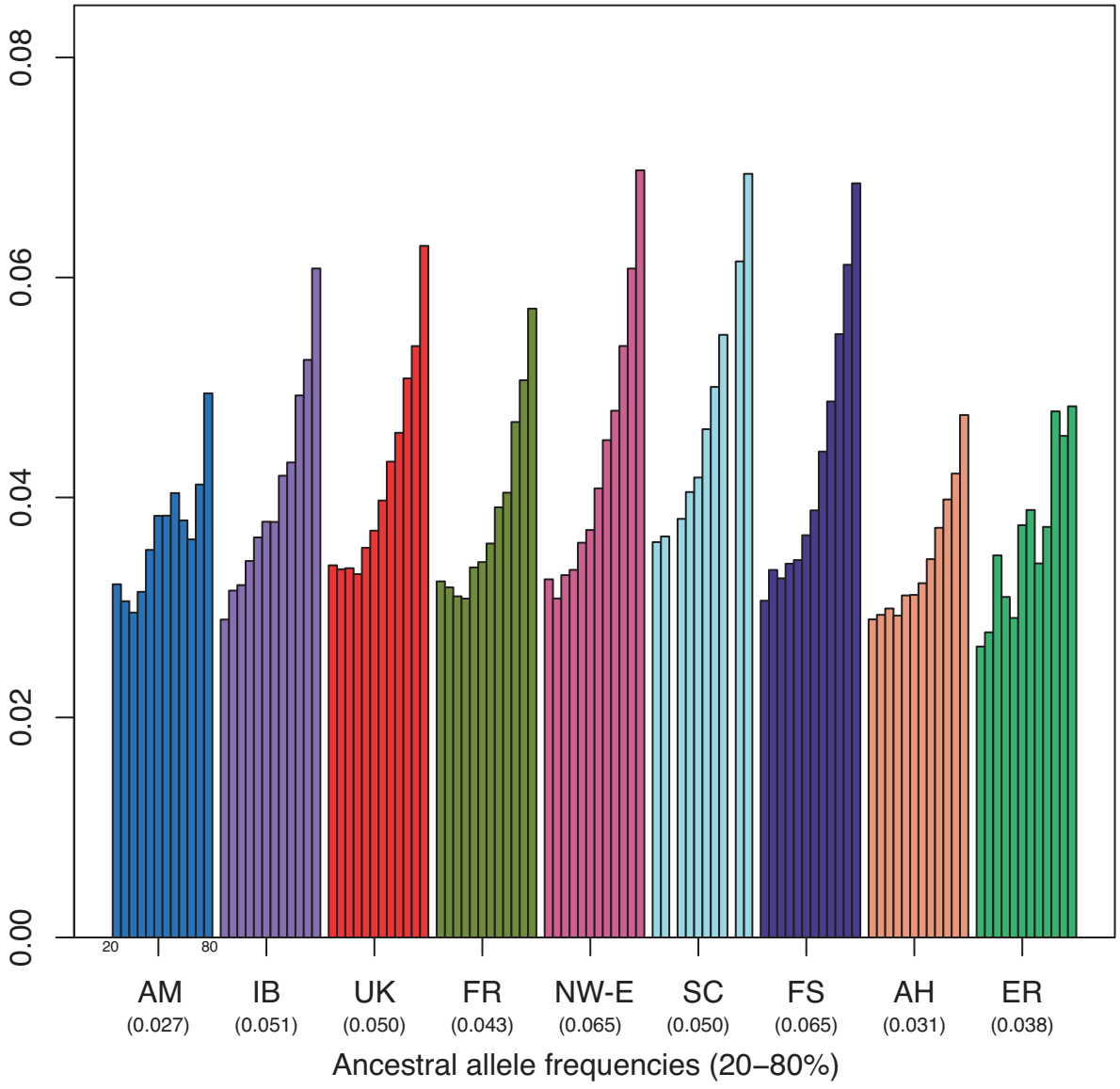
Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel.

Supplementary Material.

