SM2 - Adaptation to climate in European lines of *A. thaliana*

This supplementary material file explains how we ran ecological association tests to evaluate the correlation of individual genotype to climatic variables in the model species *A. thaliana*. *A. thaliana* lines were genotyped using a SNP-chip containing 230k SNPs with density of one SNP per 500 bp (data from Atwell *et al.* 2010). In this study, we restricted our analysis to chromosome 1 (26k SNPs, density of one SNP per kbp). The analysis of chromosomes 2-5 could follow the same guidelines as shown in this supplementary file.

Eighteen bioclimatic variables were extracted for each of the 170 samples using the WorldClim data set at the 30 arcsecond (1km^2) resolution (Hijmans *et al.* 2005). A statistical summary of the 18 variables was computed as the first principal component of the bioclimatic data, and this climatic gradient was used when we tested for associations between loci and environment. The genotypes for each individual were recorded in a matrix stored in the file "chr1geno.lfmm" and the climatic variable was stored in the file "chr1var.env".

Running 5 genome scans for selection with LFMM

We performed genome scans for selection using 5 distinct latent factor mixed models (LFMMs). A preliminary analysis of population genetic structure from the genotypic matrix suggested that $K = 6$ clusters (or 5 principal components) best explained the genetic data. We used $K = 6$ latent factors in subsequent analysis of the data with LFMMs.

The R command line was run as follows.

```
Library(LEA)

object.lfmm = lfmm("chr1geno.lfmm", "chr1var.env", K = 6, rep = 5)
```

Then, the z-scores resulting from the fit of each of the 5 models were recorded in a single matrix **zs** as follows.

```
zs = z.scores(object.lfmm, K = 6)
```

Combining LFMM tests

Next we applied a meta-analysis of the 5 LFMM models and combined their z-scores using a robust variant of the Stouffer method. In this method, the mean value of the scores was replaced by the median value at each locus.

```
zs.stouffer = apply(zs, MARGIN = 1, median)
```

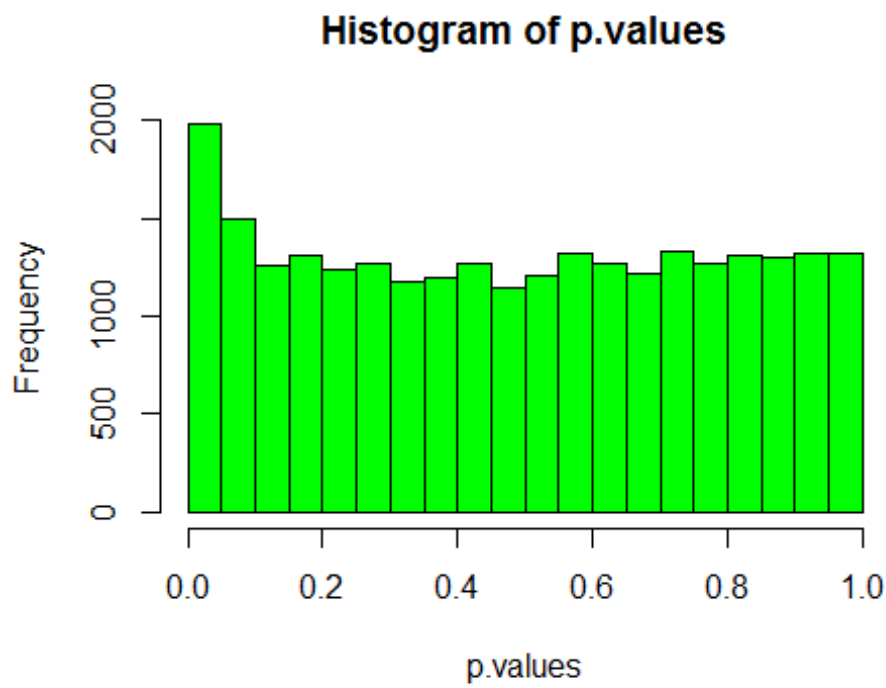
Then we corrected for the bias created by the use of combined tests by estimating the genomic inflation factor as follows

```
lambda = median(zs.stouffer^2)/0.456
lambda
```

```
## [1] 0.8155406
```

