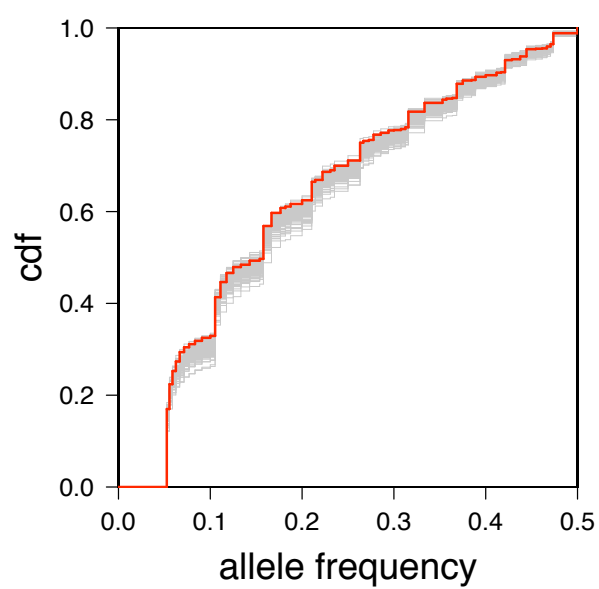


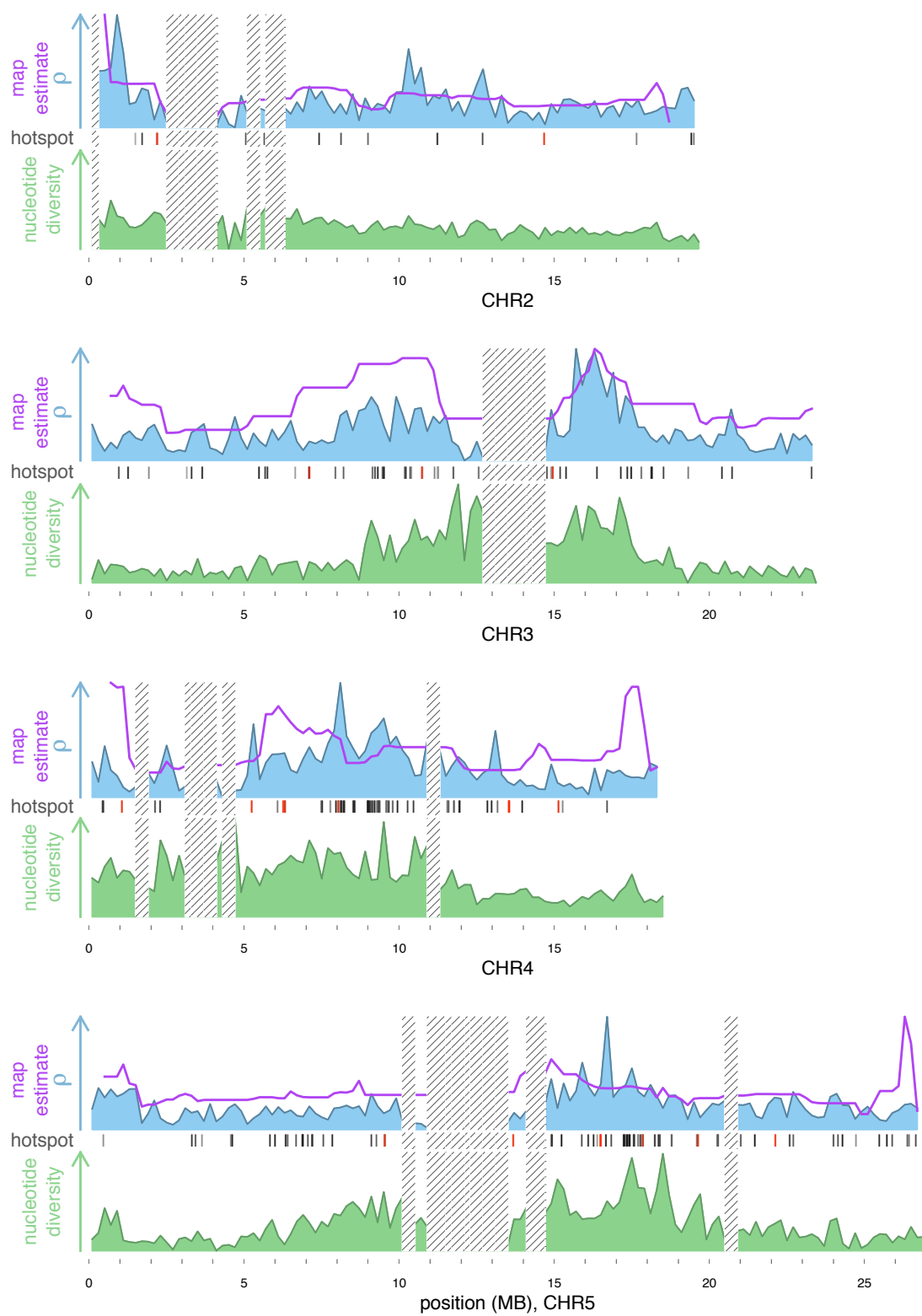
# **Recombination and linkage disequilibrium in *Arabidopsis thaliana***