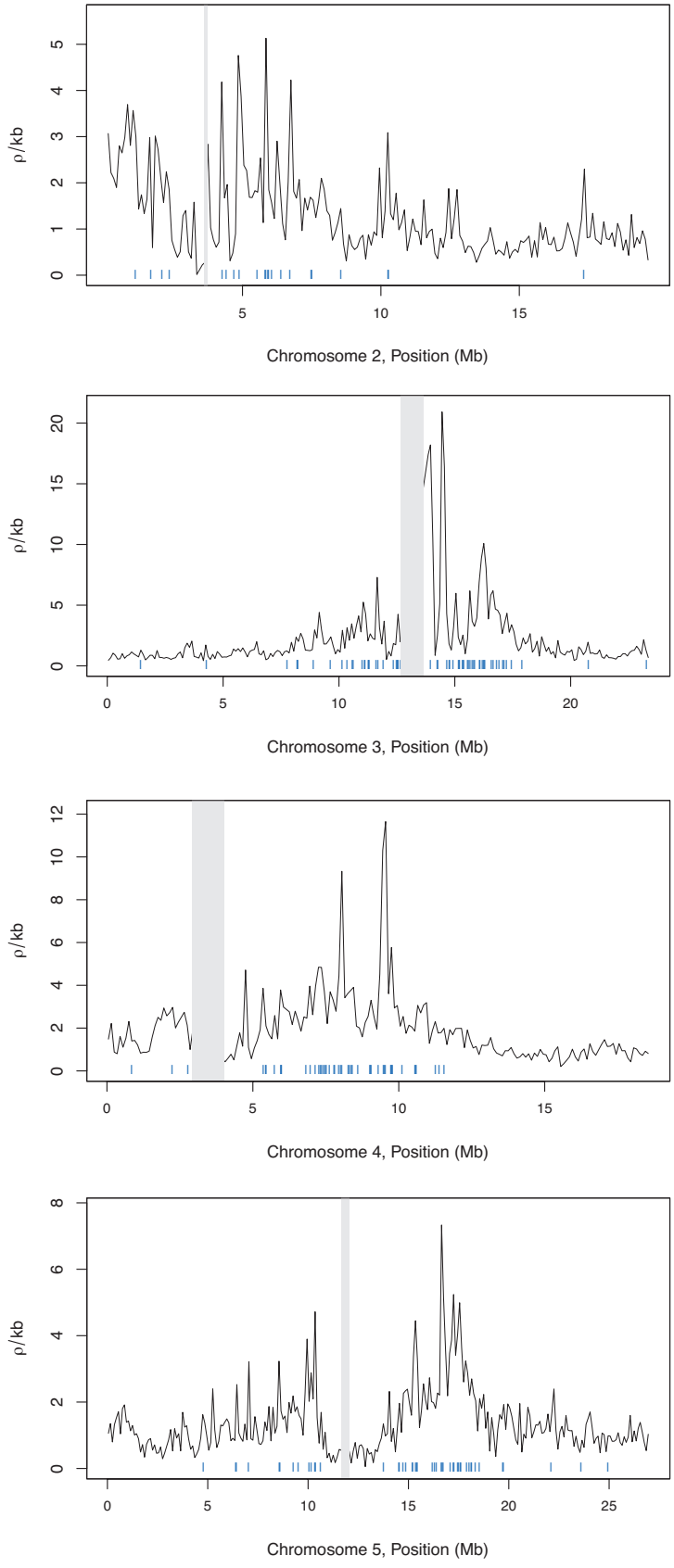
Matthew W. Horton, Angela M. Hancock, Yu S. Huang, Christopher Toomajian, Susanna Atwell, Adam Auton, N. Wayan Muliyati, Alexander Platt, F. Gianluca Sperone, Bjarni J. Vilhjálmsson, Magnus Nordborg, Justin O. Borevitz, Joy Bergelson

This file includes:

