$-values before (blue) and after (green) GC.



b: Genome-wide association mapping for climate adaptability.

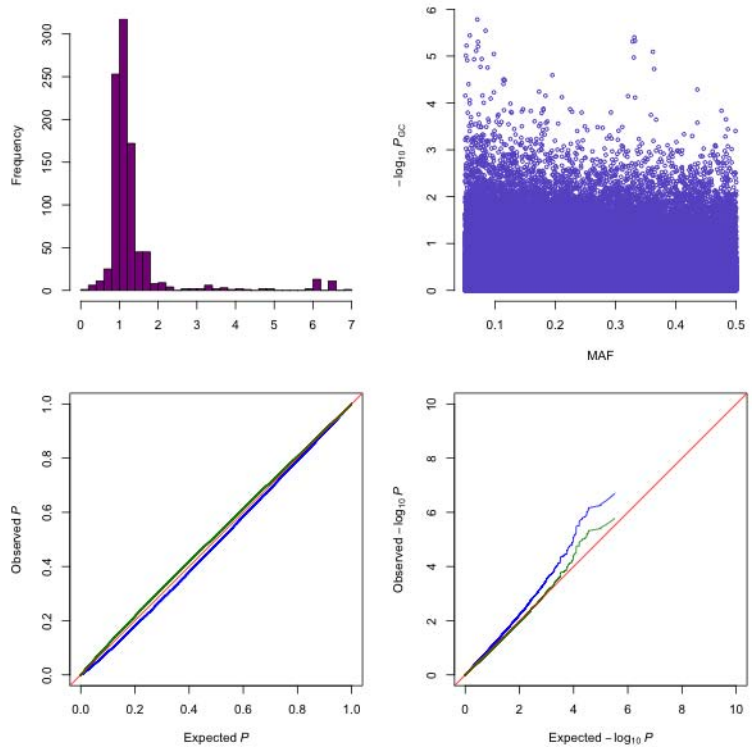
The plotted -$\log_{10}p$-values are genomic controlled. Markers with minor allele frequencies less than 5% are removed. Chromosomes are distinguished by colors. The Bonferroni-corrected significance threshold is marked by the horizontal line.



Supplementary Figure 12 - Summary of results for daylength in spring.

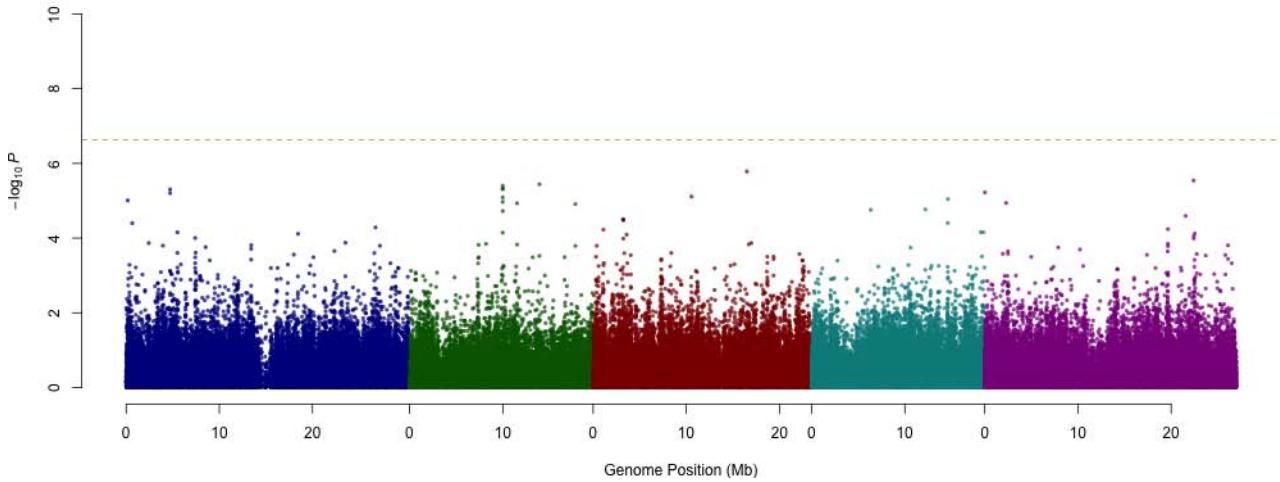
a: Phenotypic and *p*-value distributions.

Top-left: phenotypic distribution; Top-right: -$\log_{10}p$-values after genomic control (GC) against minor allele frequencies (MAF); Bottom panels: Quantile-quantile plots of *p*-values and -$\log_{10}p$-values before (blue) and after (green) GC.



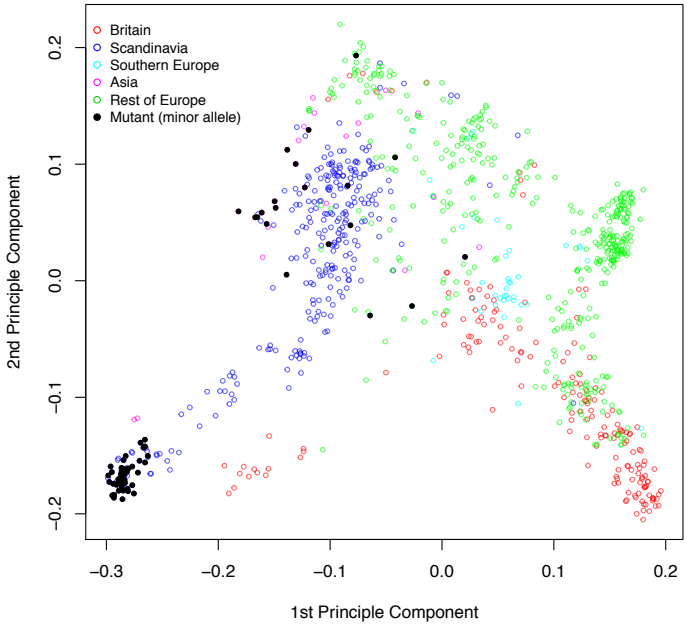
b: Genome-wide association mapping for climate adaptability.

The plotted -$\log_{10}p$-values are genomic controlled. Markers with minor allele frequencies less than 5% are removed. Chromosomes are distinguished by colors. The Bonferroni-corrected significance threshold is marked by the horizontal line.



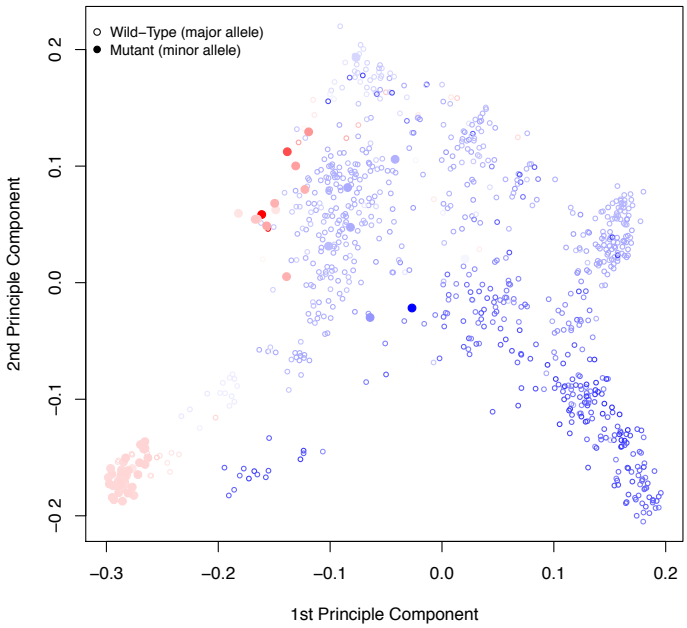
Supplementary Figure 13 - Summary of results for aridity index.

Genomic kinship principle components categorized based on geographical regions.



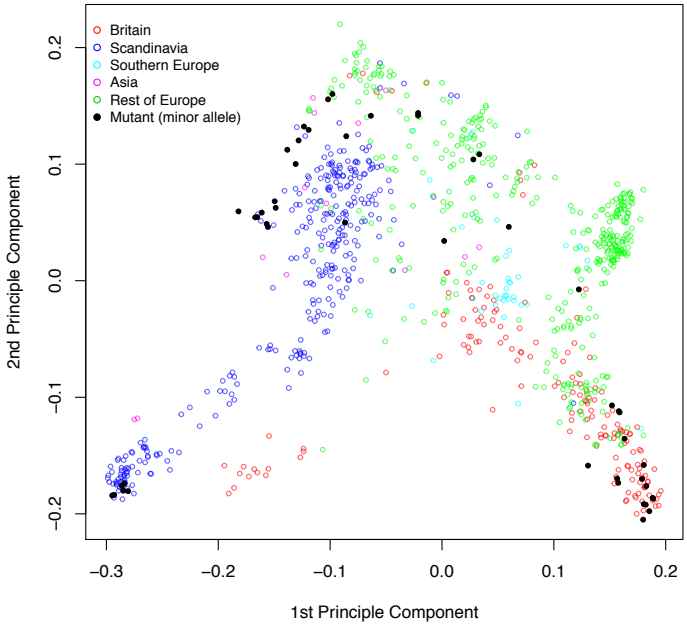
Genomic kinship principle components colored based on the scale of the climate variable.

The colors scale from pure blue (the minimum climate variable value) to pure red (the maximum value).



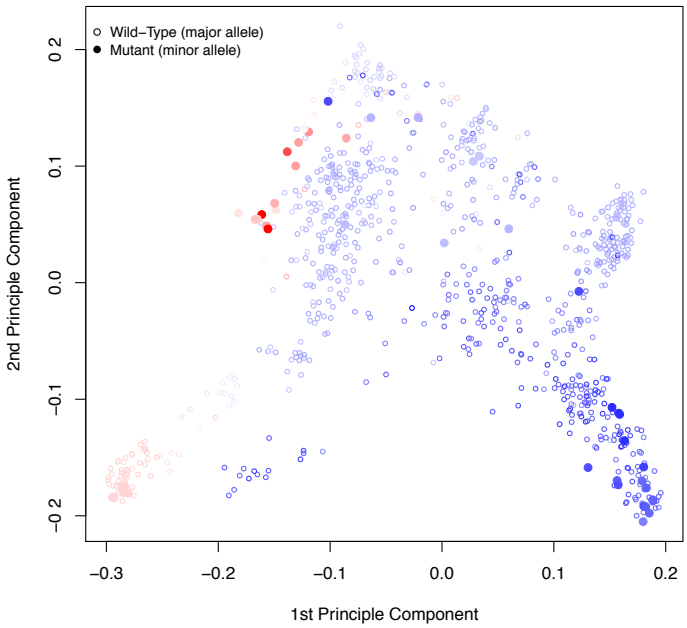
Supplementary Figure 14 - Principle components of the genomic kinship for the two alleles on chromosome 2 at 12169701 bp. Corresponding climate variable: temperature seasonality.

Genomic kinship principle components categorized based on geographical regions.



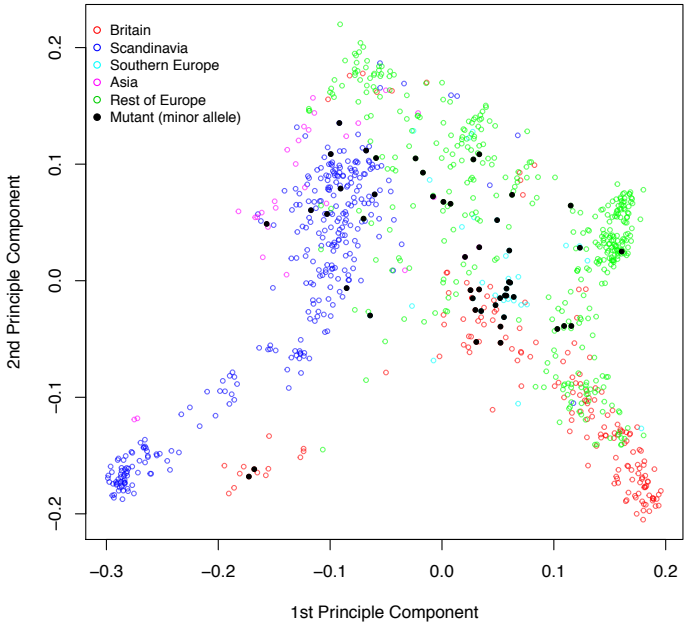
Genomic kinship principle components colored based on the scale of the climate variable.

The colors scale from pure blue (the minimum climate variable value) to pure red (the maximum value).



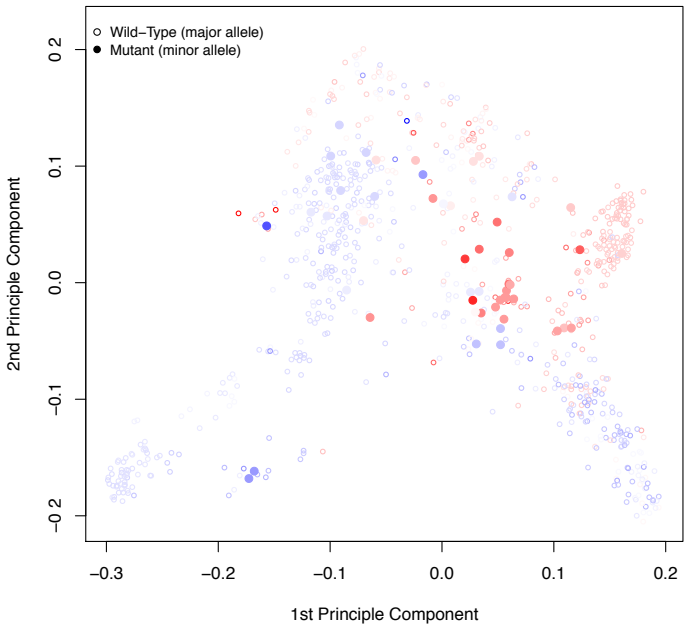
Supplementary Figure 15 - Principle components of the genomic kinship for the two alleles on chromosome 4 at 10406018 bp. Corresponding climate variable: temperature seasonality.

Genomic kinship principle components categorized based on geographical regions.



