# Supplemental Information

**Liu *et al.* Selecting closely-linked SNPs based on local epistatic effects for haplotype construction improves power of association mapping**

# Supplemental Tables

## Table S1. Summary of results for three haplotype-based GWAS approaches.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data set | FT10 | | | FT16 |
| GWAS approach | SWH | LDH | FH | FH |
| Number of candidate SNPs | 756005 | 756005 | 10214 | 9661 |
| Number of SNPs grouped in haplotypes | 756005 | 756005 | 3790 | 3265 |
| Total number of haplotypes | 755995 | 533551 | 157526 | 81187 |
| Phaplotype < PSNP | 122194 | 34145 | 126238 | 72606 |
| PSNP < Phaplotype | 633801 | 63367 | 31288 | 8581 |

## Table S2. Impact of SNPs in perfect LD on GWAS results for data set FT10. For each group of SNPs that were in perfect in LD to each other only a single SNP was considered in SNPLD GWAS. Haplotypes solely differentiated by SNPs in perfect LD to each other were assessed just once in FHLD GWAS. “NA” stands for not applicable.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| GWAS approach | SNP | SNPLD | FH | FHLD |
| No of SNPs for GWAS | 756005 | 576544 | 756005 | 576544 |
| No of candidate SNPs | NA | NA | 10214 | 7607 |
| No of SNPs grouped in haplotypes | NA | NA | 3790 | 2753 |
| No of haplotye combinations | NA | NA | 8932265 | 2460933 |
| No of funtional haplotyes | NA | NA | 157256 | 44759 |
| Significance threshold (‑log10*P*) | 7.18 | 7.06 | 8.29 | 7.79 |
| No of significant associations in region I | 4 | 4 | 71 | 183 |
| No of representative significant associations in region I | NA | NA | 2 | 2 |
| No of significant associations in region II | 0 | 0 | 15 | 28 |
| No of representative significant associations in region II | NA | NA | 2 | 2 |
| No of significant associations in region III | 1 | 2 | 701 | 1099 |
| No of representative significant associations in region III | NA | NA | 6 | 5 |
| No of significant associations in region IV | 5 | 5 | 19030 | 24779 |
| No of representative significant associations in region IV | NA | NA | 3 | 3 |
| No of significant associations in region V | 1 | 1 | 4952 | 6240 |
| No of representative significant associations in region V | NA | NA | 5 | 8 |
| No of significant associations in region on chr. 3 | 0 | 0 | 0 | 1 |
| Total number of significant associations | 11 | 12 | 24769 | 32330 |

## Table S3. Sizes of significant functional haplotypes and regions covered by overlapping significant functional haplotypes determined for data set FT10 by FH GWAS.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Minimum size of functional haplotypes | Median size of functional haplotypes | Mean size of functional haplotypes | Maximum size of functional haplotypes | Size of region spanned by overlapping significant haplotypes |
| Region I | 3925 | 8125 | 10552.73 | 48281 | 54329 |
| Region II | 27836 | 47828 | 41291.40 | 48542 | 61460/48542 |
| Region III | 5459 | 37964 | 34309.58 | 49986 | 119318 |
| Region IV | 2683 | 19390 | 26223.99 | 49341 | 61723 |
| Region V | 1903 | 26914.5 | 31940.29 | 49999 | 167210 |

## Table S4. Characteristics of representative functional haplotypes (RH) in two data sets applying FH GWAS.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Data set | No of RH | Mininimum size of RH | Maximum size of RH | Size of region in which RH were found | Start position of region | End position of region |
| Region I | FT10 | 2 | 7813 | 9646 | 17426 | 24330357 | 24347783 |
| Region II | FT10 | 2 | 27836 | 37134 | 61460 | 159026 | 220486 |
| Region II | FT16 | 2 | 33351 | 44682 | 45055 | 157993 | 203048 |
| Region III | FT10 | 6 | 21577 | 49265 | 98917 | 3128717 | 3227634 |
| Region III | FT16 | 8 | 6627 | 45547 | 96491 | 3127425 | 3223916 |
| Region IV | FT10 | 3 | 19394 | 41732 | 58538 | 18552017 | 18610555 |
| Region IV | FT16 | 5 | 18650 | 41555 | 41555 | 18552194 | 18593749 |
| Region V | FT10 | 5 | 14774 | 43889 | 53182 | 23203762 | 23256944 |