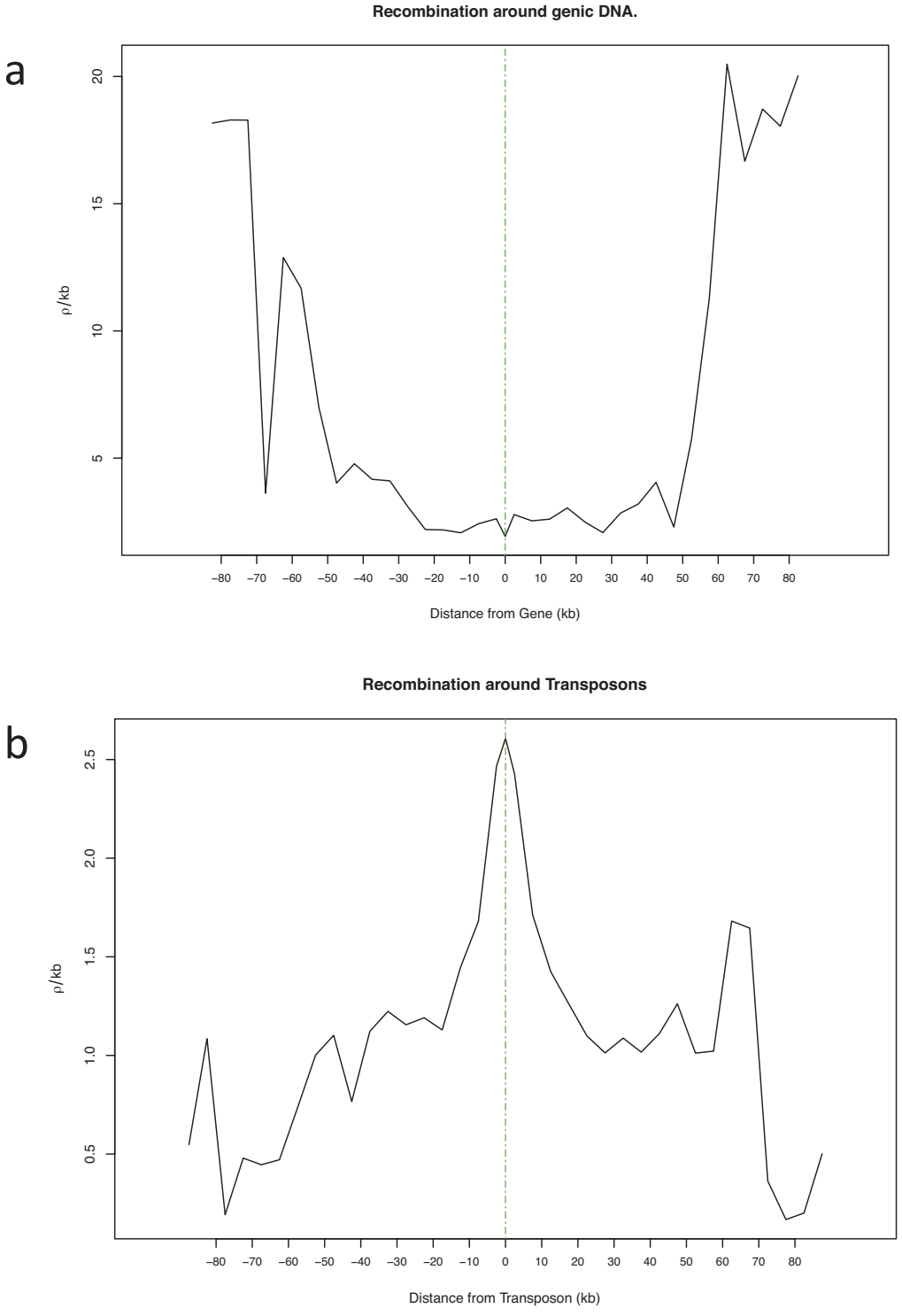
Supplementary Figures.