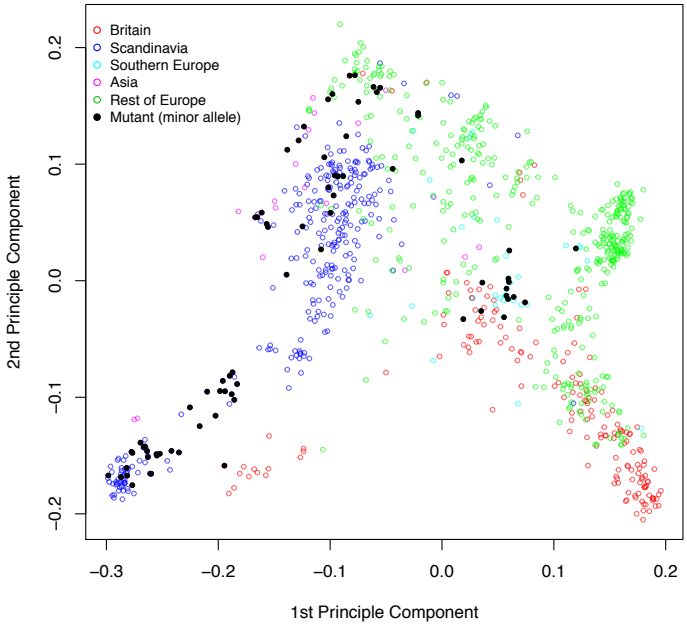
Genomic kinship principle components colored based on the scale of the climate variable.

The colors scale from pure blue (the minimum climate variable value) to pure red (the maximum value).



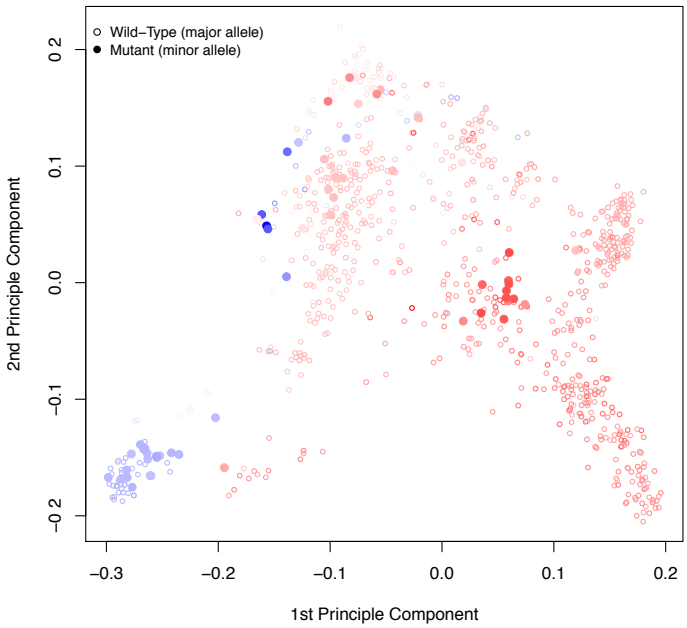
Supplementary Figure 16 - Principle components of the genomic kinship for the two alleles on chromosome 1 at 6936457 bp. Corresponding climate variable: maximum temperature in the warmest month.

Genomic kinship principle components categorized based on geographical regions.



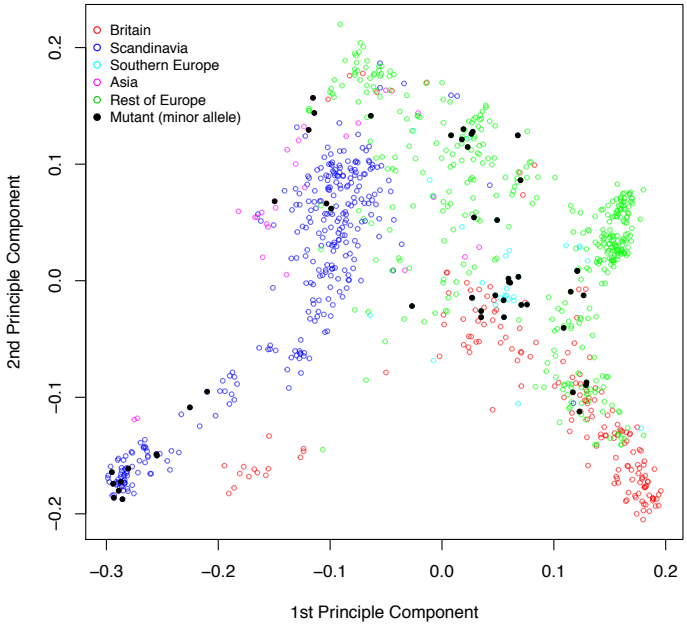
Genomic kinship principle components colored based on the scale of the climate variable.

The colors scale from pure blue (the minimum climate variable value) to pure red (the maximum value).



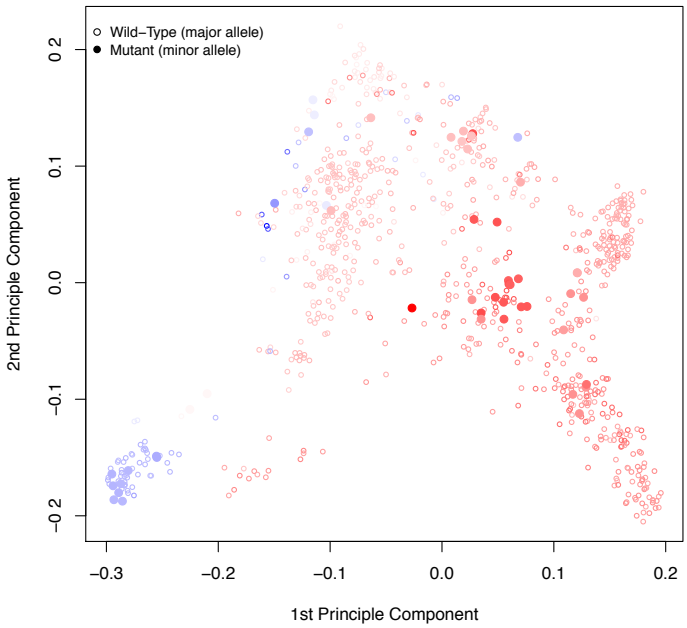
Supplementary Figure 17 - Principle components of the genomic kinship for the two alleles on chromosome 2 at 18620697 bp. Corresponding climate variable: minimum temperature in the coldest month.

Genomic kinship principle components categorized based on geographical regions.



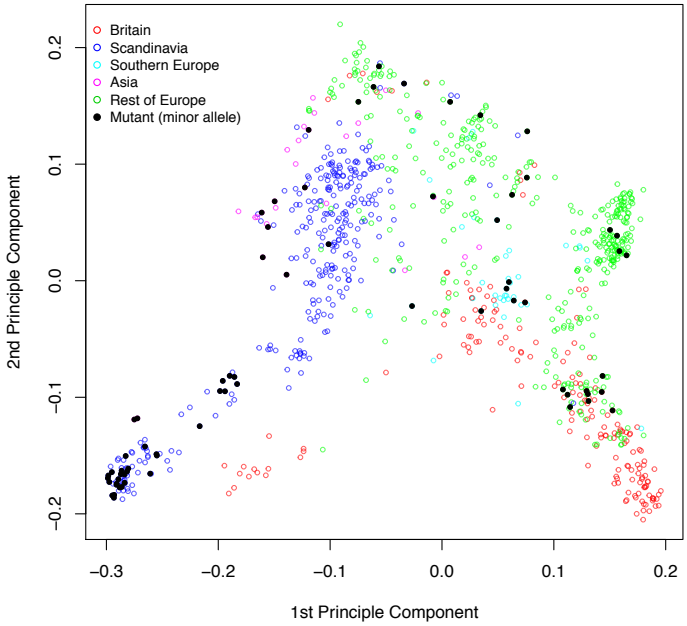
Genomic kinship principle components colored based on the scale of the climate variable.

The colors scale from pure blue (the minimum climate variable value) to pure red (the maximum value).



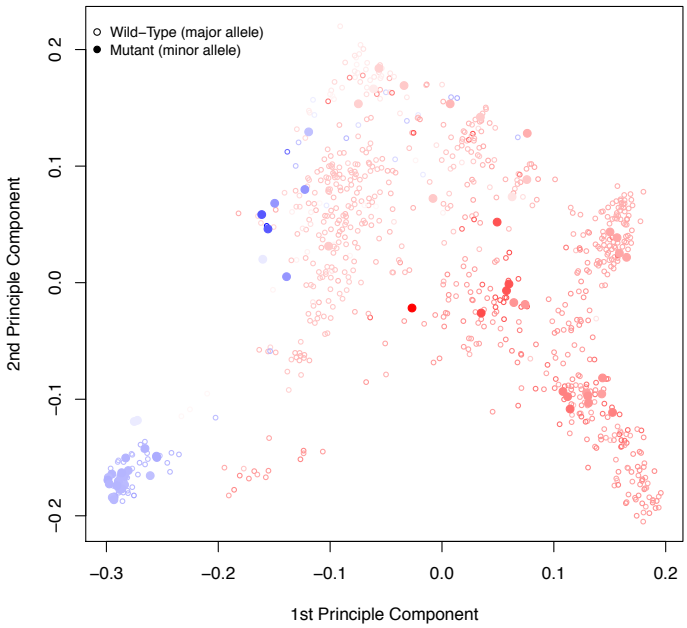
Supplementary Figure 18 - Principle components of the genomic kinship for the two alleles on chromosome 2 at 19397389 bp. Corresponding climate variable: minimum temperature in the coldest month.

Genomic kinship principle components categorized based on geographical regions.



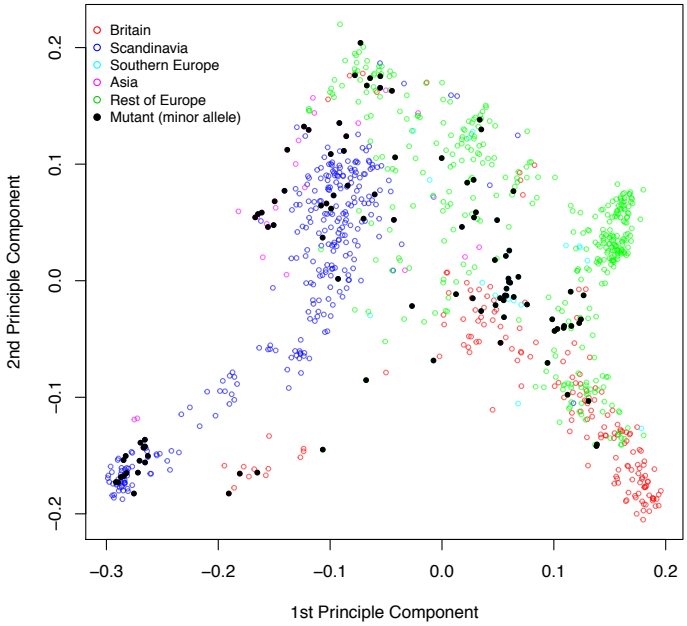
Genomic kinship principle components colored based on the scale of the climate variable.

The colors scale from pure blue (the minimum climate variable value) to pure red (the maximum value).



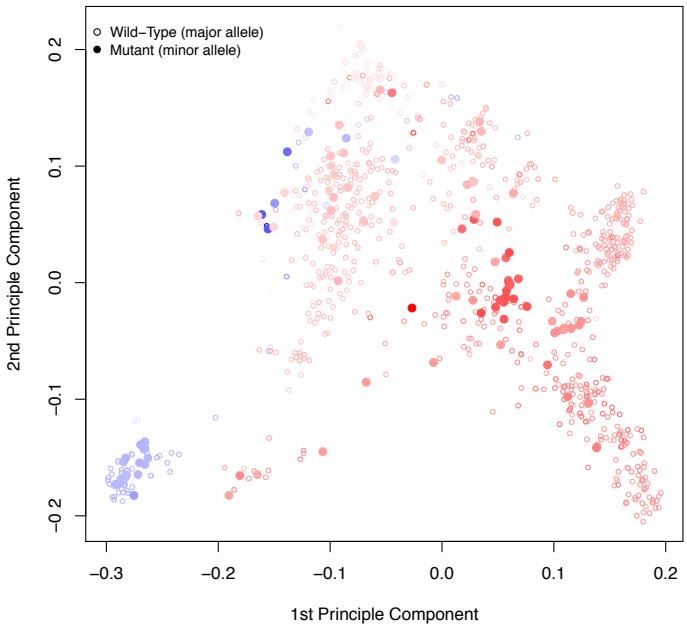
Supplementary Figure 19 - Principle components of the genomic kinship for the two alleles on chromosome 5 at 14067526 bp. Corresponding climate variable: minimum temperature in the coldest month.

Genomic kinship principle components categorized based on geographical regions.



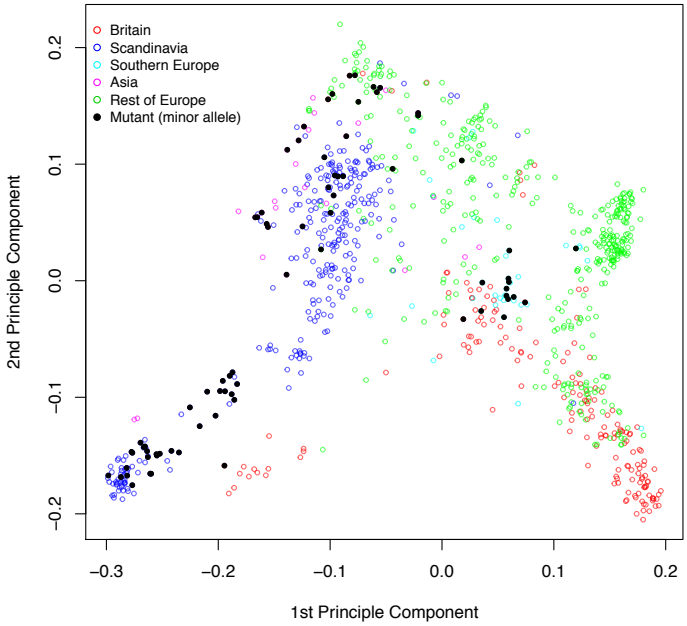
Genomic kinship principle components colored based on the scale of the climate variable.

The colors scale from pure blue (the minimum climate variable value) to pure red (the maximum value).



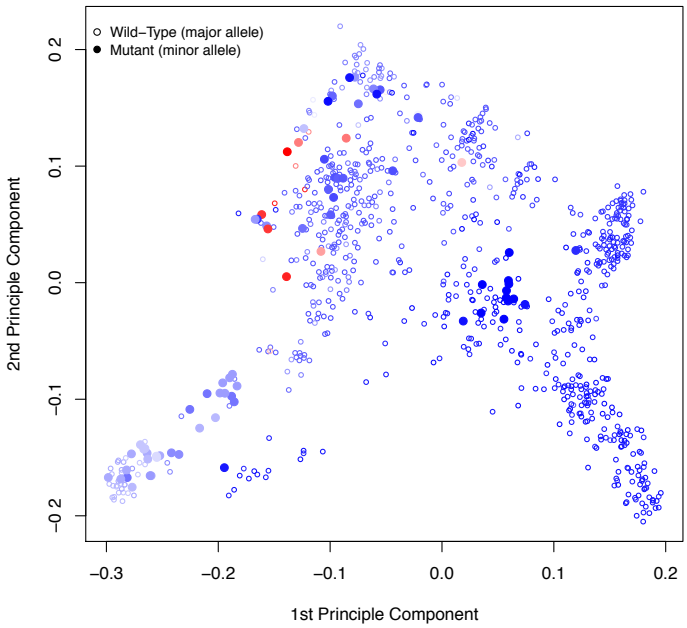
Supplementary Figure 20 - Principle components of the genomic kinship for the two alleles on chromosome 5 at 18397418 bp. Corresponding climate variable: minimum temperature in the coldest month.