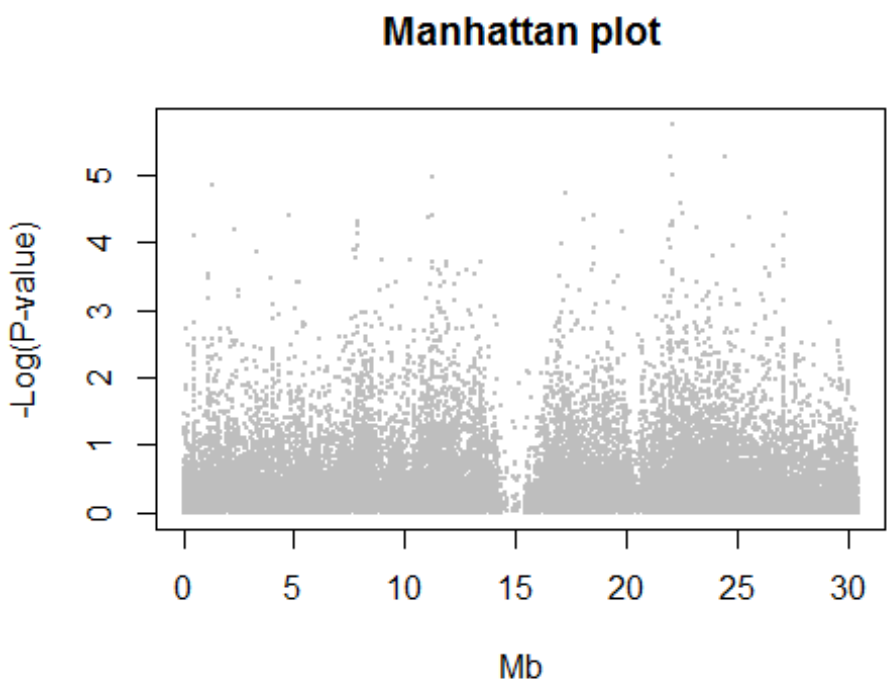
We computed significance values from the combined z-scores using the chi-squared distribution with one degree of freedom as follows. Note that we proposed a subtle modification of the inflation factor that led to an improved correction using $\lambda = 0.78$ instead of $\lambda = 0.81$.

```
lambda = .78
p.values = pchisq(zs.stouffer^2/lambda, df = 1, lower = F)
hist(p.values, col = "green")
```



The histogram of significance values had the desired shape. It was flat and exhibited a peak close to zero. This plot provides evidence that the false discoveries could be controlled. Using the genetic map, a Manhattan plot for the significance values can be displayed as follows (log scale).

```
pos1 = read.table("poschr1.txt")

p1 = as.numeric(pos1)/1000000
plot(p1, -log10(p.values), main = "Manhattan plot", xlab = "Mb",
col = "grey", pch = 19, cex=.2, ylab = "-Log(P-value)")
```

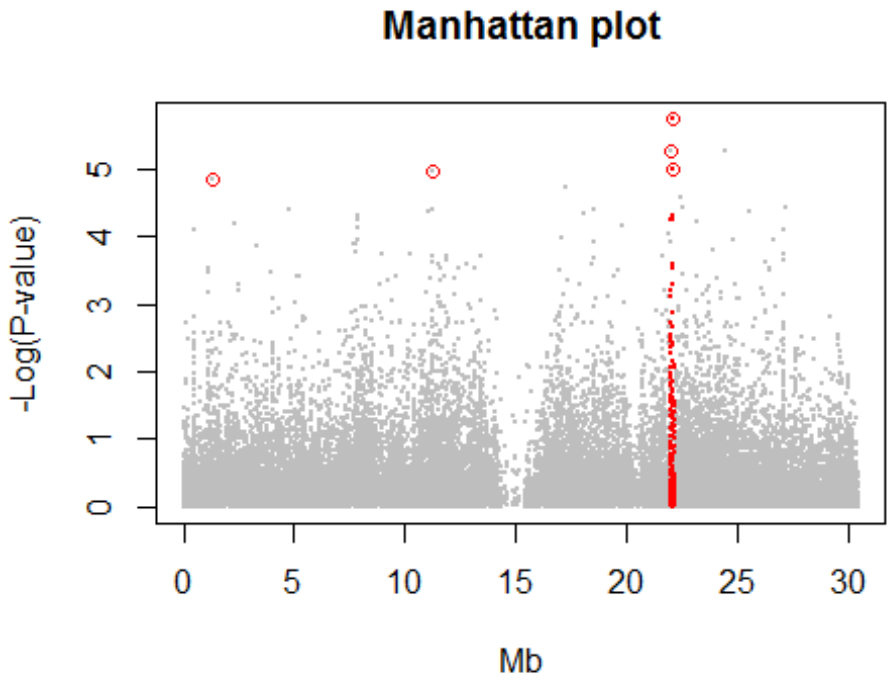


We eventually decided to control the FDR at the level of 7%. The resulting list of candidate loci was stored in the object **candidates** as follows.

```
q.level = 0.07
L = length(p.values)

#Benjamini-Hochberg algorithm
w = which(sort(p.values) < q.level * (1:L) / L)
candidates = order(p.values)[w]
```

The list of candidate loci was very short. Their positions in chromosome 1 were identified as follows.

```
pos1[candidates]

## [1] 22066552 21940236 22064235 11194101  1297710

plot(p1, -log10(p.values), main = "Manhattan plot", xlab = "Mb",
col = "grey", pch = 19, cex=.2, ylab = "-Log(P-value)")
points(p1[candidates], -log10(p.values)[candidates], col = "red")
points(p1[18000:18200], -log10(p.values[18000:18200]), main =  "Manhattan
plot", xlab = "Mb", col = "red", pch = 19, cex=.2, ylab = "-Log(P-value)")
```



The top hits in the candidate list identified a protein coding region corresponding to the gene *ARR3* (locus AT1G59940), which is a type A response regulator gene involved in the circadian rhythm. It is also involved in the cytokinin-activated signaling pathway. The protein coding genes around positions 1,297,710 and 11,194,101 had unknown functions.