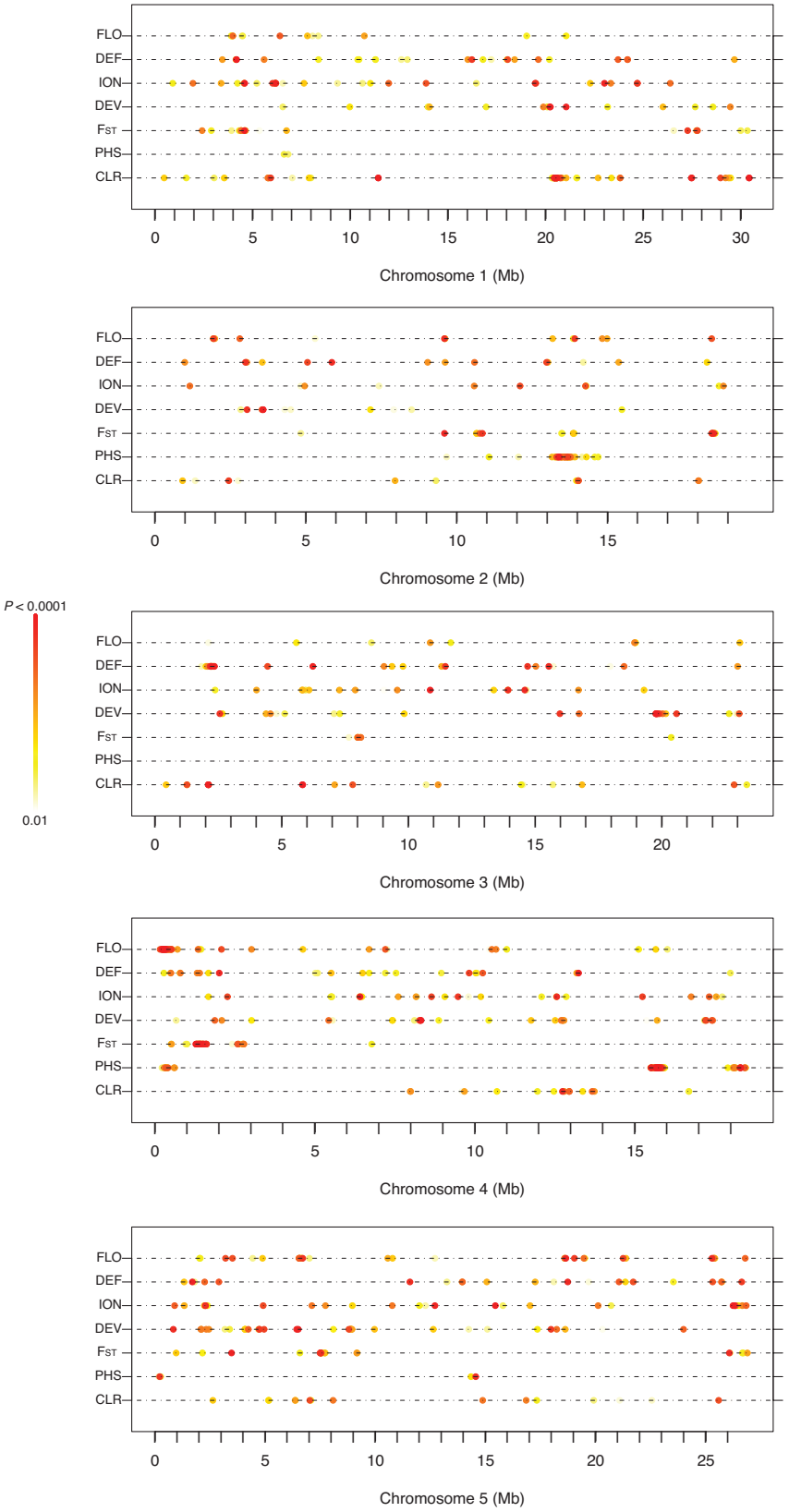
Supplementary Figure 1. Ancestral allele frequency spectra for 9 regional samples. AM: *the Americas* (Canada, United States); IB: *Iberia* (Portugal, Spain); UK: *British-Isles*; FR: *France*; NW-E: *North-West Europe* (Belgium, Netherlands, Denmark, Germany, Poland); SC: *South-Central* (Switzerland, Italy); FS: *Fennoscandia* (Norway, Sweden, Finland); AH: *Austria-Hungary* (Austria, Czech Republic, Romania); ER: *Eastern-Range* (Estonia, Lithuania, Belarus, Ukraine, Georgia, Azerbaijan, Russia, Tajikistan, Kashmir, Kazakhstan). The estimate for the slope is listed below the population label.