Genomic kinship principle components categorized based on geographical regions.



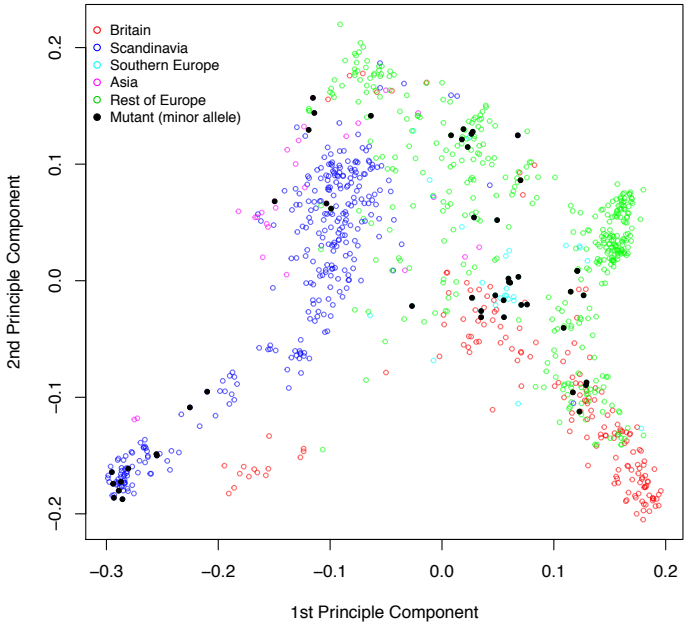
Genomic kinship principle components colored based on the scale of the climate variable.

The colors scale from pure blue (the minimum climate variable value) to pure red (the maximum value).



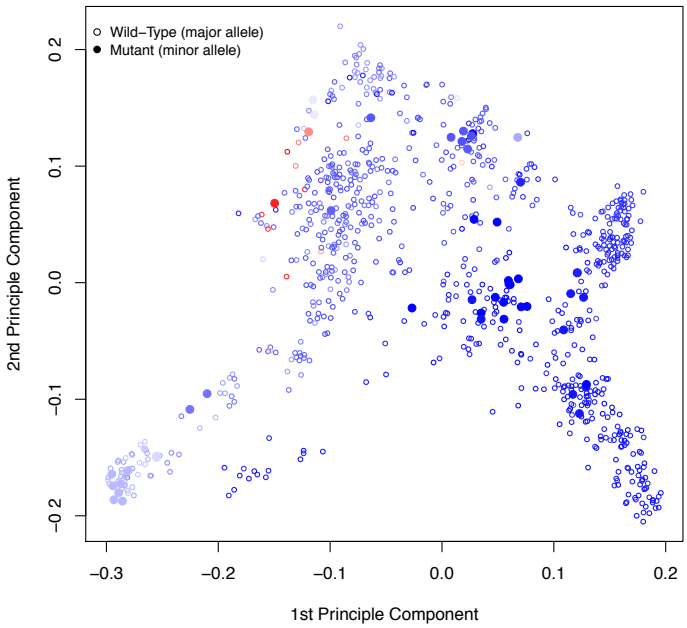
Supplementary Figure 21 - Principle components of the genomic kinship for the two alleles on chromosome 2 at 18620697 bp. Corresponding climate variable: number of consecutive cold days.

Genomic kinship principle components categorized based on geographical regions.



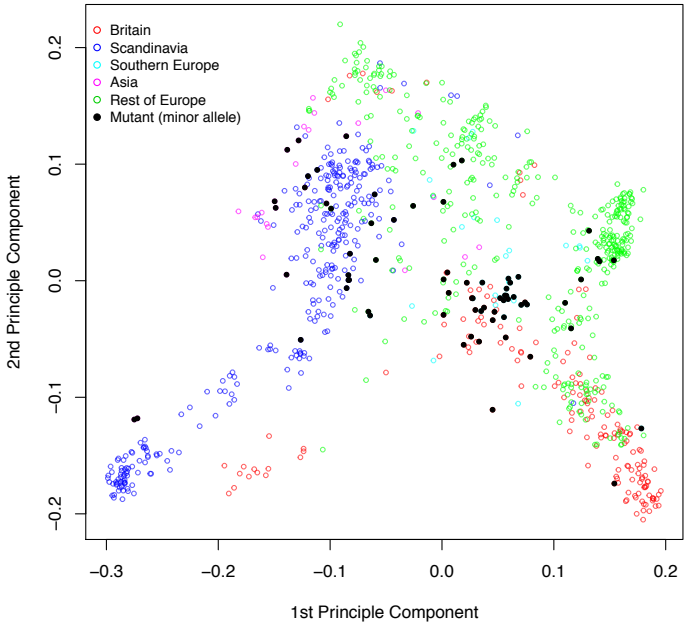
Genomic kinship principle components colored based on the scale of the climate variable.

The colors scale from pure blue (the minimum climate variable value) to pure red (the maximum value).



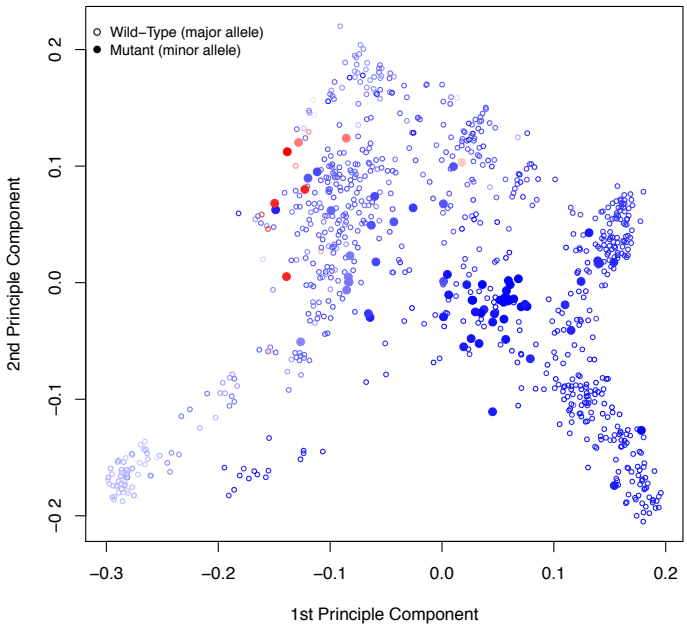
Supplementary Figure 22 - Principle components of the genomic kinship for the two alleles on chromosome 2 at 19397389 bp. Corresponding climate variable: number of consecutive cold days.

Genomic kinship principle components categorized based on geographical regions.



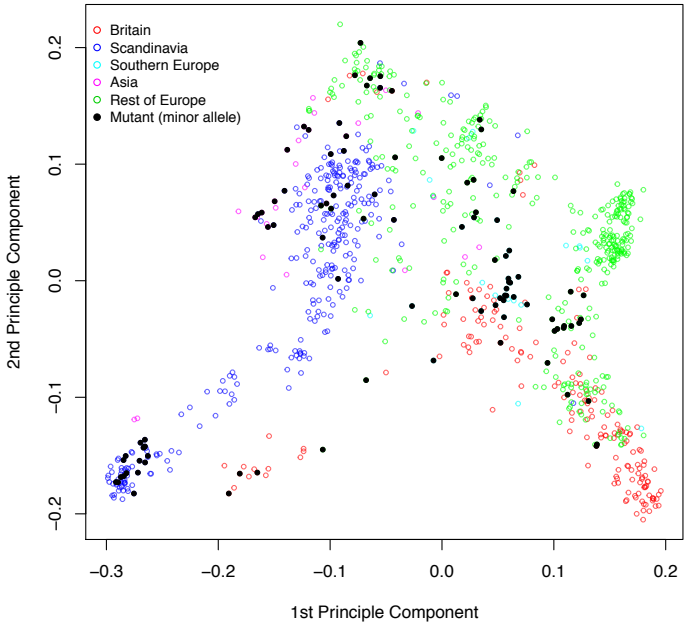
Genomic kinship principle components colored based on the scale of the climate variable.

The colors scale from pure blue (the minimum climate variable value) to pure red (the maximum value).



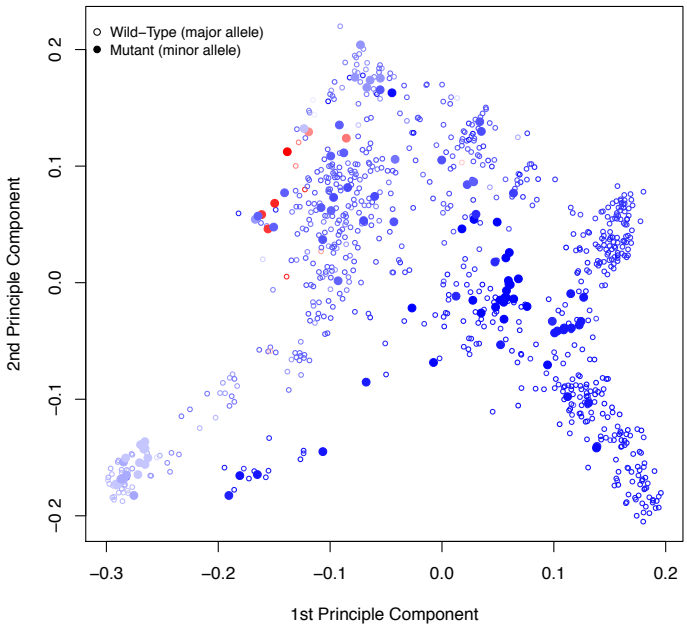
Supplementary Figure 23 - Principle components of the genomic kinship for the two alleles on chromosome 5 at 7492277 bp. Corresponding climate variable: number of consecutive cold days.

Genomic kinship principle components categorized based on geographical regions.



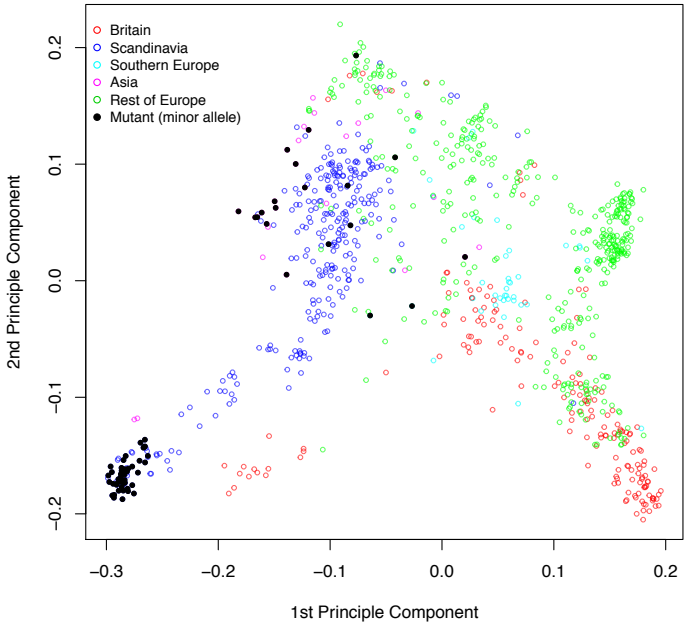
Genomic kinship principle components colored based on the scale of the climate variable.

The colors scale from pure blue (the minimum climate variable value) to pure red (the maximum value).



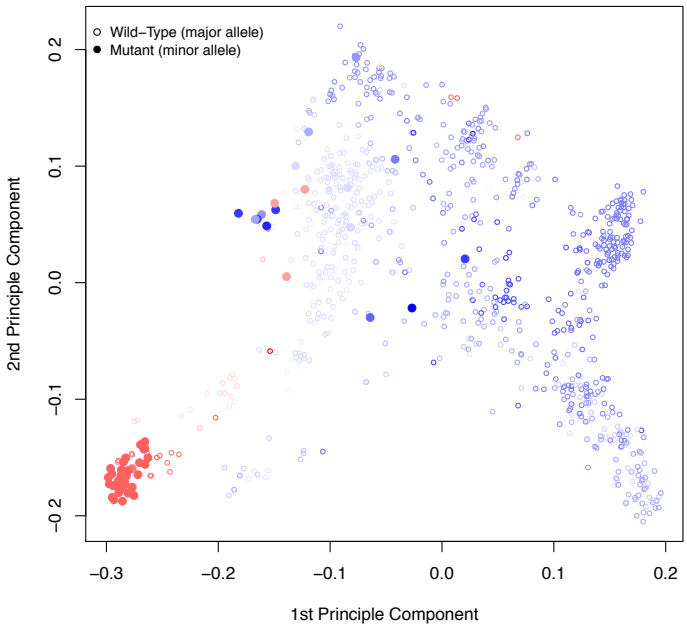
Supplementary Figure 24 - Principle components of the genomic kinship for the two alleles on chromosome 5 at 18397418 bp. Corresponding climate variable: number of consecutive cold days.

Genomic kinship principle components categorized based on geographical regions.



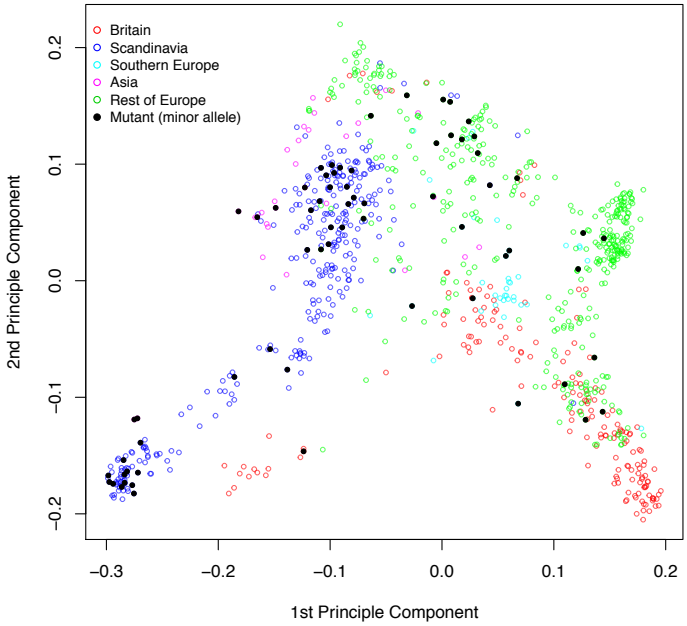
Genomic kinship principle components colored based on the scale of the climate variable.