## *Supplementary Material*

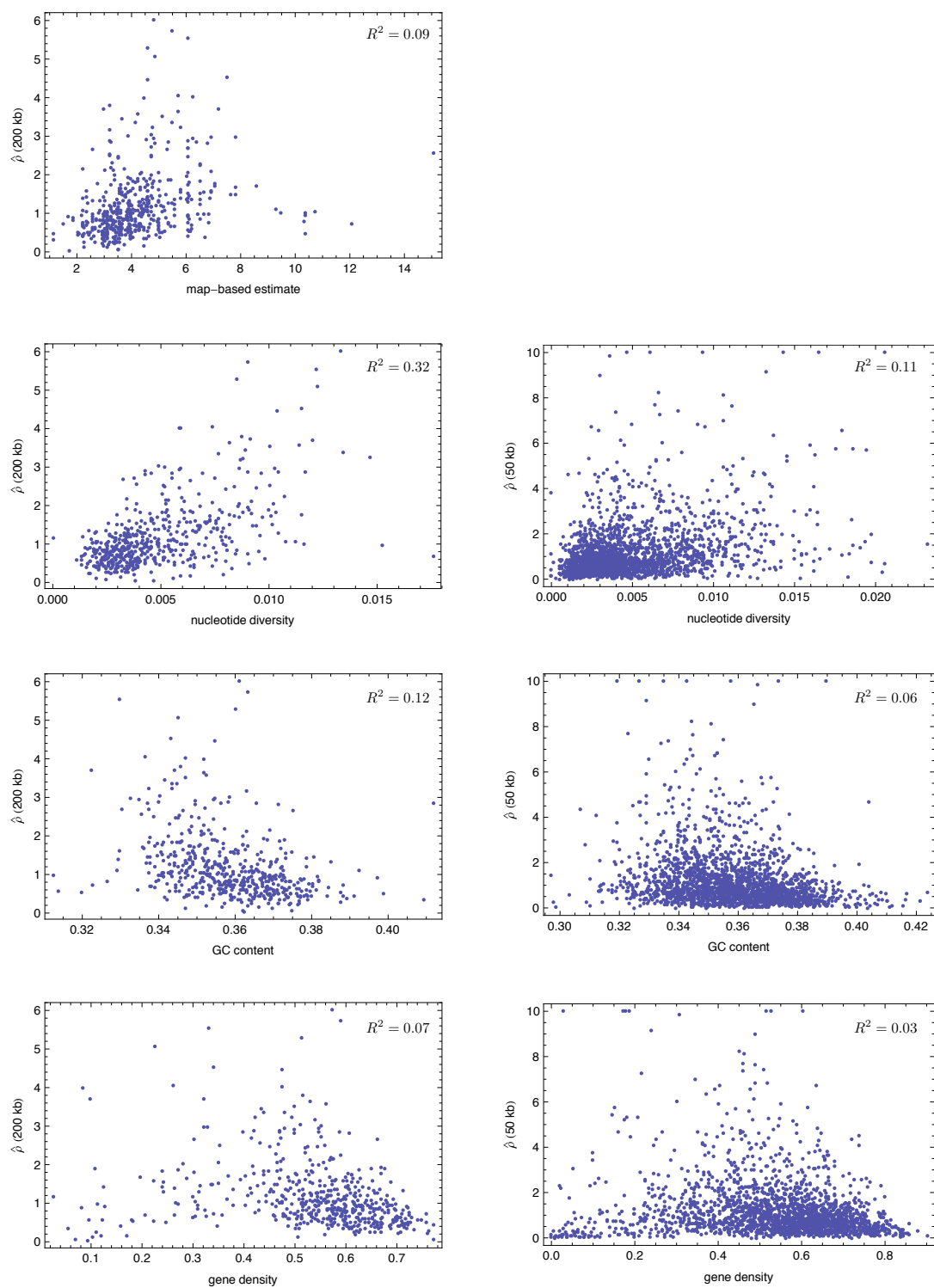
Sung Kim, Vincent Plagnol, Tina T. Hu, Christopher Toomajian, Richard M. Clark,  
Stephan Ossowski, Joseph R. Ecker, Detlef Weigel, Magnus Nordborg