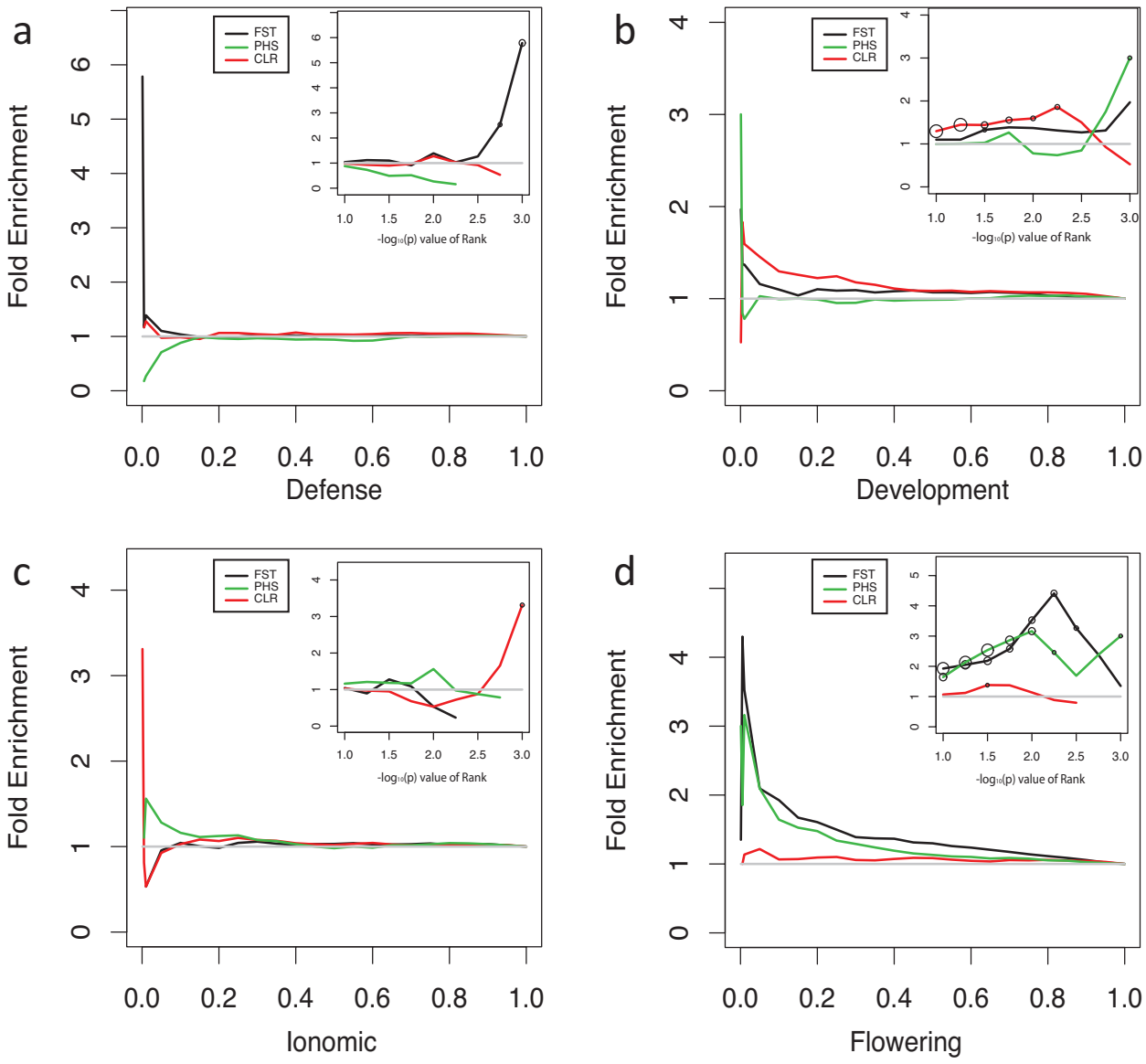
Supplementary Figure 2. Recombination rate estimates for chromosomes 2-5. cf Figure 2. Estimates were smoothed using 100-kb windows.



Supplementary Figure 3. Recombination rate estimates surrounding genomic features. Average recombination rate increases with increasing distance from the nearest gene (a) and decreases moving away from transposable elements (b). Distances are the midpoint of SNP-intervals, 5’ or 3’ from the focal point. Estimates were smoothed using 5-kb windows.



Supplementary Figure 4. Overlap in signals of selection with the top results from GWAS of 107 phenotypes. GWAS results (Atwell et al., 2010; Methods) were separated into four phenotypic categories either related to flowering (FLO), defense (DEF), ionomics (ION) or development (DEV). The top 1% of ranked 10-kb windows are shown for each test.



Supplementary Figure 5. Enrichment analyses of GWAS results with selection scans. Enrichment of windows containing SNPs associated with 107 phenotypes (GWAS p-values less than $1x10^{-4}$) across the distributions of the selection scan results. The inset shows the extreme tails of the selection scan distributions (10% - 0.1% of ranks) as a $-\log_{10}p$ value. The sizes of the circles denote significance based on 1000 permutations. Shown are phenotypes related to defense (a), development (b), ionomic concentration (c) and flowering-time (d). For defense, the circles correspond to p=0.034 and p=0.006 for the smallest and largest circles, respectively. Development: p=0.047 and p=0.001. Ionomics: p=0.034. Flowering-time: p=0.049 and p=0.001.