The colors scale from pure blue (the minimum climate variable value) to pure red (the maximum value).



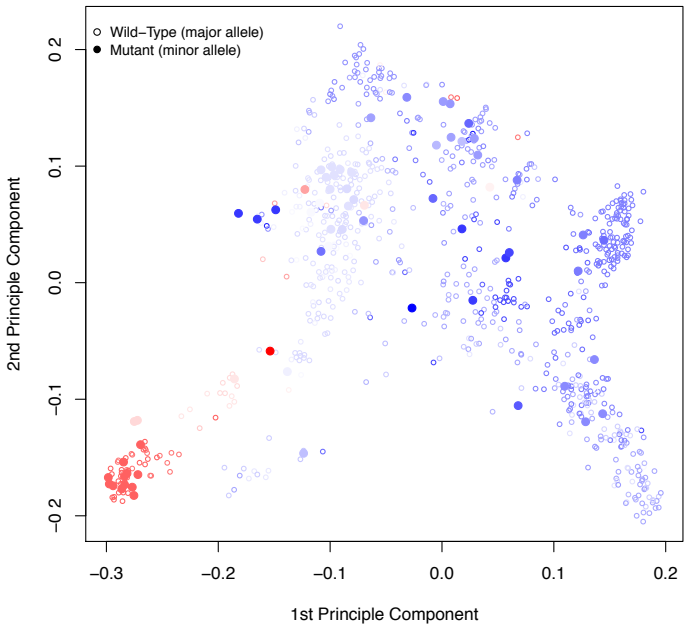
Supplementary Figure 25 - Principle components of the genomic kinship for the two alleles on chromosome 2 at 12169701 bp. Corresponding climate variable: day length in spring.

Genomic kinship principle components categorized based on geographical regions.



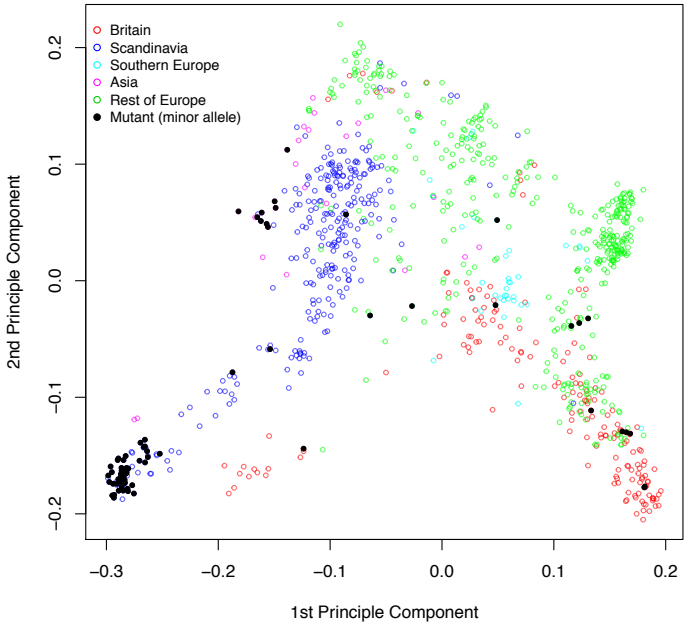
Genomic kinship principle components colored based on the scale of the climate variable.

The colors scale from pure blue (the minimum climate variable value) to pure red (the maximum value).



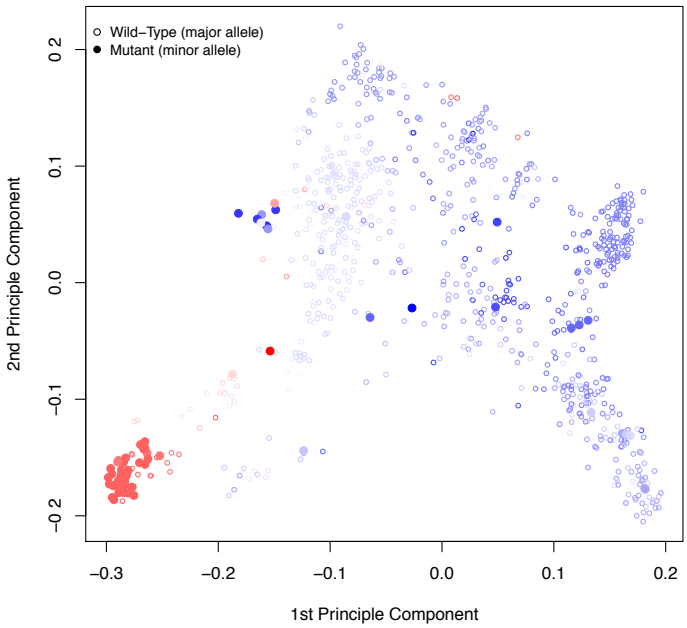
Supplementary Figure 26 - Principle components of the genomic kinship for the two alleles on chromosome 3 at 12642006 bp. Corresponding climate variable: day length in spring.

Genomic kinship principle components categorized based on geographical regions.



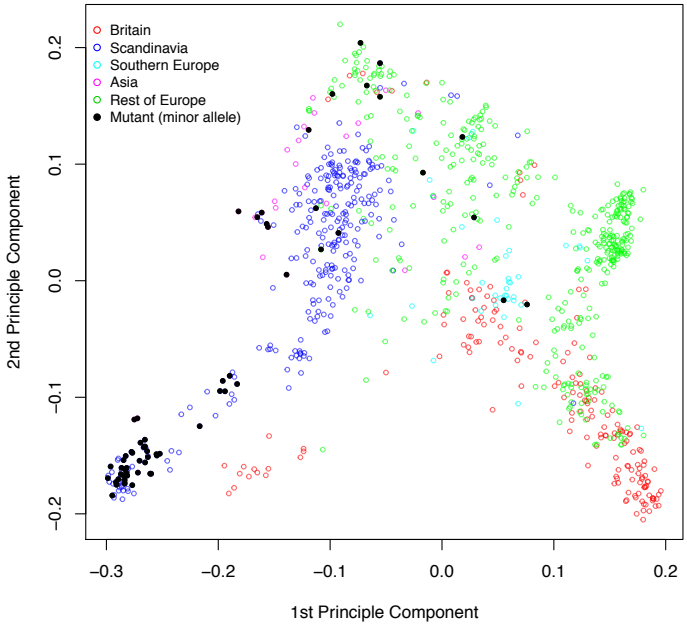
Genomic kinship principle components colored based on the scale of the climate variable.

The colors scale from pure blue (the minimum climate variable value) to pure red (the maximum value).



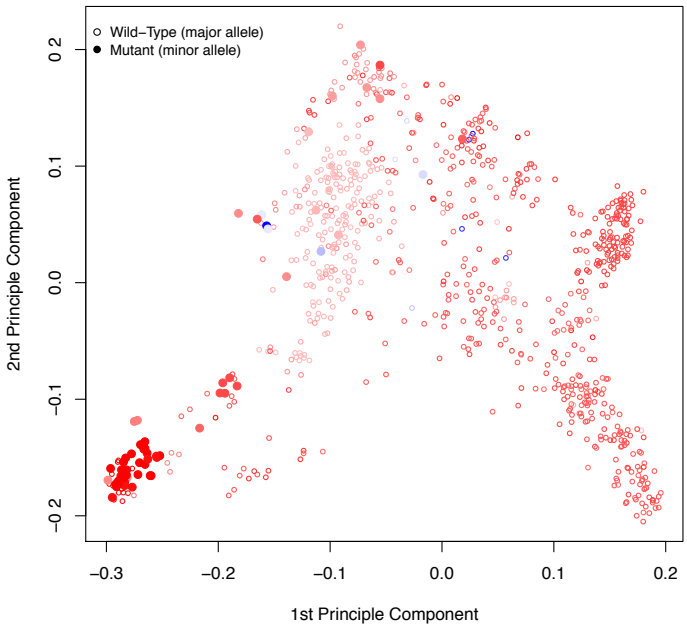
Supplementary Figure 27 - Principle components of the genomic kinship for the two alleles on chromosome 4 at 14788320 bp. Corresponding climate variable: day length in spring.

Genomic kinship principle components categorized based on geographical regions.



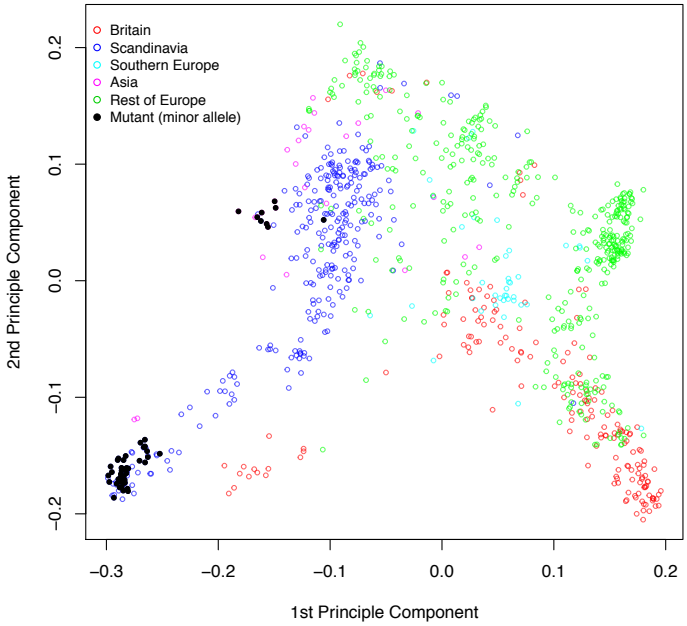
Genomic kinship principle components colored based on the scale of the climate variable.

The colors scale from pure blue (the minimum climate variable value) to pure red (the maximum value).



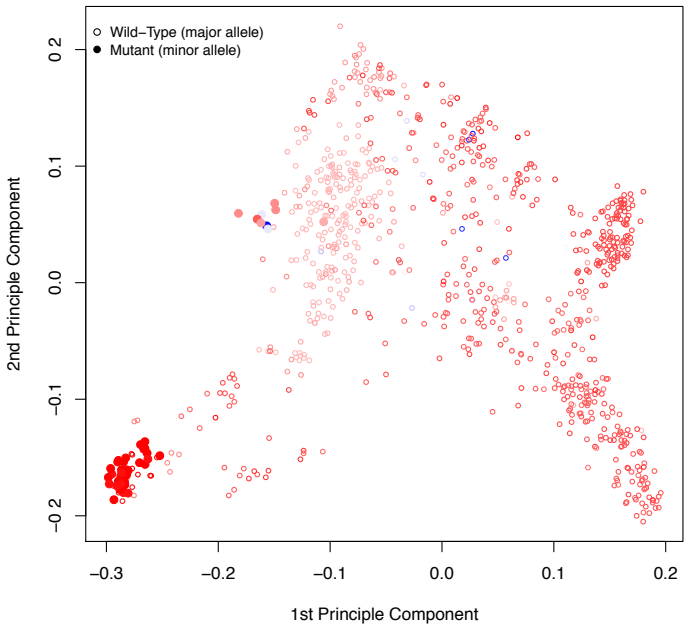
Supplementary Figure 28 - Principle components of the genomic kinship for the two alleles on chromosome 3 at 1816353 bp. Corresponding climate variable: relative humidity in spring.

Genomic kinship principle components categorized based on geographical regions.



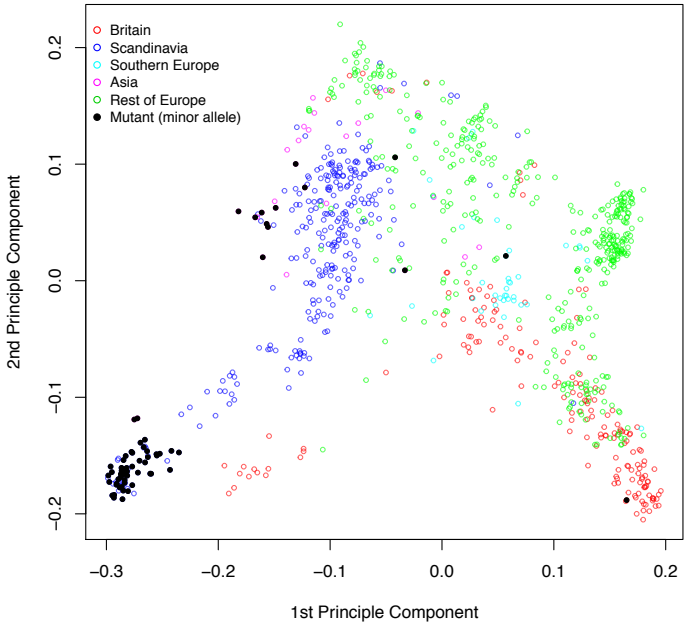
Genomic kinship principle components colored based on the scale of the climate variable.

The colors scale from pure blue (the minimum climate variable value) to pure red (the maximum value).



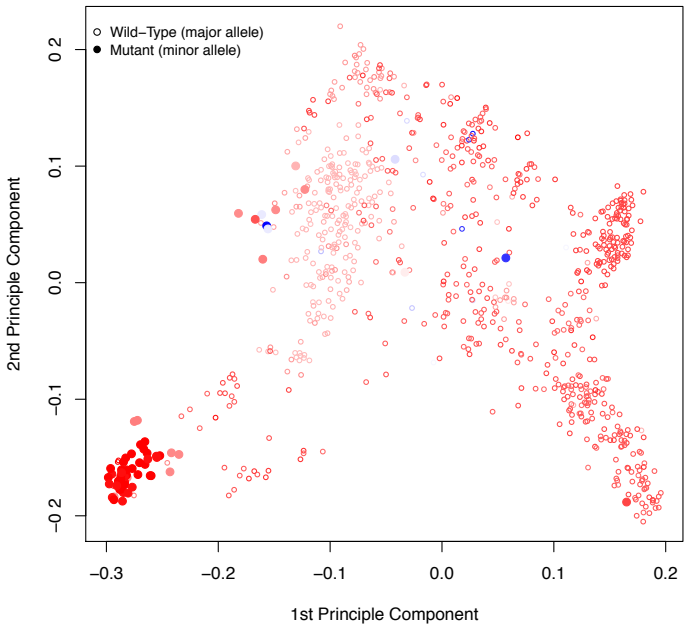
Supplementary Figure 29 - Principle components of the genomic kinship for the two alleles on chromosome 4 at 14834441 bp. Corresponding climate variable: relative humidity in spring.

Genomic kinship principle components categorized based on geographical regions.