## Table S5. Comparisons of the percentages of phenotypic variance (Pv) explained by SNPs analyzed in SNP-based GWAS and significant functional haplotypes detected by FH GWAS using data set FT10.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | SNPs | Significant functional haplotypes | | | | Representative significant functional haplotypes | |
| Pv | Maximum | Minimum | Median | Mean | Maximum | Minimum | Maximum |
| Region I | 12.95 | 10.72 | 13.35 | 12.79 | 15.37 | 15.04 | 15.34 |
| Region II | 14.34 | 8.08 | 8.84 | 11.66 | 18.31 | 15.00 | 18.31 |
| Region III | 19.96 | 5.75 | 13.94 | 18.20 | 29.93 | 20.18 | 28.99 |
| Region IV | 23.28 | 17.27 | 29.81 | 28.82 | 39.23 | 25.21 | 39.23 |
| Region V | 14.74 | 1.58 | 12.39 | 12.37 | 26.10 | 13.41 | 26.10 |

## Table S6. Comparisons of FH GWAS characteristics for regions significantly associated with the trait flowering time in two data sets.

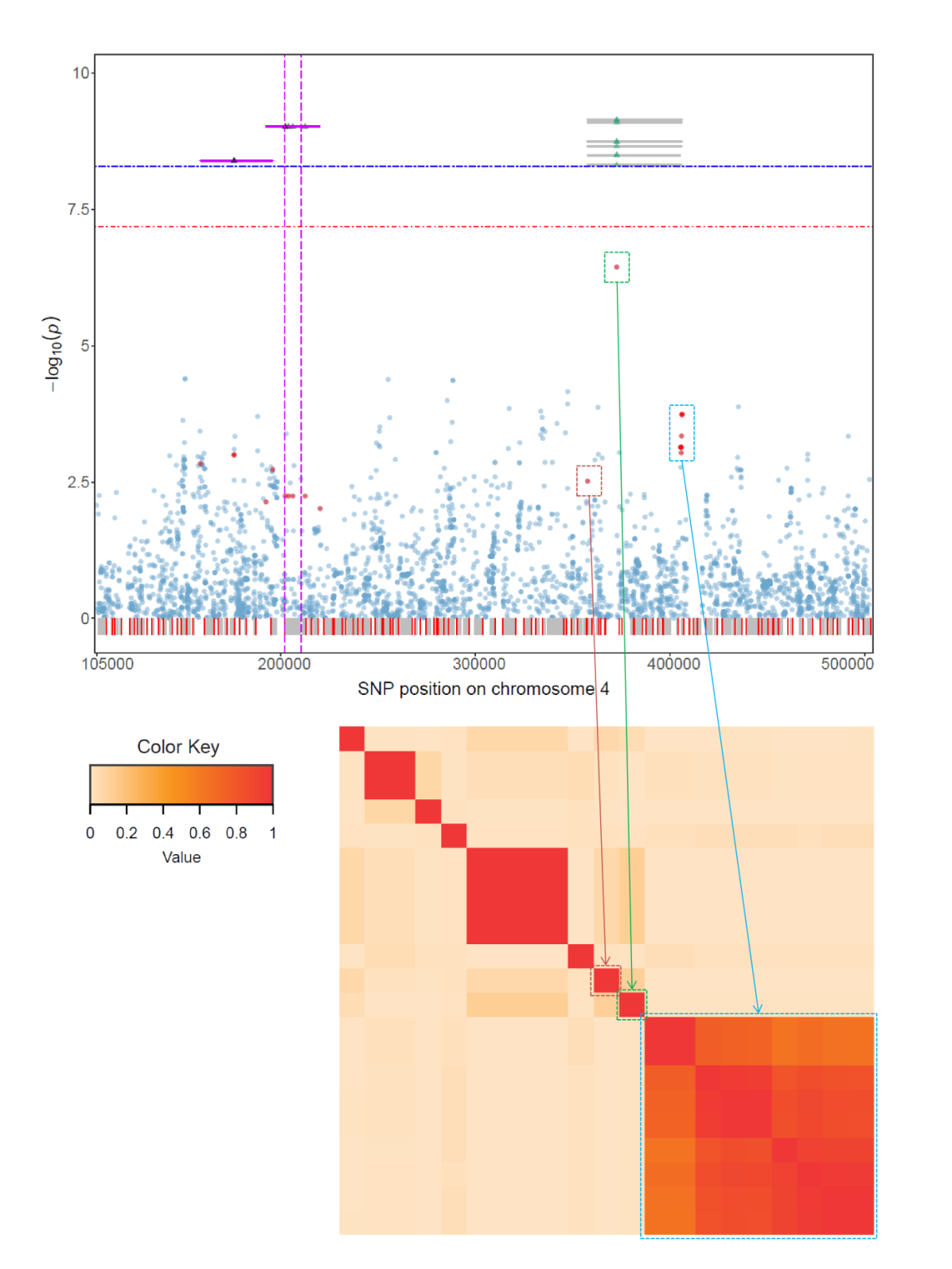
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | No of associations in data set FT10 | No of associations in data set FT16 | No of identical associations in both data sets |
| Region II | Candidate SNPs | 300 | 302 | 151 |
|  | SNPs in the functional haplotypes | 268 | 261 | 123 |
|  | SNPs in the significant functional haplotypes | 21 | 23 | 1 |
|  | SNPs in the representative haplotypes | 6 | 6 | 0 |