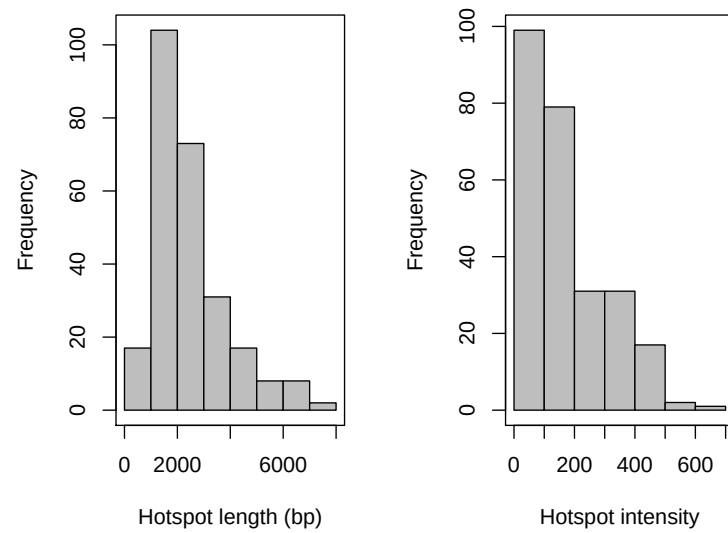
**Supplementary Figure 1** Minor allele frequency distribution in the array sample and random samples. The cumulative densities of minor allele frequencies of dideoxy SNPs in the 19 array accessions (red), and for random subsamples (grey) of size 19.



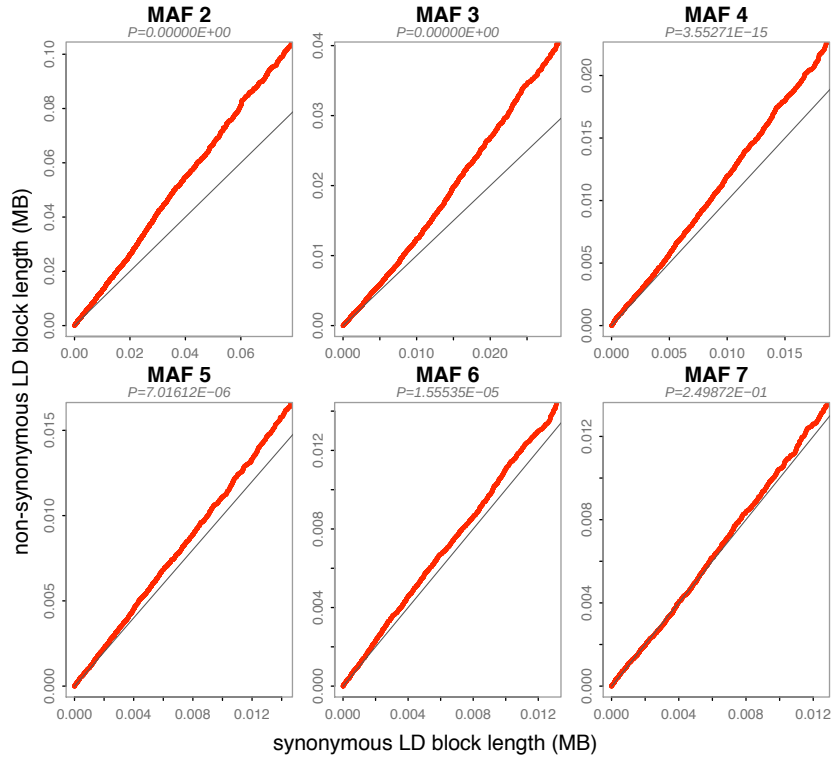
**Supplementary Figure 2** Recombination rate variation along chromosomes 2 to 5 (cf. Fig. 2 for chromosome 1).