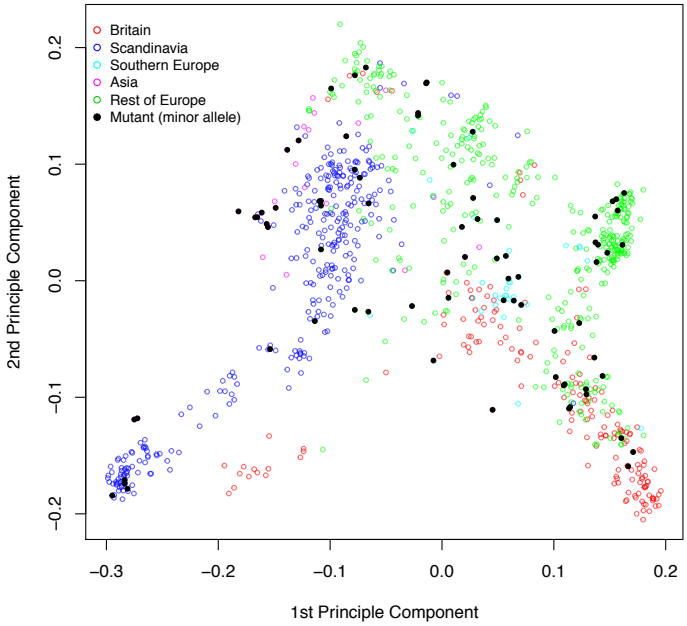
Genomic kinship principle components colored based on the scale of the climate variable.

The colors scale from pure blue (the minimum climate variable value) to pure red (the maximum value).



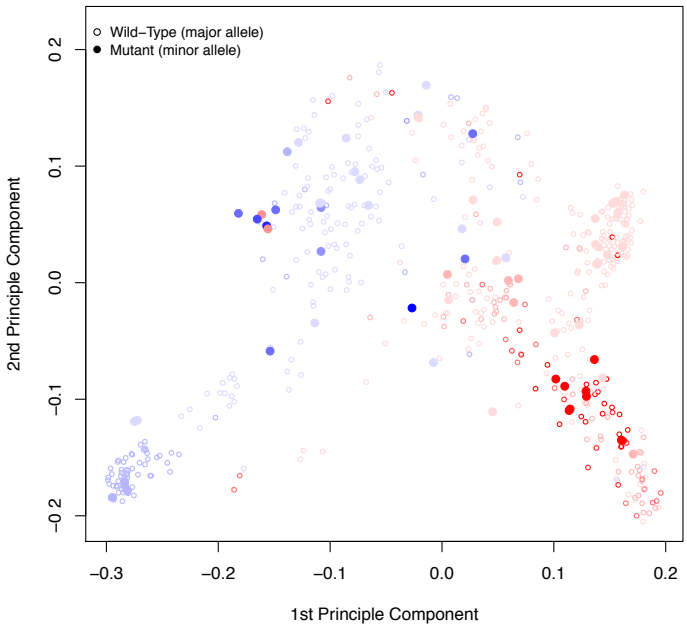
Supplementary Figure 30 - Principle components of the genomic kinship for the two alleles on chromosome 5 at 8380640 bp. Corresponding climate variable: relative humidity in spring.

Genomic kinship principle components categorized based on geographical regions.



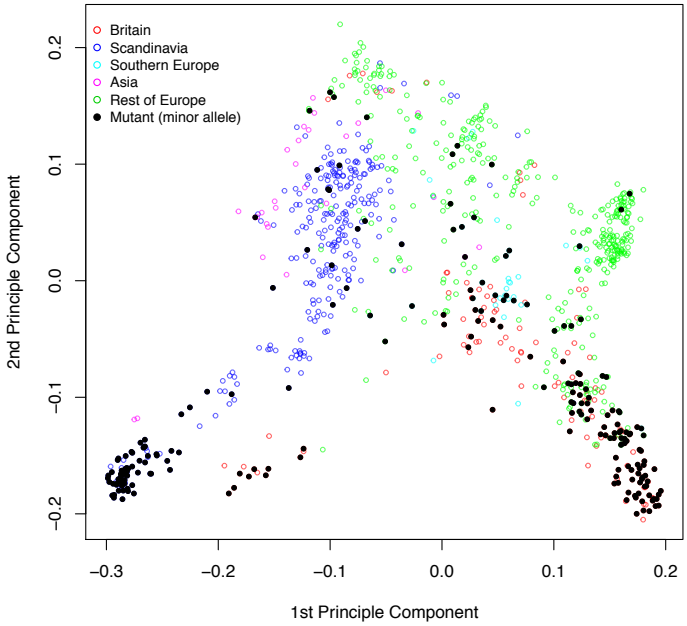
Genomic kinship principle components colored based on the scale of the climate variable.

The colors scale from pure blue (the minimum climate variable value) to pure red (the maximum value).



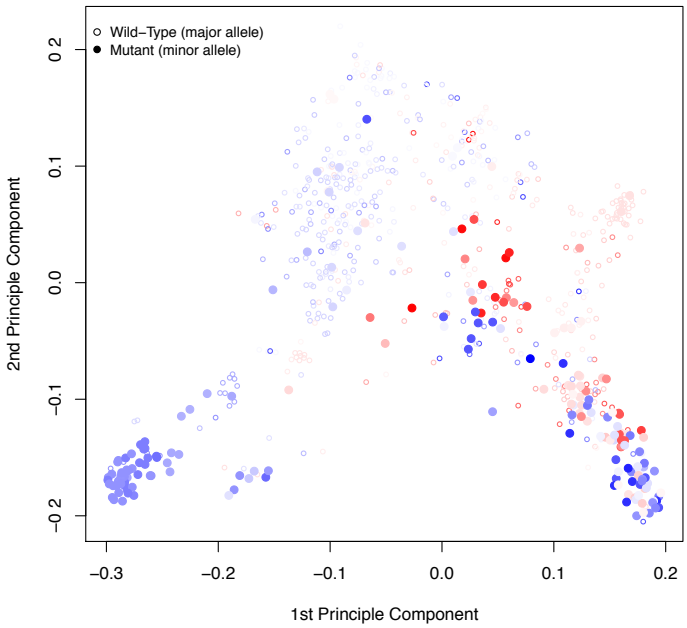
Supplementary Figure 31 - Principle components of the genomic kinship for the two alleles on chromosome 3 at 576148 bp. Corresponding climate variable: length of the growing season.

Genomic kinship principle components categorized based on geographical regions.



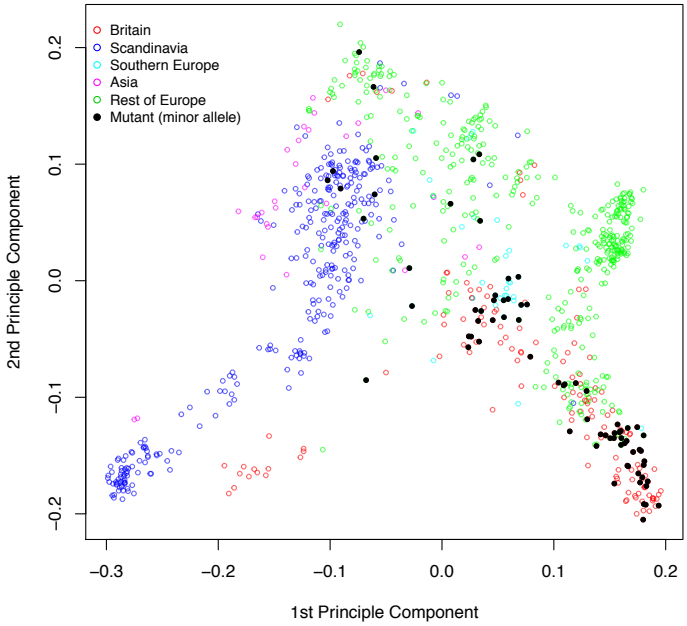
Genomic kinship principle components colored based on the scale of the climate variable.

The colors scale from pure blue (the minimum climate variable value) to pure red (the maximum value).



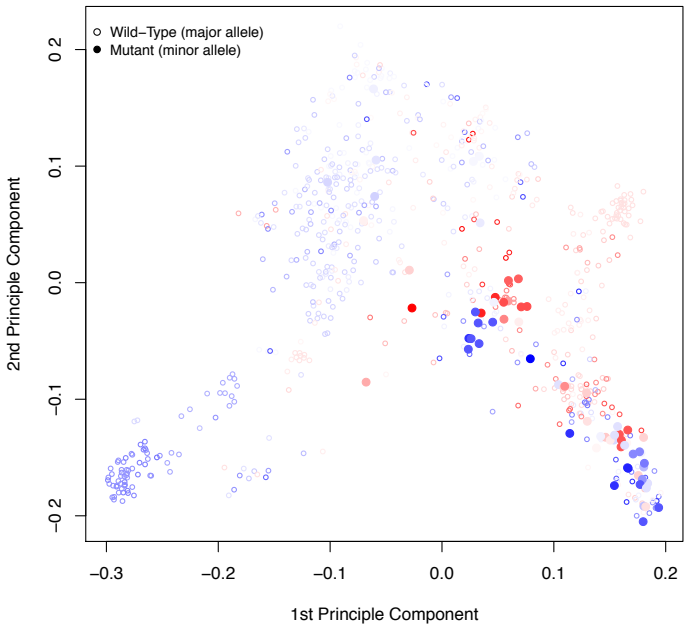
Supplementary Figure 32 - Principle components of the genomic kinship for the two alleles on chromosome 1 at 953031 bp. Corresponding climate variable: number of consecutive frost-free days.

Genomic kinship principle components categorized based on geographical regions.



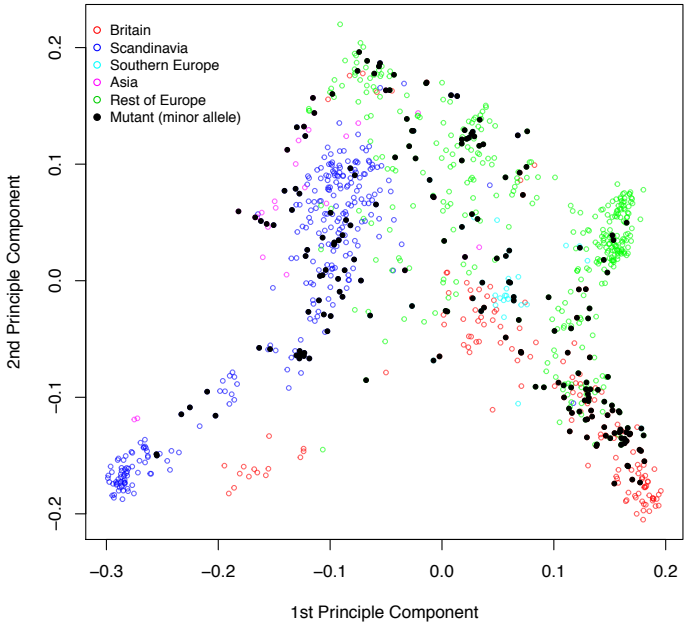
Genomic kinship principle components colored based on the scale of the climate variable.

The colors scale from pure blue (the minimum climate variable value) to pure red (the maximum value).



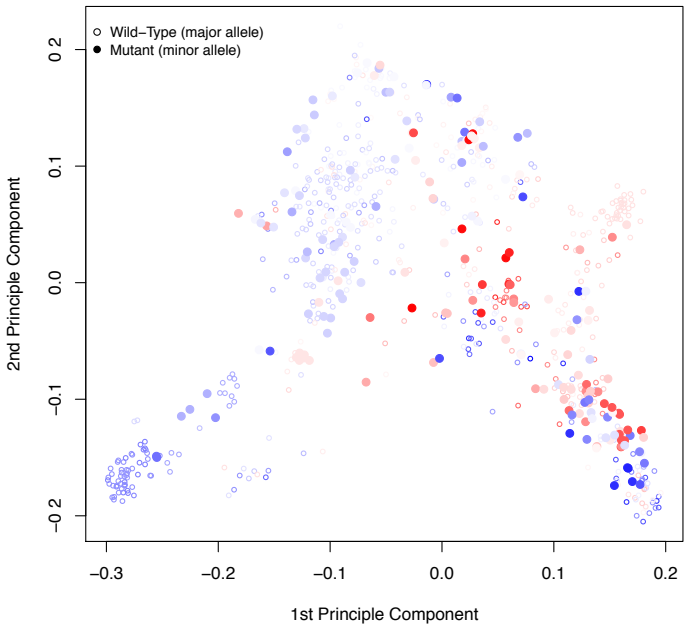
Supplementary Figure 33 - Principle components of the genomic kinship for the two alleles on chromosome 1 at 6463065 bp. Corresponding climate variable: number of consecutive frost-free days.

Genomic kinship principle components categorized based on geographical regions.



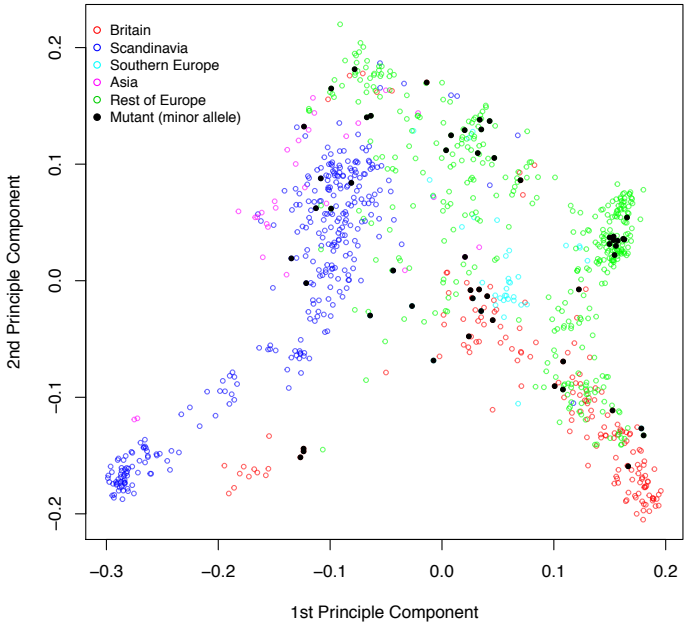
Genomic kinship principle components colored based on the scale of the climate variable.

The colors scale from pure blue (the minimum climate variable value) to pure red (the maximum value).