|  | Functional haplotypes | 3754 | 6138 | 209 |
|  | Significant functional haplotypes | 15 | 20 | 0 |
|  | Representative functional haplotypes | 2 | 2 | 0 |
| Region III | Candidate SNPs | 134 | 200 | 113 |
|  | SNPs in the functional haplotypes | 130 | 200 | 111 |
|  | SNPs in the significant functional haplotypes | 114 | 163 | 88 |
|  | SNPs in the representative haplotypes | 13 | 17 | 4 |
|  | Functional haplotypes | 7700 | 36681 | 2687 |
|  | Significant functional haplotypes | 701 | 3404 | 307 |
|  | Representative functional haplotypes | 6 | 8 | 0 |
| Region IV | Candidate SNPs | 206 | 162 | 126 |
|  | SNPs in the functional haplotypes | 204 | 162 | 124 |
|  | SNPs in the significant functional haplotypes | 178 | 112 | 87 |
|  | SNPs in the representative haplotypes | 8 | 13 | 2 |
|  | Functional haplotypes | 56411 | 14616 | 710 |
|  | Significant functional haplotypes | 19030 | 1161 | 26 |
|  | Representative functional haplotypes | 3 | 5 | 0 |
| Total | Candidate SNPs | 640 | 664 | 390 |
|  | SNPs in the functional haplotypes | 602 | 623 | 358 |
|  | SNPs in the significant functional haplotypes | 313 | 298 | 176 |
|  | SNPs in the representative haplotypes | 27 | 36 | 6 |
|  | Functional haplotypes | 67865 | 57435 | 3606 |
|  | Significant functional haplotypes | 19746 | 4585 | 333 |
|  | Representative functional haplotypes | 11 | 15 | 0 |