**Supplementary Figure 3** Scatterplots illustrating the correlation (or lack thereof) between recombination estimates and genomic features in 200 kb or 50 kb windows.



**Supplementary Figure 4** Distribution of the length and intensity of hotspots.



**Supplementary Figure 5** Quantile-quantile plots of synonymous and non-synonymous LD block lengths, with p-values calculated using the Kolmogorov-Smirnov test. Each plot corresponds to a particular minor allele frequency (maf) at the focal polymorphism: conditioning on allele frequencies is necessary to avoid confounding due to the fact that allele frequency distributions differ between synonymous and non-synonymous sites.

**Supplementary Table 1** Bin distribution<sup>a</sup>.

| Size  | Avg. length | Number of bins | Number of SNPs |
|-------|-------------|----------------|----------------|
| 1     | 0 bp        | 101,039        | 101,039        |
| 2-5   | 5,182 bp    | 27,897         | 79,711         |
| 6-10  | 10,908 bp   | 6,221          | 46,778         |
| 11-20 | 12,479 bp   | 3,295          | 47,001         |
| 21-30 | 13,262 bp   | 984            | 24,255         |
| >30   | 14,679 bp   | 915            | 42,818         |
| Total |             | 140,351        | 341,602        |

<sup>a</sup>Generated by LDBIN with a 20 kb window and an  $r^2$  cut-off of 0.8

## Supplementary Methods

**Tag-SNP selection.** We applied four different tag-SNP selection algorithms to investigate how many SNPs would be required to capture all the 341,602 non-singleton polymorphisms in the array data (**Table 1**).

The first algorithm, LDBIN<sup>1,2</sup>, greedily selects SNPs such that each has an  $r^2$  greater than some threshold to a collection (or “bin”) of SNPs within a maximum window size. We chose a window size of 20 kb to obtain a false discovery rate threshold of 5% estimated as described in<sup>1</sup>, and an  $r^2$  threshold of 0.8. The resulting bin distribution is shown in **Supplementary Table ??** online. Each bin of size greater than 1 is a collection of SNPs spanning no more than 20 kb, such that one member has an  $r^2$  greater than 0.8 to all the others. Bins of size one contain SNPs that could not be captured by any other SNP. These results appeared relatively robust to the choice of maximum window length and  $r^2$  cut-off. Reducing the maximum window length to 10 kb increased the number of size 1 bins by approximately 4%, but were otherwise very comparable. Increasing the  $r^2$  threshold to 1 produced nearly identical results (not shown).

The second algorithm, TAGGER<sup>3</sup>, implemented in Haploview<sup>4</sup>, is similar to LDBIN, but each tag-SNP can be replaced by a multi-marker predictor providing a multi-staged SNP selection approach. We used a window size of 20 kb, an  $r^2$  threshold of 0.8, and forced all markers to be tagged using the “aggressiveTagging” option for analysis.

The third algorithm, HAPBLOCK<sup>5</sup>, implemented in HapBlock<sup>6</sup>, determines the minimum SNP set that explains 80% of the observed haplotype diversity within a “haplotype block” as defined by<sup>7</sup>. All parameters used in the analyses were “default” values.