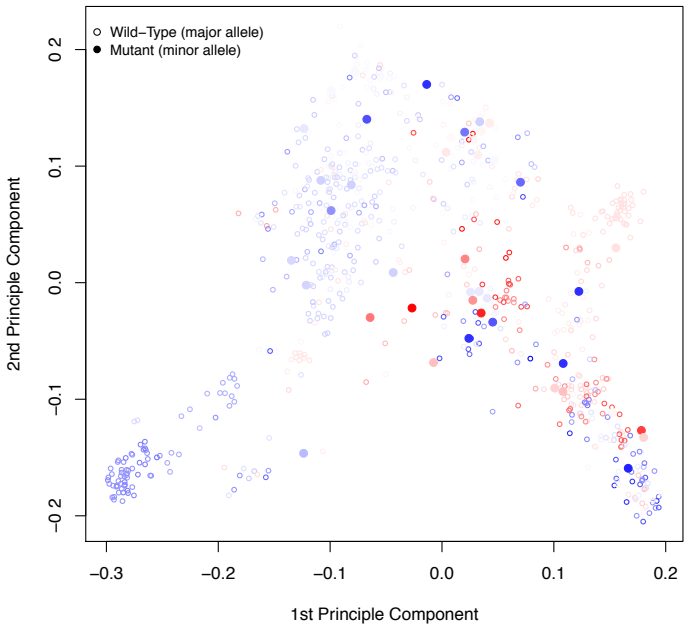
Supplementary Figure 34 - Principle components of the genomic kinship for the two alleles on chromosome 2 at 9904076 bp. Corresponding climate variable: number of consecutive frost-free days.

Genomic kinship principle components categorized based on geographical regions.

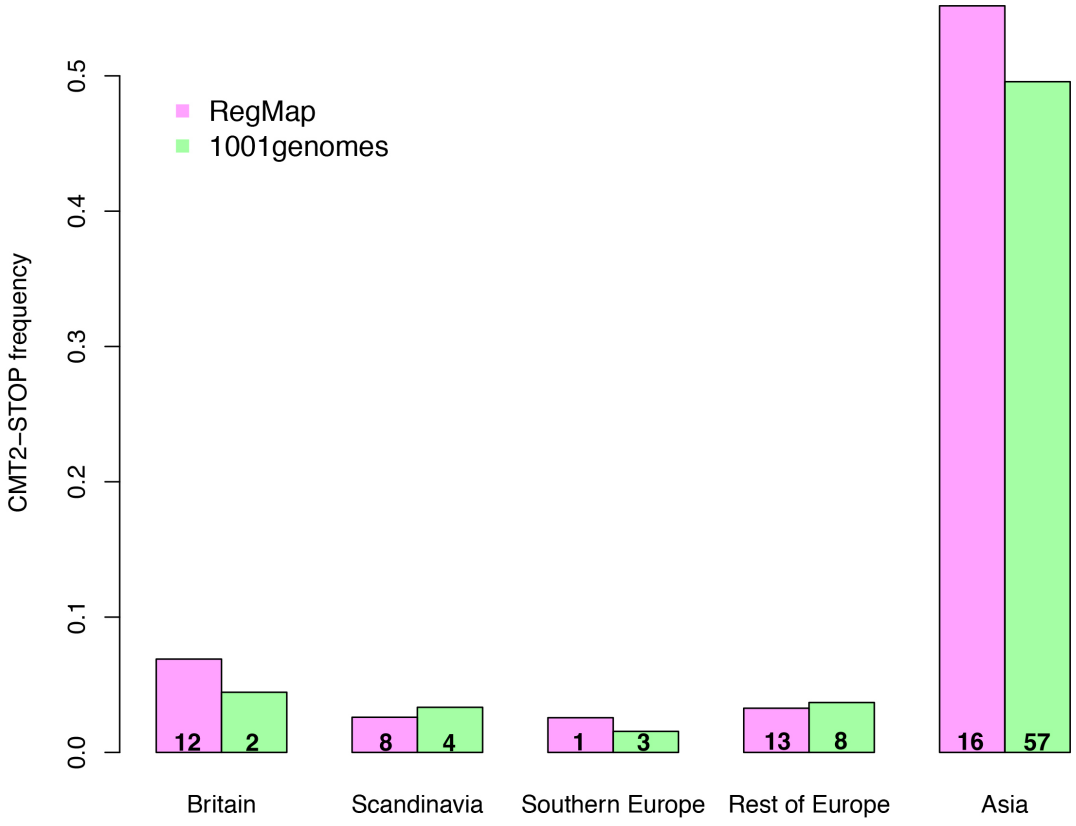


Genomic kinship principle components colored based on the scale of the climate variable.

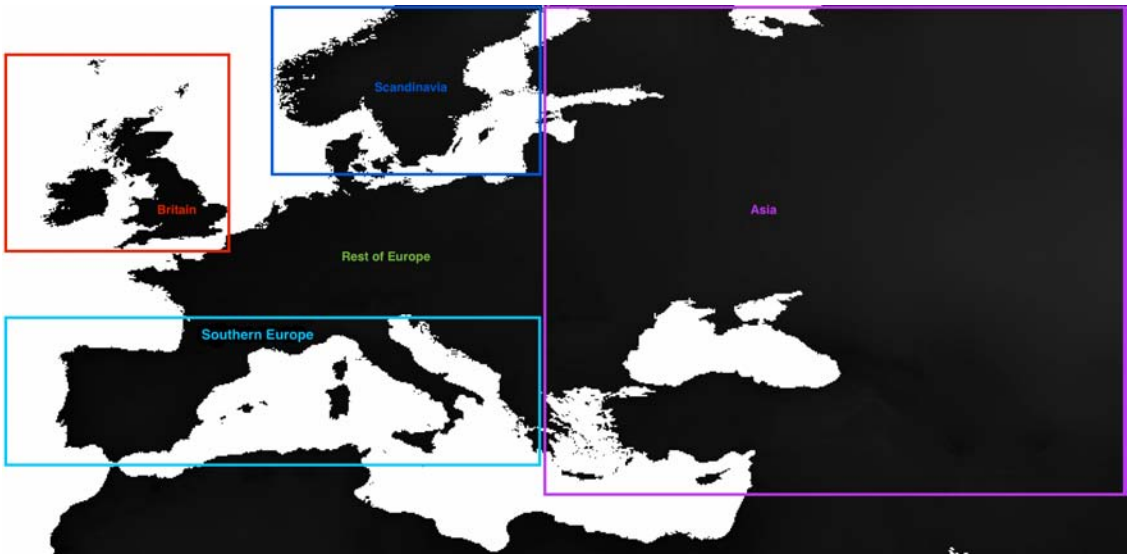
The colors scale from pure blue (the minimum climate variable value) to pure red (the maximum value).



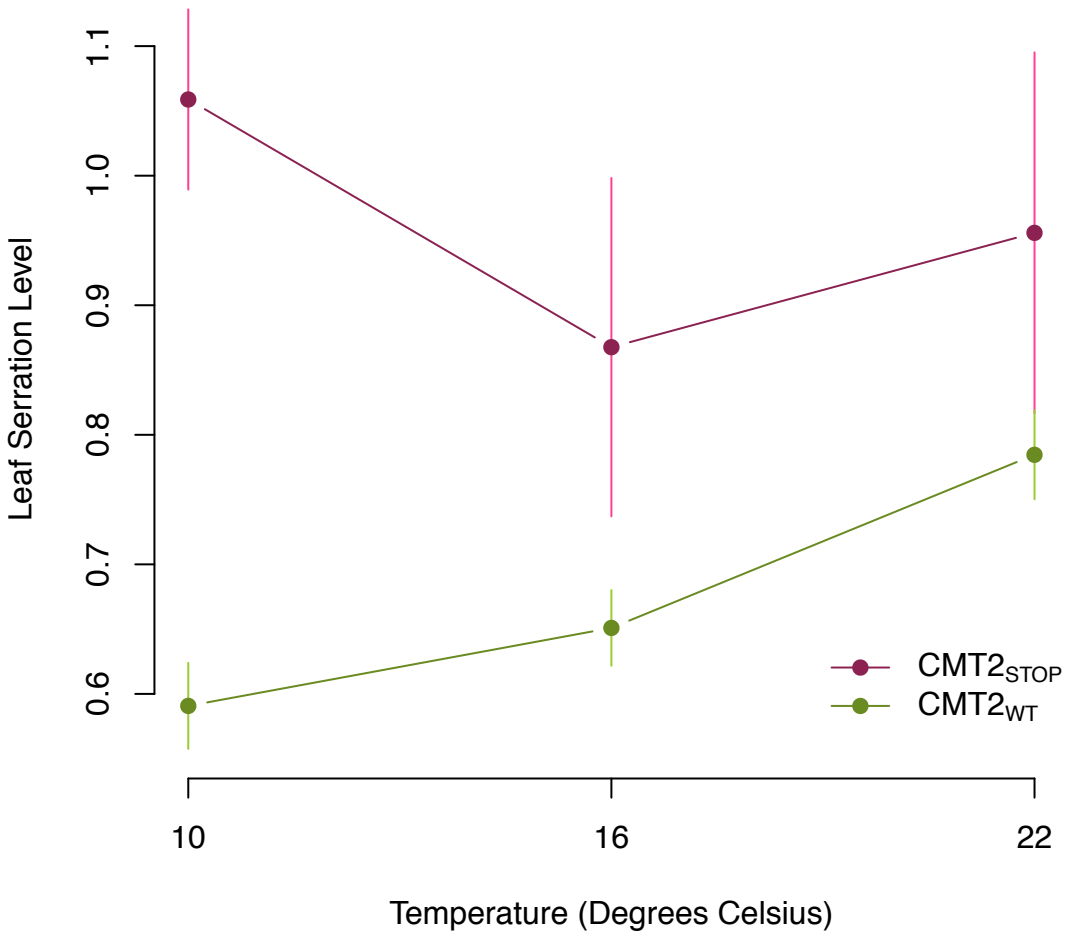
Supplementary Figure 35 - Principle components of the genomic kinship for the two alleles on chromosome 5 at 18061531 bp. Corresponding climate variable: number of consecutive frost-free days.



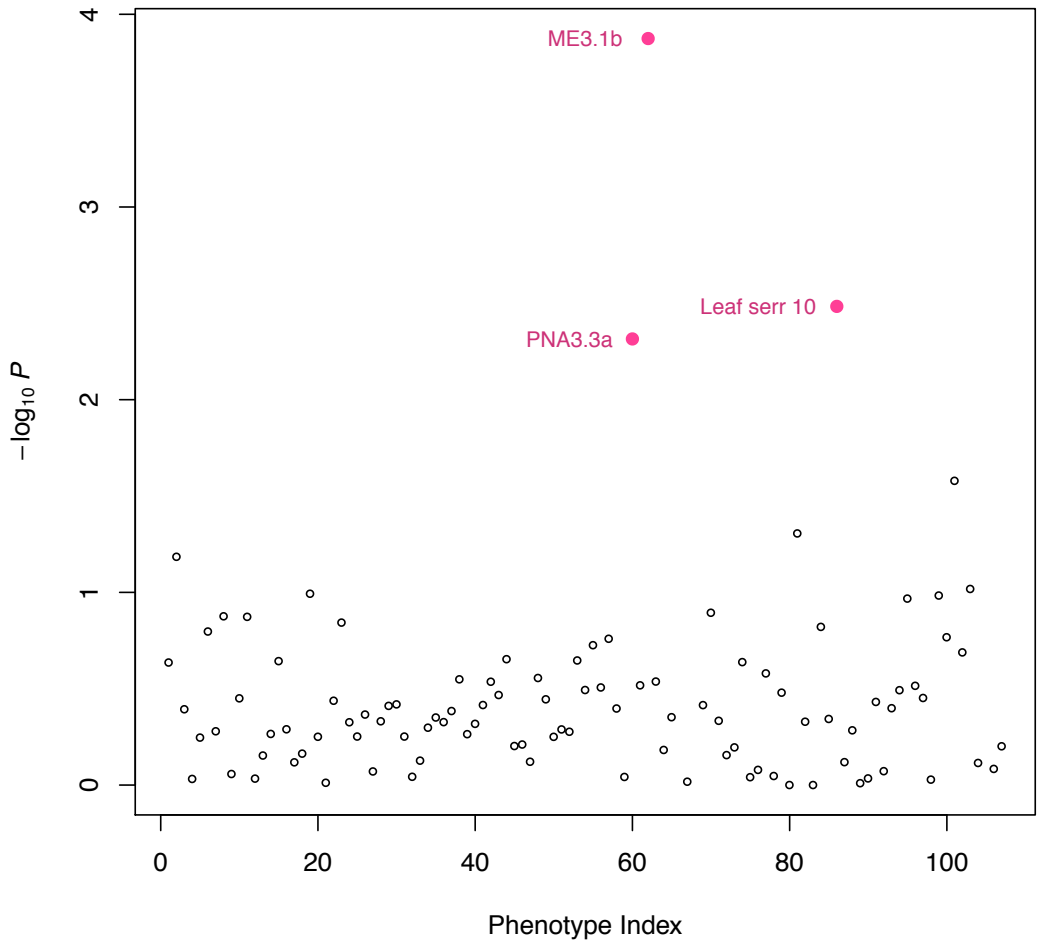
Supplementary Figure 36: Comparison between the RegMap and 1001genomes collections in terms of the allele-frequency of *CMT2$_{STOP}$* across different geographic regions in the Euroasian *A. thaliana* population. The numbers in the bars are the number of *CMT2$_{STOP}$* alleles in this area.