## Table S7. Percentage of functional haplotypes outperforming the most significant SNP of a particular haplotype in nine different simulation scenarios.

|  |  |  |  |
| --- | --- | --- | --- |
|  | MAF range:  0.4-0.5 | MAF range:  0.2-0.3 | MAF range:  0-0.1 |
| LD range: 0-0.2 | 99.60 | 97.20 | 29.40 |
| LD range: 0.3-0.6 | 97.88 | 96.18 | 72.50 |
| LD range: 0.7-1 | 43.39 | 42.42 | 34.40 |

# Supplemental FiguresE:\liufang\study\data\gwas\haplotype\simulation\Molecular Plant\MP\Supplementary Figure 2.png

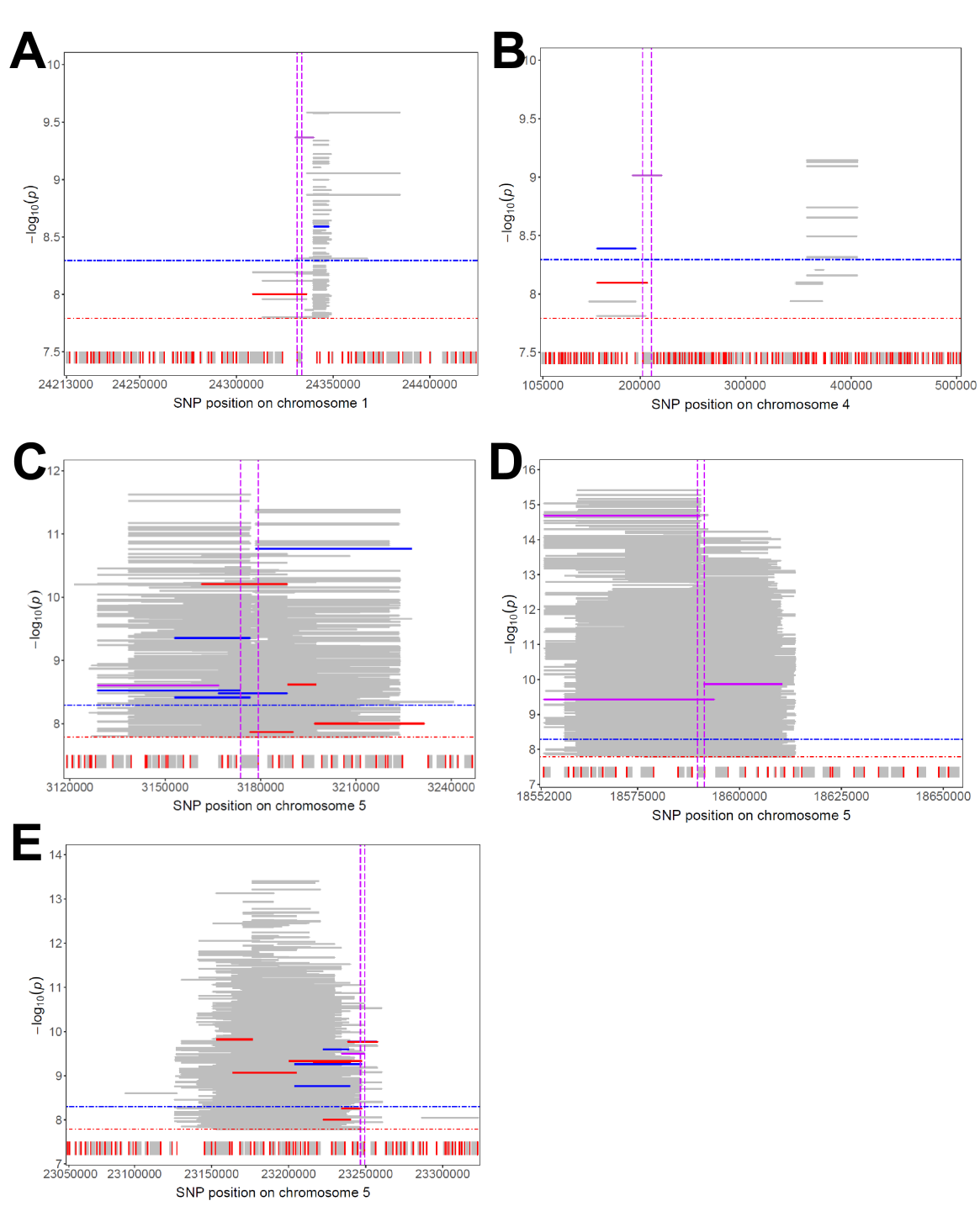
**Figure S1. Haplotype construction based on overlapping sliding-windows (A) or linkage disequilibrium (B).** Using the sliding window approach consecutive SNPs are grouped into haplotypes based on a fixed window length. Windows are moved along the chromosome according to the step size. In the example shown window and step size corresponded to three and one, respectively. In the LD approach LD between consecutive SNPs is considered for haplotype construction. In all cases in which successive SNPs equal or exceed the LD threshold, they are grouped into a haplotype. SNPs not included into haplotypes were defined as haplotypes with a length of one.