The last and most conservative algorithm, which we refer to as CORRBIN, simply removes all adjacent markers more strongly correlated than some  $r^2$  threshold. The algorithm greedily selects SNPs that have the largest number of adjacent markers with  $r^2$  greater than the threshold (we used 0.8) and repeats until no further SNPs can be tagged.

**Genetic maps.** We generated a genetic map using  $F_2$  data obtained by crossing Col-0 with seven different accessions: RRS-10, Kno-18, Lov-1, Ull2-5, Var2-6, Edi-0<sup>8</sup> and Pu2-8 (Toomajian et al., unpublished). The data consists of 185  $F_2$  offspring per cross (90  $F_2$  for Pu2-8) for a total of 1,200 meioses in the 120 Mb genome, comparable to the best available human map (1,257 meioses for a 3 Gb genome)<sup>9</sup>.

However, the marker density is significantly lower (on average one marker per 2 Mb compared to one marker every 500 kb in human) and the set of markers is not the same for all crosses. To combine these data we used the following approach: First, we estimated a genetic map separately for each of the crosses using the R/qtl software<sup>10</sup>. Then we combined the crosses using a weighted average (using weights proportional to the number of offspring in each cross). The recombination rate between two markers for each cross were assumed to be constant.

**Correlates of  $\rho$ .** We used a standard linear regression to determine the correlation between genomic features and recombination estimates in windows of different sizes. First, we tested each genomic feature separately. Second, we used a forward regression where at each step the most significant of the remaining co-factors was added. The co-factors we considered were: (1) polymorphism levels defined as the average number of

pairwise difference at 4-fold degenerate coding sites<sup>11</sup>; (2) GC content, i.e., the percentage of GC base pairs in the window, based on the reference genome; (3) gene density, calculated as the fraction of introns and exons in that window (based on the reference genome); (4) repeats element fraction (TAIR annotation, version 6).

## References

1. Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A., and Cox, D. R. Whole-genome patterns of common DNA variation in three human populations. *Science*, 307:1072–1079, 2004.
2. Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L., and Nickerson, D. A. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*, 74:106–120, 2004.
3. de Bakker, P. I. W., Yelensky, R., Pe’er, I., Gabriel, S. B., Daly, M. J., and Altshuler, D. Efficiency and power in genetic association studies. *Nature Genet.*, 37:1217–23, 2005.
4. Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21(2):263–265, 2005.
5. Johnson, G. C., Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., Genova, G. D., Ueda, H., Cordell, H. J., Eaves, I. A., Dudbridge, F., Twells, R. C., Payne, F., Hughes, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tuomilehto, J., Gough, S. C., Clayton, D. G., and Todd, J. A. Haplotype tagging for the identification of common disease genes. *Nature Genet.*, 29:233–237, 2001.
6. Zhang, K., Qin, Z., Chen, T., Liu, J. S., Waterman, M. S., and Sun, F. Hapblock: haplotype block partitioning and tag snp selection software using a set of dynamic programming algorithms. *Bioinformatics*, 21:131–34, 2005.
7. Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., Nguyen, B. T., Norris, M. C., Sheehan, J. B., Shen, N., Stern, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P., and Cox, D. R. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294:1719–1732, 2001.
8. Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajian, C., Zheng, H., Dean, C., Marjoram, P., and Nordborg, M. An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet.*, 3:e4, 2007.
9. Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S. T., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R., and Stefansson,

- K. A high-resolution recombination map of the human genome. *Nature Genet.*, 31: 241–247, 2002.
10. Broman, K., Wu, H., Sen, S., and Churchill, G. R/qtl: Qtl mapping in experimental crosses. *Bioinformatics*, 19:889–890, 2003.
  11. Clark, R. M., Schweikert, G., Ossowski, S., Zeller, G., Toomajian, C., Shinn, P., Warthmann, N., Hu, T. T., Fu, G., Hinds, D. A., Chen, H., Frazer, K. A., Huson, D. H., Schölkopf, B., Nordborg, M., Rättsch, G., Ecker, J. R., and Weigel, D. Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. To appear in *Science*, 2007.