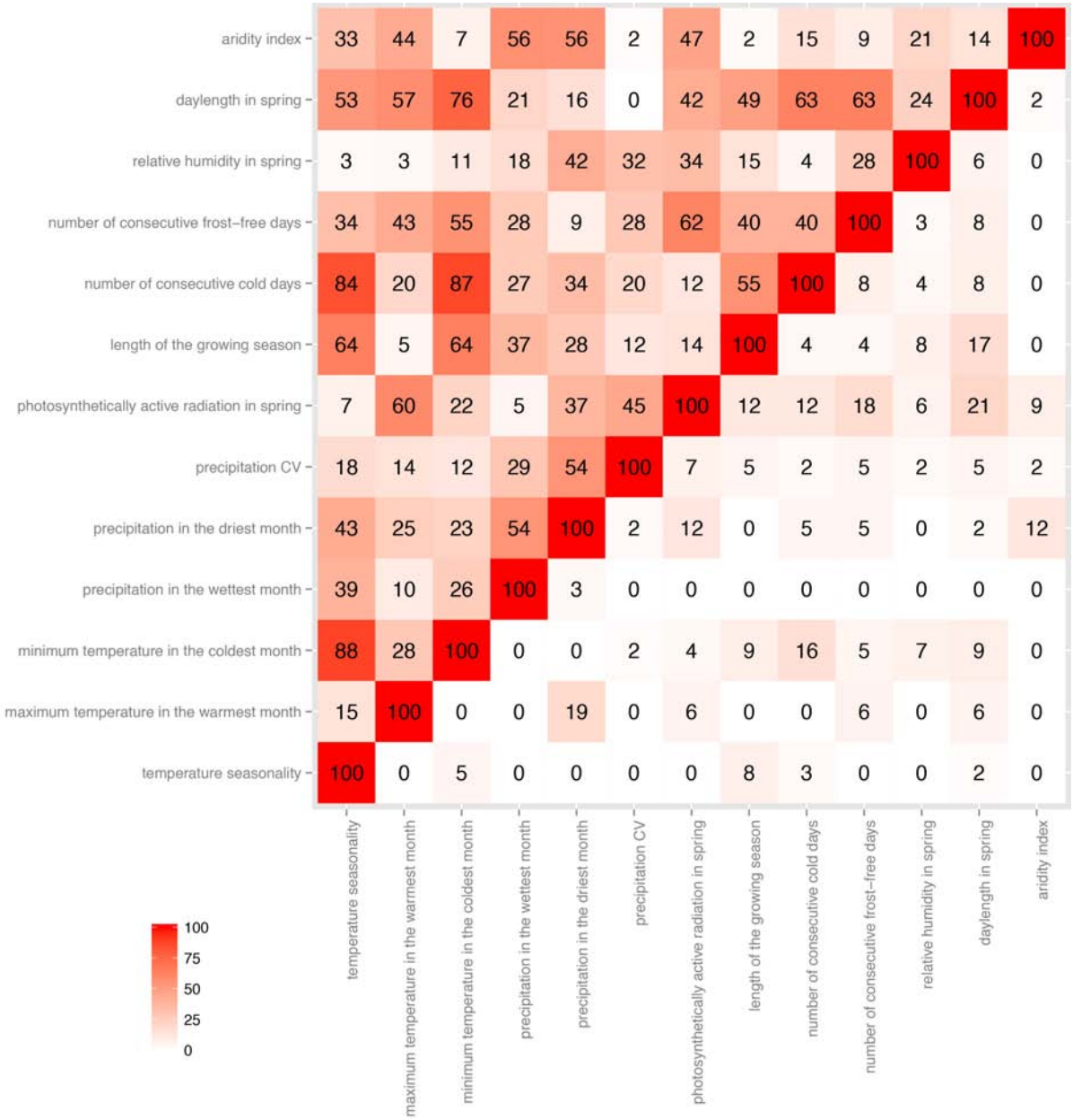
Supplementary Figure 37 - Defined geographical regions across the Euro-Asia sampling area.



Supplementary Figure 38: *CMT2*-by-temperature interaction effects on leaf serration. The analysis was performed using the genome-wide association data reported by Atwell et al. (2010). Each point is the mean leaf serration level of a combination of *CMT2* genotype and temperature. The vertical bars represent standard errors of the mean estimates.

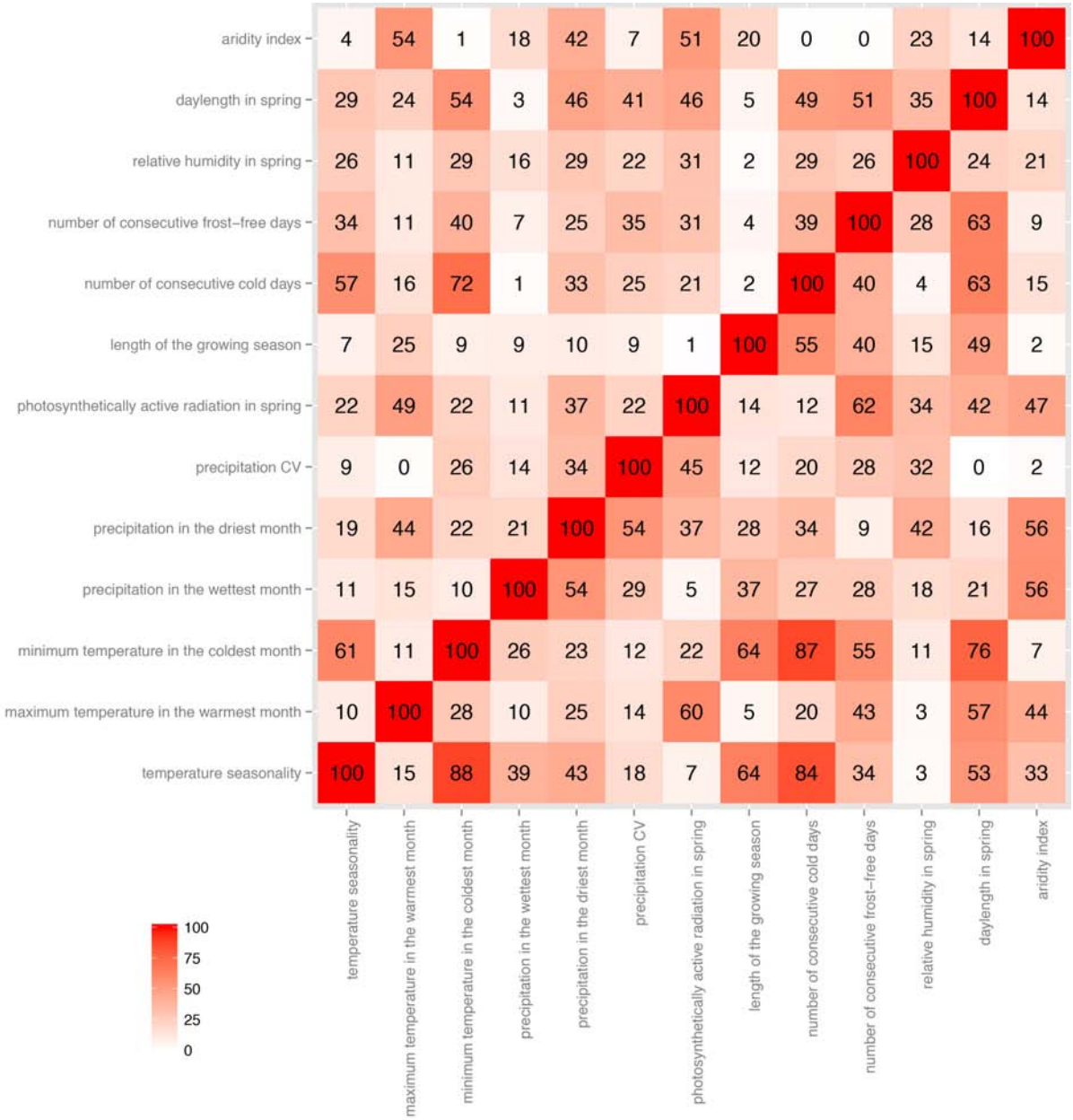


Supplementary Figure 39: Associations between the *CMT2$_{STOP}$* genotype and the 107 scored phenotypes in Atwell et al. (2010). The most significant three associations are labeled in pink, with a false discovery rate of 0.17. The definition of each labeled phenotype should be referred to the Supplementary Tables in Atwell et al. (2010).



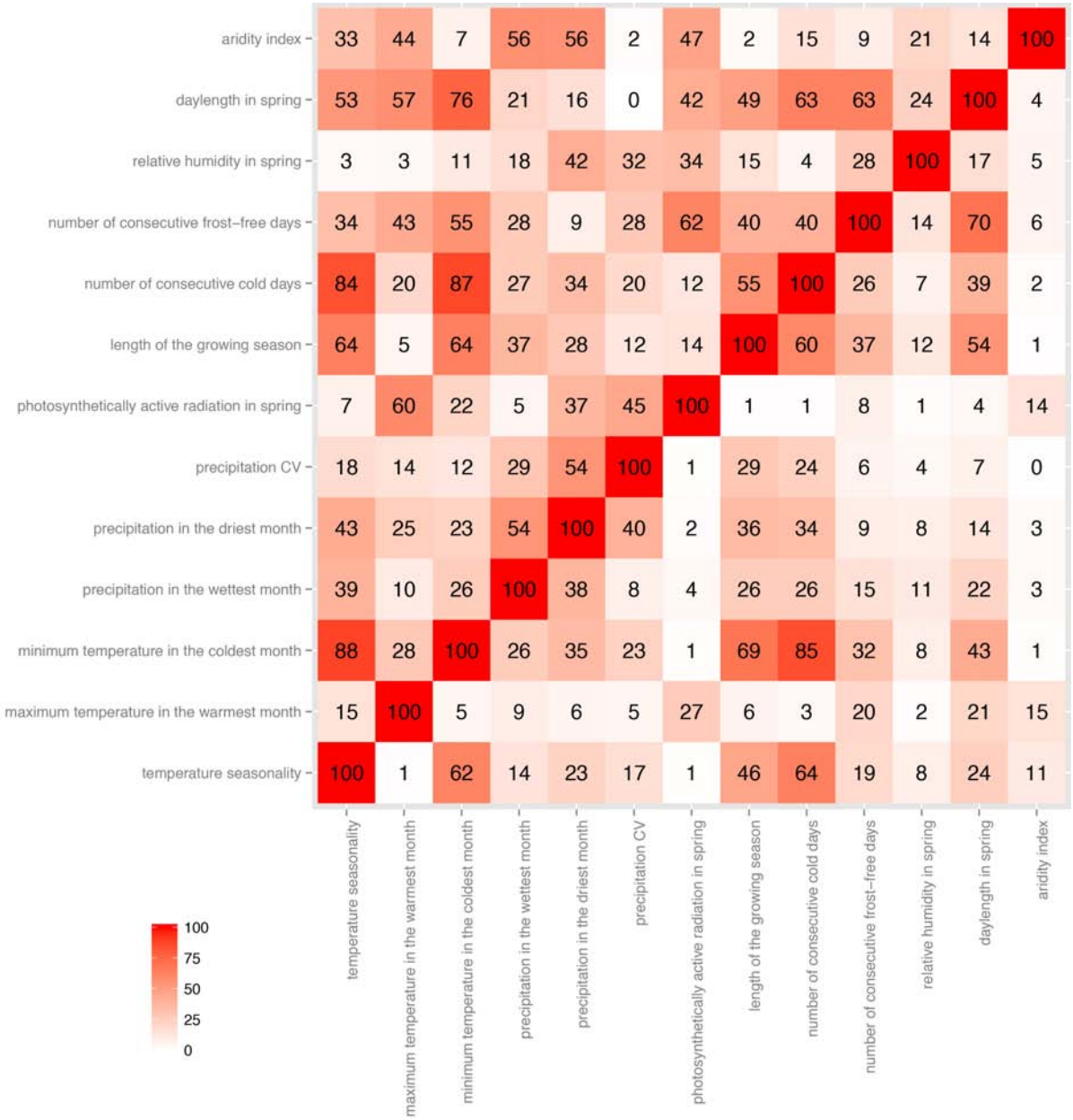
Supplementary Figure 40 - Comparison between the correlations among the climate variables (upper triangle) and the overlap in variance-heterogeneity GWA profiles (lower triangle). Numbers shown in the figure are percentages. Pearson's correlation coefficients were calculated for each pair of the climate variables. Overlaps in GWA profiles were calculated as the proportion of shared SNPs above the threshold of 1.0×10^{-4}.

Supplementary Figure 41 - Comparison between the correlations among the residual climate variables after genomic kinship correction (upper triangle) and the overlap in variance-heterogeneity GWA profiles (lower triangle). Numbers shown in the figure are percentages. Pearson's correlation coefficients were calculated for each pair of the climate variables. Overlaps in GWA profiles were calculated as the proportion of shared SNPs above the threshold of 1.0×10^{-4}.

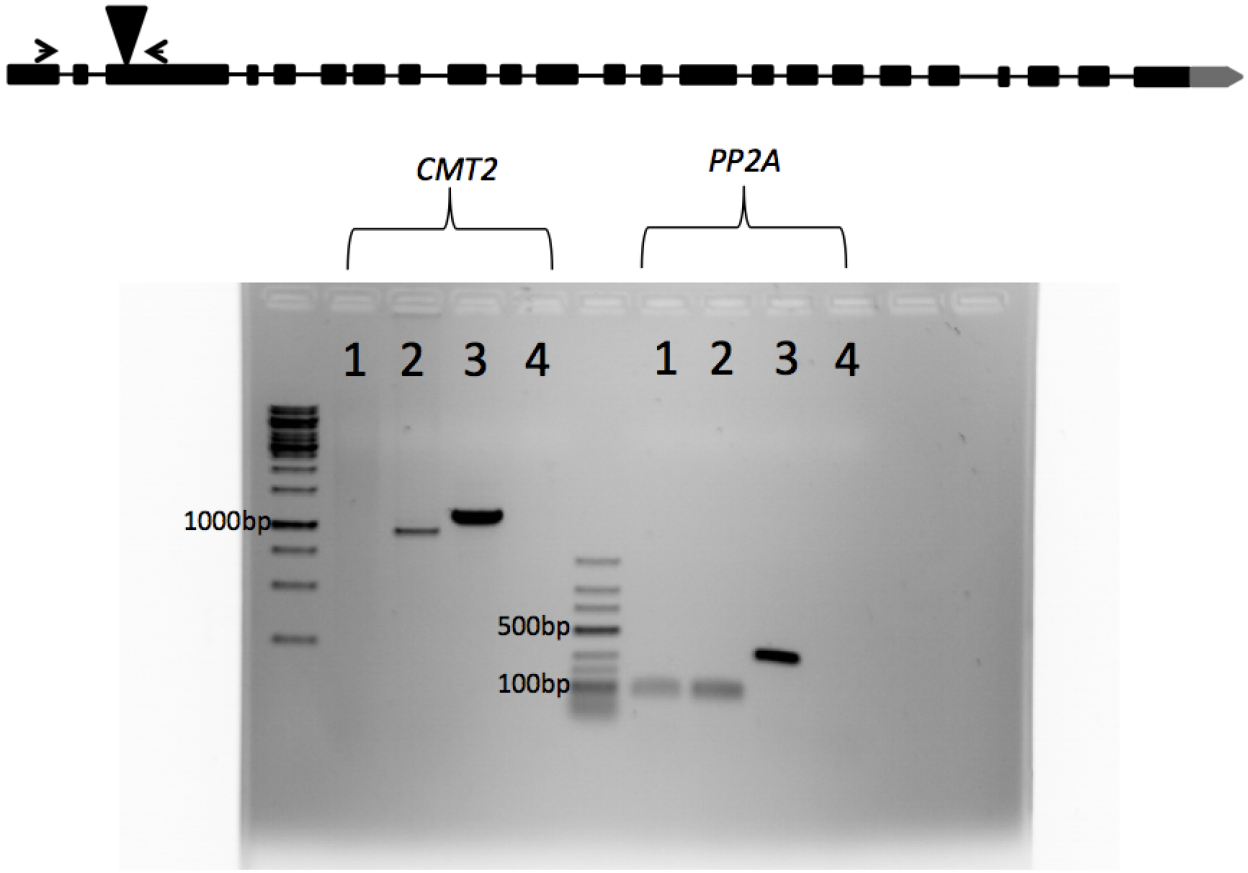


Supplementary Figure 42 - Comparison between the correlations among the residual climate variables after genomic kinship correction (upper triangle) and the correlations among the original climate variables (lower triangle). Numbers shown in the figure are percentages. Pearson's correlation coefficients were calculated for each pair of the climate variables.