Supplementary Tables

Supplementary Table 1. Distribution of simple sequence repeats (SSRs) in recombination hotspots (RR > 1) and coldspots (RR < 1). Bonferonni correction was performed using the number of total tests applied. Shown are the motifs for which the corrected P < 0.05, and for which the count of the motif is n > 1.

Candidate Hot-motifs.

Motif	Number overlapping hotspots.	Number overlapping coldspots.	Relative-Risk	Bonferonni corrected p-value
AAAAT	180	49	3.67	4.28E-17
AAAT	423	231	1.83	2.63E-12
AATT	198	84	2.36	4.38E-10
ACG	347	211	1.64	4.57E-07
AAATT	64	20	3.20	7.77E-05
AAAAAG	32	4	8.00	9.51E-05
AATC	177	105	1.69	0.001
CCG	230	149	1.54	0.0018
AAAATT	14	1	14.00	0.0478

Candidate Cold-motifs.

Motif	Number overlapping hotspots.	Number overlapping coldspots.	Relative-Risk	Bonferonni corrected p-value
AGG	484	614	0.79	0.0047
AGC	427	545	0.78	0.0084
AAGC	65	116	0.56	0.0091
AAAC	291	385	0.76	0.0166

Supplementary Table 2. The results from an exhaustive motif search (5-9 bp) in non-repetitive (TE or pseudogene) DNA. Adenine-rich microsatellites are overrepresented in hotspots of recombination. Significance was assessed through Fisher's Exact Test and Bonferonni adjusted. The top 5 candidates are shown for each motif size; results are sorted based on the difference between the count of the motif in hot and in coldspots.

Candidate Hot-motifs.

Length	Motif	Number in hotspots	Number in coldspots	Diff	RR	Bonferonni corrected p-values
9	AAAAAAAAA	471	252	219	2.30	7.96E-13
9	ATTTTTTTT	315	172	143	2.26	1.98E-08
9	AAAAAAAGA	240	99	141	2.99	2.58E-12
9	AAAAAAAAG	272	132	140	2.54	7.56E-10
9	CAAAAAAAA	340	204	136	2.05	1.45E-06
8	AAAAAAAA	610	411	199	1.83	1.31E-07
8	ATTTTTTT	509	323	186	1.94	1.88E-08
8	TTTATTTT	451	274	177	2.03	5.77E-09
8	TTTATATA	302	138	164	2.70	3.16E-13
8	AATATATA	315	152	163	2.55	3.84E-12
7	TAATTAA	480	222	258	2.66	2.23E-19
7	AATTAAA	537	344	193	1.92	4.12E-09
7	TAAAATA	538	347	191	1.91	6.25E-09
7	ATAATTA	434	266	168	2.01	6.45E-09
7	TATTAAA	435	270	165	1.99	1.09E-08
6	GTCGAG	264	121	143	2.69	9.03E-13
6	TACTCG	275	148	127	2.29	1.97E-09
6	CGCCGT	187	77	110	2.99	3.59E-11
6	ACGACG	266	169	97	1.94	1.29E-05
6	CGACGA	317	222	95	1.76	0.0002
5	CGACG	562	462	100	1.50	0.0015
5	GTACG	563	472	91	1.47	0.0043
5	CCGCG	295	209	86	1.74	0.0001
5	CGCGT	358	286	72	1.54	0.0061
5	CGCCG	429	361	68	1.46	0.0202

Supplementary Note

Population Structure in *A. thaliana*.

We examined the population structure of this sample using principal components analysis (PCA). In order to minimize artifacts due to linkage disequilibrium[1,2], we filtered our genome-wide SNP data using PLINK[3] to exclude SNPs in high pairwise linkage disequilibrium (r^2 > 80%). PCA was performed on the remaining 165,579 SNPs using the software smartpca[2].

Overall, PCA distinguishes our regional samples and provides high-level inferences, but patterns that are consistent with earlier analyses[9, 46]. Genetic admixture among samples is often illustrated by straight lines in plots of principal components[2], and Figure1a is suggestive of admixture between the American sample and Western-Europe/British-Isles and separately, Northern Sweden with Central Sweden. The fine-scale pattern of population structure is more easily discerned in a PCA analysis of the native range of *A. thaliana* (Figure 1b). However, it is clear that PCA is susceptible to our irregular sampling scheme[4], leading to some differences between these PC plots and the one that might be expected based on the geographic origin of individual accessions. Based on our collection strategy and these results, we separated our panel into 9 regional samples; correcting for sample size differences among these regions removes most but not all of the overlap among them (results not shown). For example, most of the accessions from the British-Isles project in PC-space closest to accessions collected from France; however, as noted earlier[11] a fraction of this sample clusters with lines from the Nordic countries and is consistent with a model of different routes of migration into the British-Isles.