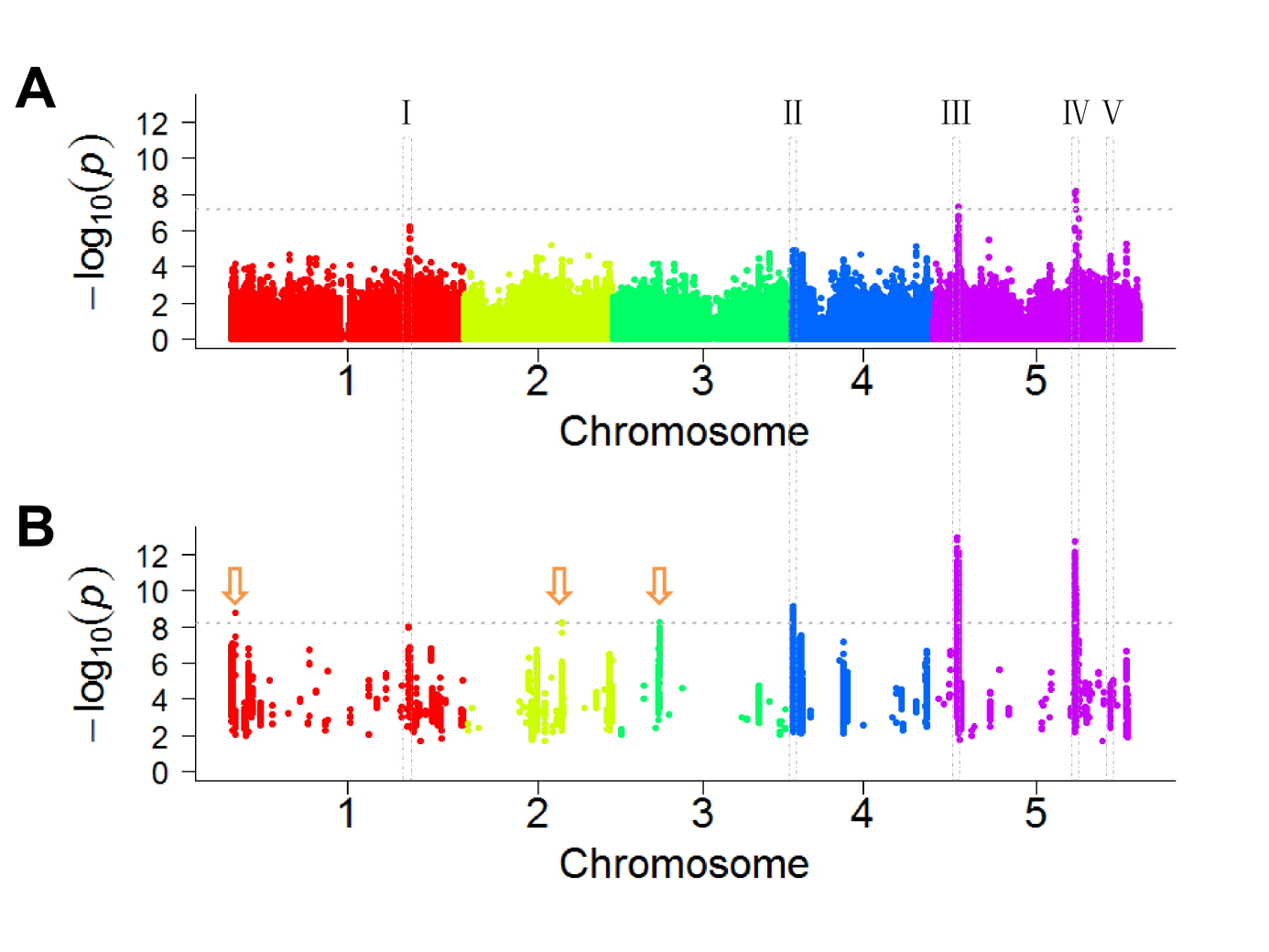
## Figure S2. Significant associations for the trait flowering time revealed by FH GWAS in region II of *Arabidopsis thaliana* chromosome 4 using data set FT0. In the top panel, the region in which overlapping functional haplotypes were found is displayed together with the position of the candidate gene as described in Figure 2. The bottom panel shows an LD heat map plot for those SNPs that made up the significant haplotypes. The color key for the LD as measured by the *r*2 statistic is shown on the left.