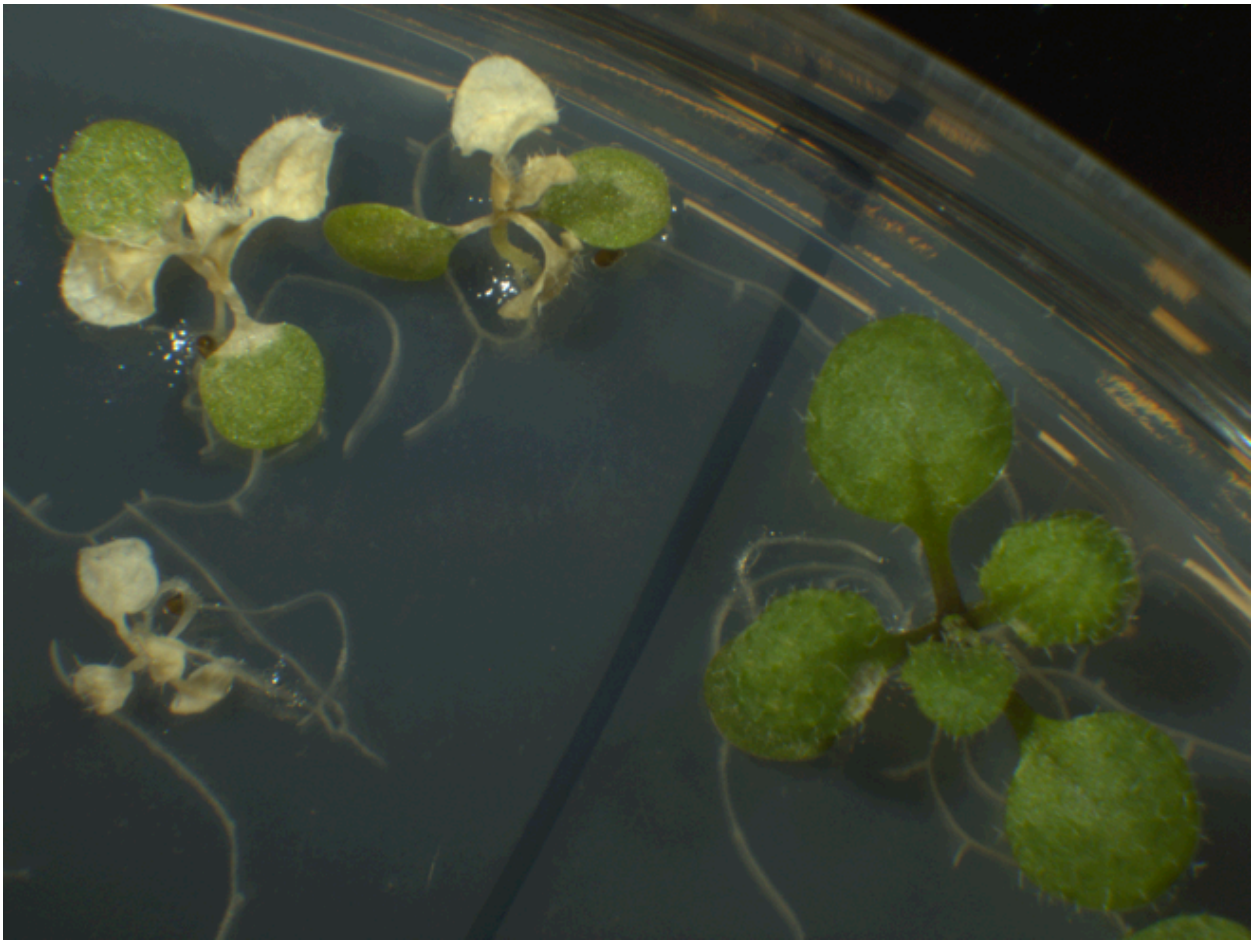
Supplementary Figure 43 - Comparison between the correlations among the climate variables (upper triangle) and the overlap in ordinary GWA profiles (lower triangle). Numbers shown in the figure are percentages. Pearson's correlation coefficients were calculated for each pair of the climate variables. Overlaps in GWA profiles were calculated as the proportion of shared SNPs above the threshold of 1.0×10^{-4}.



Supplementary Figure 44 - Comparison between the correlations among the climate variables (upper triangle) and the overlap in simple GWA profiles without correction for population structure (lower triangle). Numbers shown in the figure are percentages. Pearson's correlation coefficients were calculated for each pair of the climate variables. Overlaps in GWA profiles were calculated as the proportion of shared SNPs above the threshold of 1.0×10^{-4}.



Supplementary Figure 45: Gene-model of *CMT2* and T-DNA insertion confirmation. Boxes indicate exons, lines represent introns. The triangle shows the T-DNA insertion site. Arrow heads indicate the location of primers that were used to assay *CMT2* transcripts. *CMT2*: PCR reaction with *CMT2*-specific primers, PP2A: PCR reaction with PP2A-specific primers. Lanes 1: *cmt2*-5 cDNA, lanes 2: Col cDNA, lanes 3: Col genomic DNA, lanes 4: no template controls. *CMT2* cDNA and genomic DNA are predicted to give 940bp and 1159bp bands, respectively. *PP2A* cDNA and genomic DNA are predicted to give 84 bp and 210 bp bands, respectively.



Supplementary Figure 46: Prolonged heat stress is often lethal. 10-days-old seedlings were heat-stressed at 37.5°C for 24 h based on a published protocol (Ito et al. 2011). Plants were counted as non-viable if shoot apices were completely bleached. Note that the lamina of cotyledons often remains green for a longer time but no recovery was observed if apices were bleached.

Table S1. Detailed information about the missense mutations significantly associated with climate adaptability of Arabidopsis thaliana.

Trait	Chromosome	Position (bp)	Gene		Consequence	Position in cDNA	Position in CDS	Position in protein	Amino acid change	Codon change	Mutant frequency	Pase score	
												Physicochemical Property Score	MSA conservation Score
Temperature seasonality													
	4	10405599	AT4G19000	TFIIS protein family	missense variant	97	97	33	S/T	Tcg/Acg	83/728	0.21	0.35
	4	10410039	AT4G19006	PCI domain protein	missense variant	837	719	240	H/L	cAc/cTc	81/728	0.52	0.41
	4	10411801	AT4G19010	AMP-dependent synthetase and ligase family protein	missense variant	1643	1615	539	N/H	Aac/Cac	82/728	0.29	0.08
	4	10412877	AT4G19010		missense variant	1146	1118	373	S/L	tCg/tTg	80/728	0.52	0.14
	4	10413290	AT4G19010		missense variant	960	932	311	A/V	gCa/gTa	87/728	0.35	0.68
	4	10413374	AT4G19010		missense variant	876	848	283	K/R	aAg/aGg	88/728	0.23	0.2
	4	10414556	AT4G19020	CMT2	stop gained	31	31	11	E/*	Gag/Tag	79/728	NA	0.11
	4	10414640	AT4G19020		frameshift variant feature truncation	115	115	39	-	-	2/728	NA	0.17
	4	10414640	AT4G19020		missense variant	115	115	39	E/K	Gaa/Aaa	85/728	0.63	0.17
	4	10414747	AT4G19020		missense variant	222	222	74	N/K	aaC/aaG	81/728	0.45	0.11
	4	10415805	AT4G19020		missense variant	1061	1061	354	G/D	gGc/gAc	78/728	0.58	0.11
	4	10415833	AT4G19020		missense variant	1089	1089	363	L/F	ttA/ttT	80/728	0.2	0.14
	4	10416047	AT4G19020		missense variant	1303	1303	435	G/S	Ggt/Agt	80/728	0.26	0.14
	4	10445425	AT4G19060	P-loop containing nucleoside triphosphate hydrolases superfamily protein	missense variant	876	809	270	I/T	aTt/aCt	79/728	0.37	0.44
	4	10447160	AT4G19070	Putative membrane lipoprotein	missense variant	529	418	140	S/R	Agt/Cgt	78/728	0.73	0.11
	4	10450001	AT4G19090	unknown protein	missense variant	49	49	17	R/G	Aga/Gga	80/728	0.93	0.07
	4	10455784	AT4G19110	Protein kinase, superfamily protein	missense variant	610	576	192	L/F	ttG/ttT	76/728	0.2	0.45
	4	10455784	AT4G19110		missense variant	1682	744	248	L/F	ttG/ttT	76/728	0.2	0.44
	4	10455784	AT4G19110		missense variant	1652	744	248	L/F	ttG/ttT	76/728	0.2	0.44
Maximum temperature in the warmest month													
	1	6936457	AT1G19990	unknown protein	missense variant	657	548	183	S/F	tCt/tTt	63/728	0.64	0.2
Minimum temperature in the coldest month													
	5	14067526	AT5G35930	AMP-dependent synthetase and ligase family protein	missense variant	3153	2887	963	E/Q	Gaa/Caa	87/728	0.3	0.05
Number of consecutive cold days													
	5	7492033	AT5G22560	unknown protein	missense variant	1144	1065	355	E/D	gaG/gaC	99/728	0.26	0.22
	5	7492259	AT5G22560		missense variant	918	839	280	R/T	aGa/aCa	96/728	0.63	0.11
	5	7492277	AT5G22560		missense variant	900	821	274	N/T	aAt/aCt	100/728	0.27	0.11
	5	7492277	AT5G22560		missense variant	900	821	274	N/S	aAt/aGt	100/728	0.3	0.11
Relative humidity in spring & Day length in spring													
	4	14788320	AT4G30200	VEL1	missense variant	963	708	236	E/D	gaA/gaT	79/728	0.26	0.06
	4	14788320	AT4G30200		missense variant	1057	759	253	E/D	gaA/gaT	79/728	0.26	0.06
	4	14829581	AT4G30290	XTH19	missense variant	349	298	100	I/V	Att/Gtt	54/728	0.14	0.49
Minimum temperature in the coldest month & Number of consecutive cold days													
	2	19397389	AT2G47240	AMP-dependent synthetase and ligase family protein	missense variant	1966	1859	620	K/R	aAa/aGa	47/728	0.23	0.13
	2	19397389	AT2G47240		missense variant	2068	1859	620	K/R	aAa/aGa	47/728	0.23	0.13
Temperature seasonality & Day length in spring													
	2	12169734	AT2G28470	BGAL8	missense variant	2338	2016	672	E/D	gaA/gaT	77/728	0.26	0.06
	2	12169734	AT2G28470		missense variant	2336	2034	678	E/D	gaA/gaT	77/728	0.26	0.06
	2	12169828	AT2G28470		missense variant	2242	1940	647	F/Y	tTc/tAc	89/728	0.36	0.72
	2	12169828	AT2G28470		missense variant	2244	1922	641	F/Y	tTc/tAc	89/728	0.36	0.72
Number of consecutive frost-free days													
	1	954782	AT1G03790	SOM	missense variant	258	194	65	N/I	aAt/aTt	183/728	0.6	0.06
	1	955189	AT1G03790		missense variant	665	601	201	P/T	Cct/Act	181/728	0.22	0.17
	1	955268	AT1G03790		missense variant	744	680	227	S/C	tCt/tGt	180/728	0.55	0.4

Table S2. Loci significantly associated with climate adaptability of Arabidopsis thaliana but without non-synonymous mutations in high LD detected.
P-values were obtained from linear regression of squared z-scores. GC P-values were the P-values after genomic control.
Gamma P-values were obtained by fitting generalized linear models with Gamma response.
Pleiotropic loci are marked with stars. bp = base pair; MAF = minor allele frequency.

Trait	Leading SNP					
	Chromosome	*Position (bp)*	*MAF*	*P-value*	*GC P-value*	*Gamma P-value*
Relative humidity in spring						
	3	1816353	0.072	1.8E-10	1.2E-08	1.0E-14
	5	8380640	0.073	1.4E-09	6.3E-08	9.2E-17
Minimum temperature in the coldest month & Number of consecutive cold days*						
	2	18620697	0.083	7.2E-11	3.5E-08	1.2E-11
				*7.2E-10	*1.7E-07	*1.4E-11
	5	18397418	0.107	3.8E-10	1.2E-07	1.5E-13
				*8.8E-10	*1.9E-07	*3.2E-13
Length of the growing season						
	3	576148	0.084	1.1E-09	1.4E-07	8.9E-09
Number of consecutive frost-free days						
	1	953031	0.237	1.0E-08	8.0E-08	2.4E-21
	1	6463065	0.081	2.7E-08	1.9E-07	2.9E-09
	2	9904076	0.215	2.3E-09	2.2E-08	1.8E-10

Table S3. Experimental data of the heat-stress treatment on Col-0 and cmt2 individuals.

experiment	accession	total	alive	dead
1	cmt2-5	28	17	11
1	cmt2-5	25	21	4
1	Col-0	25	12	13
1	Col-0	23	5	18
2	Col-0	28	8	20
2	Col-0	30	4	26
2	cmt2-5	26	10	16
2	Col-0	25	4	21
2	cmt2-5	19	11	8
2	Col-0	29	11	18
3	cmt2-5	25	20	5
3	Col-0	29	21	8
3	cmt2-5	26	16	10
3	Col-0	29	12	17
3	Col-0	29	13	16
3	Col-0	29	14	15

Table S4. Experimental data of root growth (mm) of Col-0 and cmt2 knockouts, with and without 6 h heat stress.

Col		cmt2	
control	6h heat	control	6h heat
11.839	13.48	17.375	20.214
10.911	0.364	28.15	22.313
20.818	22.539	26.546	23.846
8.357	0	20.006	22.25
13.656	1.214	19.581	18.644
23.621	21.834	8.576	9.141
16.425	18.095	23.061	18.037
26.866	7.785	26.788	3.843
7.801	14.861	23.047	9.019
23.397	8.013	33.445	24.753
11.484	0.243	27.716	6.607
8.684	10.881	21.39	33.157
18.94	10.902	23.379	19.084
21.055	1.093	14.33	5.901
21.355	0	19.115	22.115
26.758	16.586	29.795	1.75
25.958	6.19	18.046	1.996
23.438	0.777		6.376
23.419	12.242		17.121
22.792	0.384		15.948
21.526	1.341		7.113
26.55	0.738		19.502
16.929	0.619		6.085
29.122	24.244		14.354
27.851	25.941		1.573
24.558	10.512		16.381
24.275	0.243		5.892
30.535	23.916		
30.152			
18.197			