Based on these analyses, we split our samples into 9 regional groups: **(1)** *the Americas*, or Canada and the United States (n = 183); **(2)** *Iberia*, or Portugal and Spain (n = 28); **(3)** *France* (n = 204); **(4)** *the British-Isles* (n = 171); **(5)** *NW-Europe*, or Belgium, Netherlands, Denmark, Germany and Poland (n = 92); **(6)** *South-Central*, or Switzerland and Italy (n = 17); **(7)** *Austria-Hungary*, or Austria, the Czech Republic and Romania (n = 155); **(8)** *Fennoscandia*, or Norway, Sweden, Finland (n = 303); and **(9)** *the Eastern-Range*, or Estonia, Lithuania, Belarus, Ukraine, Georgia, Azerbaijan, Russia, Tajikistan, Kashmir and Kazakhstan (n = 26). This omits accessions from Cape Verde, Libya and New Zealand.

Ancestral Allele Frequency (AAF) Spectra of 9 regional samples.

Next, we used the genome of *A. lyrata*[18] to help determine the ancestral allele frequency (AAF) spectrum for each of these 9 subsamples (Supplementary Fig. 1). Because the distribution of SNPs in the AAF spectrum is heavily influenced by demography and selection[5-8], the AAF spectrum has the potential to offer insights into the history of individual samples. To describe the AAF spectra for our 9 samples we corrected for sample size differences[9]; the smallest sized sample includes lines from Switzerland and Italy (n = 17). We resampled 17 individuals, without replacement, from each geographic region to determine these spectra. We then estimated the slope of these distributions in the midrange (20-80%) of each spectrum[10] to minimize the impact of rare alleles and selection on inference. We note that the AAF spectrum is also influenced by ascertainment; however, the SNP chip was designed using accessions from each geographic region except France and South-Central (Switzerland & Italy). Samples near the center of our panel, including North-West Europe (NW-E), South-Central Europe (SC), and Fennoscandia (FS) possess the steepest slopes. These samples contain a higher proportion of SNPs in the ancestral state relative to the other, peripheral, samples. The sample from North America, where *A. thaliana* seems to be introduced[11], contains more high frequency derived alleles than the other regions and thus its AAF spectrum exhibits the flattest slope (~0.027).

Small populations are more susceptible to genetic drift than large (or expanding) populations, and are therefore more likely to undergo increases in derived allele frequencies. In that context, the American AAF spectrum is consistent with a population bottleneck. Cao et al. (2011) measured the ratio of deleterious mutations to tolerated mutations, and argued for a recent bottleneck in a region overlapping with our sample, the 'Eastern-Range'; our results are also consistent with a population bottleneck in this region (slope ~ 0.038). The AAF spectrums for samples in the center of the species distribution (NW-E, SC and FS) suggest these populations have either maintained large population sizes or have experienced population growth. We note, however, that our samples are defined coarsely, and non-randomly distributed. More intense sampling in the regions peripheral to the sampling area considered here should further elucidate the global pattern of diversity across the range of *A. thaliana*.

References

1. Novembre, J. et al. Genes mirror geography within Europe. *Nature* **456**, 98-101 (2008).
2. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006).
3. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
4. McVean, G. A genealogical interpretation of principal components analysis. *PLoS Genet* **5**, e1000686 (2009).
5. Przeworski, M. The signature of positive selection at randomly chosen loci. *Genetics* **160**, 1179-89 (2002).
6. Kimura, M. & Ota, T. The age of a neutral mutant persisting in a finite population. *Genetics* **75**, 199-212 (1973).
7. Fay, J.C. & Wu, C.I. Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405-13 (2000).
8. Marth, G.T., Czabarka, E., Murvai, J. & Sherry, S.T. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166**, 351-72 (2004).
9. Kalinowski, S.T. Counting alleles with rarefaction: Private alleles and hierarchical sampling designs. *Conservation Genetics* **5**, 539-543 (2004).
10. Li, J.Z. et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100-4 (2008).
11. Okane, S.L. & AlShehbaz, I.A. A synopsis of Arabidopsis (Brassicaceae). *Novon* **7**, 323-327 (1997).