## 

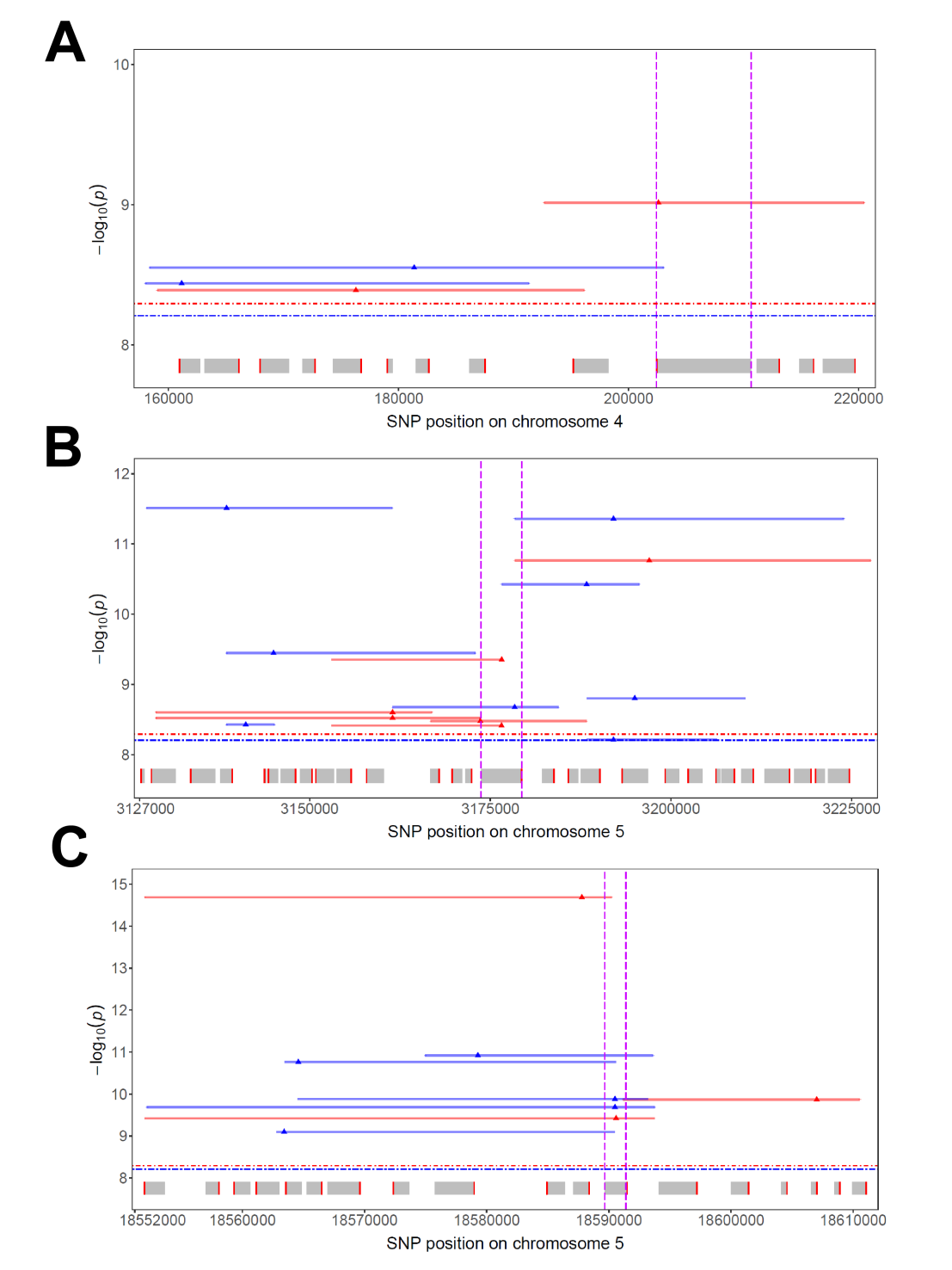
## Figure S3. Manhattan plots illustrating the results of SNPLD GWAS and FHLD GWAS for data set FT10. For each group of SNPs that were in perfect in LD to each other only a single SNP was considered in SNPLD GWAS (A). Haplotypes solely differentiated by SNPs in perfect LD to each other were assessed just once in FHLD GWAS (B). Positions of SNPs or haplotypes on the five chromosomes are shown on the x axis relative to their ‑log10(*P*) values on the y axis. In panel (A) the horizontal dotted grey line corresponds to the thresholds after Bonferroni correction for multiple testing (Dunn 1961) (*P* < 0.05). The threshold shown as a dotted pale blue line in panel (B) was adjusted after Bonferroni correction for multiple testing by taking into account the pre-testing procedure for single SNP main and epistatic effects implemented in the functional haplotype approach. The five regions in which significant associations had been identified in FH GWAS are marked by stippled lines to ease a comparison with the results shown in Figure 2. The arrow indicates a significant association that was identified in FHLD GWAS but not FH GWAS.



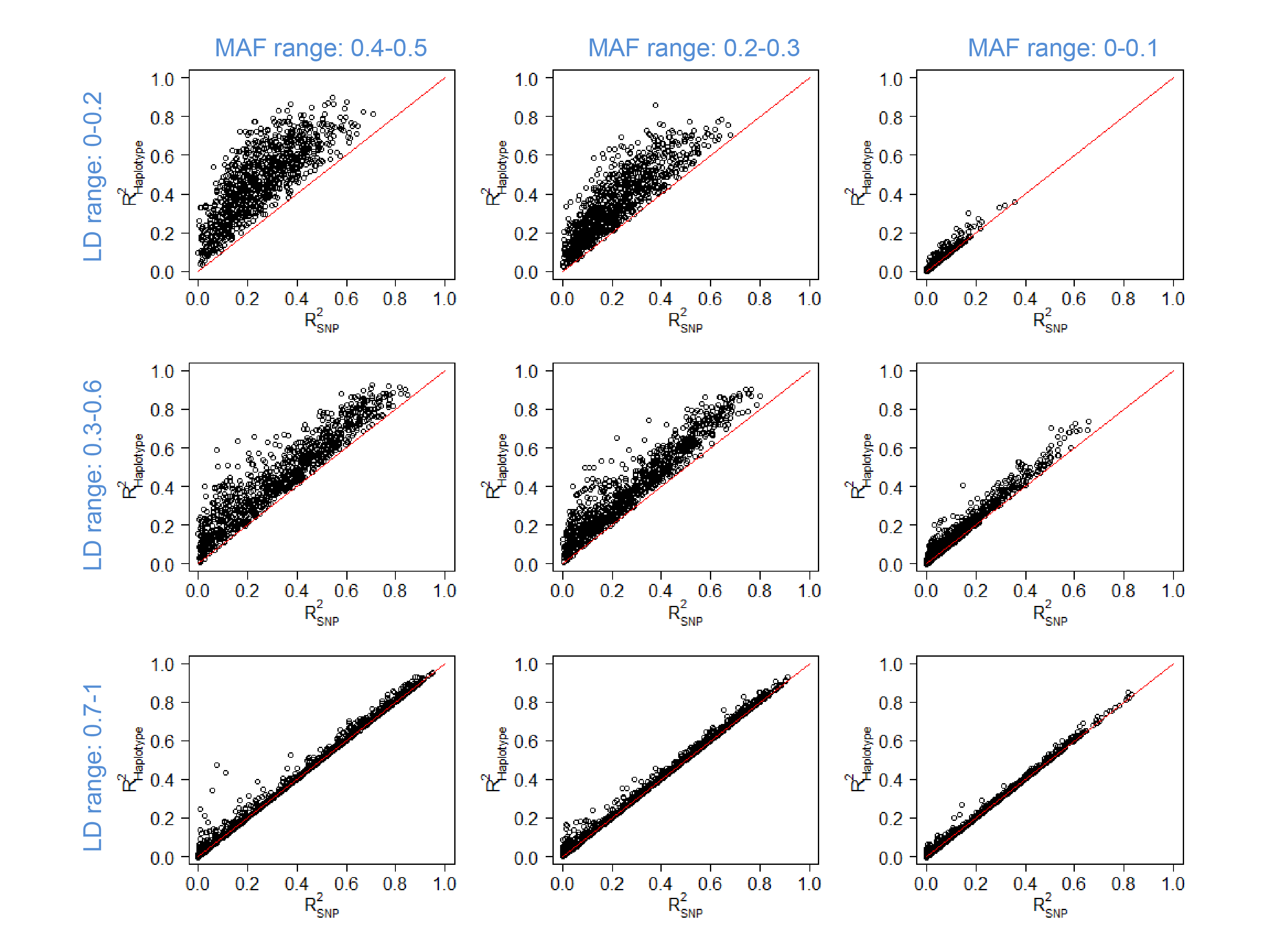
## Figure S4. Significant associations for the trait flowering time revealed by FH GWAS for data set FT10. Significant functional haplotypes in regions I, II, III, IV and V are illustrated in panels (A)-(E), respectively. The significance thresholds for FH GWAS and FHLD GWAS are shown as blue and red horizontal stippled lines, respectively. Extent of haplotypes are shown as lines relative to the SNP positions on the x axis. The ‑log10(*P*) values of the functional haplotypes are displayed on the y axis. Representative haplotypes determined for FH GWAS and FHLD GWAS are shown as blue and red lines, respectively and those that were in common between FH GWAS and FHLD GWAS are displayed as pink lines. Coding regions of genes are represented as grey boxes, red lines indicate 5’-regions of genes. Vertical pink dashed lines delimit the position of the candidate genes.



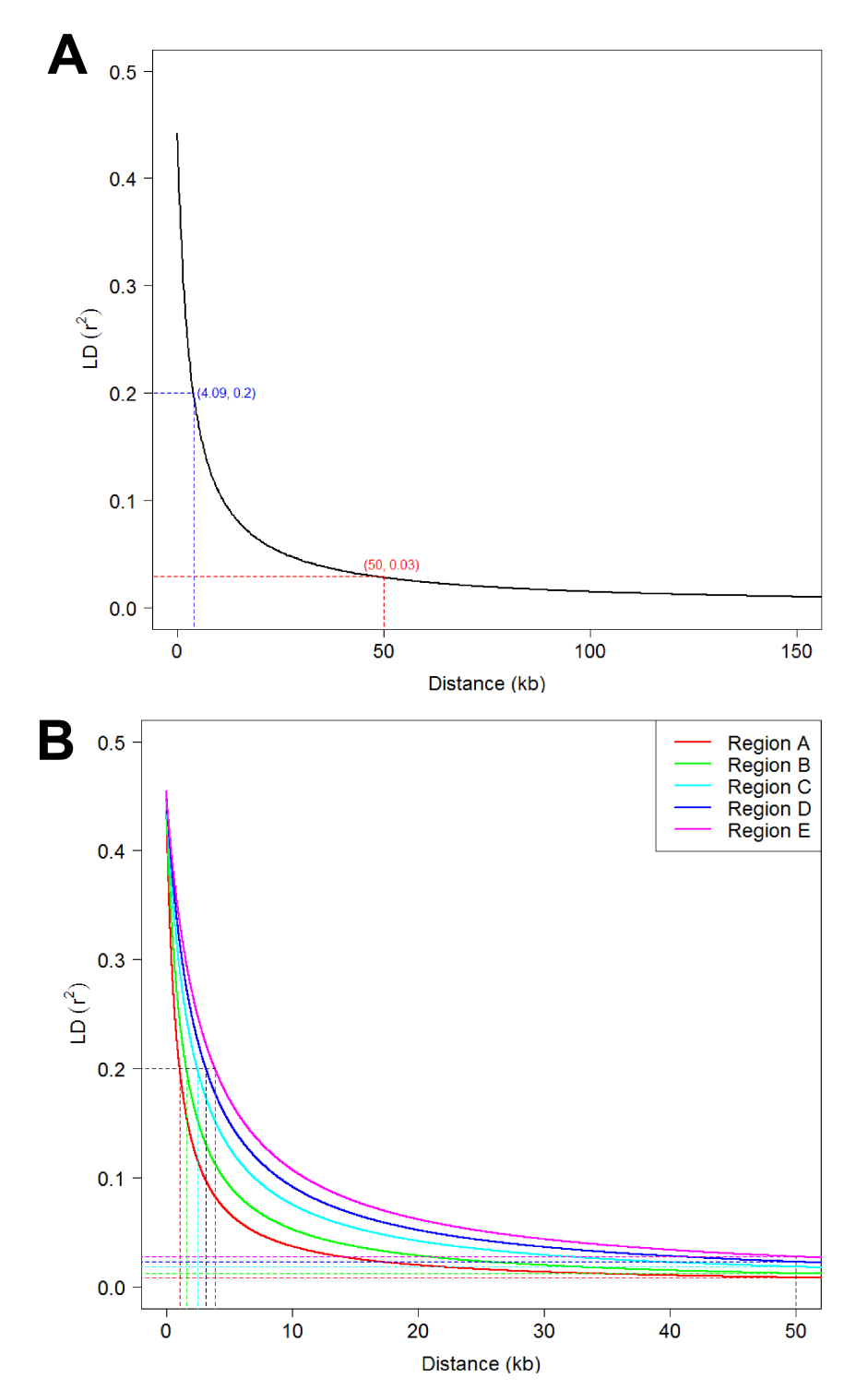
**Figure S5. Association mapping results using SNP-based and FH GWAS using data set FT16.** Panels (**A**) and (**B**) show Manhattan plots for GWAS using single SNPs and functional haplotypes, respectively. Positions of SNPs (**A**) or haplotypes (**B**) on the five chromosomes are shown on the x axis relative to their ‑log10(*P*) values on the y axis. The horizontal dotted grey line in panel (**A**) represents the significance threshold (‑log(*P*) = 7.18) after Bonferroni correction for multiple testing (*P* < 0.05). For the results of FH GWAS (**B**) a more stringent significance threshold (-‑log(*P*) = 8.21) is indicated to adjust for the pre-testing procedure for single SNP main and epistatic effects. The five regions in which significant associations had been identified for plants cultivated at 10°C are marked by stippled lines to ease a comparison with the results shown in **Figure** **2**. Arrows indicate significant associations that were only identified in data set FT16.



## Figure S6. Comparison of representative significant functional haplotypes (RH) associated with the trait flowering time in two different data sets. Representative significant functional haplotypes in regions II, III and IV are illustrated in panels (A), (B) and (C), respectively. RHs that had been identified in data sets FT10 and FT16 are shown as red and blue lines, respectively. Triangles indicate SNP positions. SNP positions are displayed on the x axis relative to ‑log10(*P*) values of the RHs on the y axis. The significance thresholds for FH GWAS are shown as red and blue horizontal stippled lines for data sets FT10 and FT16, respectively. Grey boxes represent the coding regions of genes and red lines indicate 5’-regions of genes. Candidate genes are marked by two vertical pink dashed lines.



**Figure S7. Proportions of phenotypic variance explained by functional haplotypes in nine different simulation scenarios.** For each haplotype the R2 value represented on the y axis is plotted relative to the highest adjusted R2 value that was observed for a SNP of this particular haplotype on the x axis. Plots are arranged in order of decreasing MAF range and increasing LD range.



## Figure S8. Decay of linkage disequilibrium in the population of data set FT10. Panel (A) shows the genome-wide decay based on 756,005 biallelic SNPs and panel (B) illustrates the LD decay in the five chromosome regions in which candidate genes were identified. LD had decayed at 50 kb to 0.008, 0.012, 0.018, 0.023 and 0.028 in regions I, II, III, IV and V, respectively.