nature genetics

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

Matthew W Horton[1], Angela M Hancock[1,6], Yu S Huang[2,6], Christopher Toomajian[3,6], Susanna Atwell[2], Adam Auton[4], N Wayan Muliyati[1], Alexander Platt[2], F Gianluca Sperone[1], Bjarni J Vilhjálmsson[2], Magnus Nordborg[2,5], Justin O Borevitz[1] & Joy Bergelson[1]

© 2012 Nature America, Inc. All rights reserved.

Arabidopsis thaliana is native to Eurasia and is naturalized across the world. Its ability to be easily propagated and its high phenotypic variability make it an ideal model system for functional, ecological and evolutionary genetics. To date, analyses of the natural genetic variation of _A. thaliana_ have involved small numbers of individual plants or genetic markers. Here we genotype 1,307 worldwide accessions, including several regional samples, using a 250K SNP chip. This allowed us to produce a high-resolution description of the global pattern of genetic variation. We applied three complementary selection tests and identified new targets of selection. Further, we characterized the pattern of historical recombination in _A. thaliana_ and observed an enrichment of hotspots in its intergenic regions and repetitive DNA, which is consistent with the pattern that is observed for humans but which is strikingly different from that observed in other plant species. We have made the seeds we used to produce this Regional Mapping (RegMap) panel publicly available. This panel comprises one of the largest genomic mapping resources currently available for global natural isolates of a non-human species.

A. thaliana occupies a wide range of habitats, including beaches, rocky slopes, riverbanks, roadsides and the periphery of agricultural areas. It has been collected in the Americas, New Zealand, the mountains of east Africa, on islands in both the Pacific and Atlantic Oceans and throughout its native range in Eurasia. Its wide species distribution and rich genetic resources, combined with the fact that it can be maintained as pure lines, make *A. thaliana* an attractive resource for investigating the molecular and genetic bases of ecologically relevant traits. The feasibility of genome-wide association studies (GWAS)[1–4] using these accessions facilitates the dissection of natural variation within this species and adds to its value as a model system. However, attempts to examine the genomic pattern of recombination and selection in *A. thaliana* have so far been limited by sample size[5,6]. To consider the global pattern of genetic diversity in *A. thaliana*, we genotyped 1,307 accessions collected from around the world (see URLs for the project website) on a 250K SNP chip[1,5] (Online Methods). The use of large regional populations allowed us to identify new candidate targets of selection, including the most differentiated regions in the *A. thaliana* genome. We also characterized the pattern of recombination among these samples and searched for the genetic factors that are associated with recombination hotspots.

Our samples comprised large regional panels and smaller samples combined after fine-scale principal components analysis (PCA)[7] (**Fig. 1** and **Supplementary Note**). After correcting for sample size differences among geographic regions, the population structure we found was consistent with earlier analyses[8,9]. Further analysis of the ancestral allele frequency spectrum suggested that the central populations that we examined maintained larger population sizes than the peripheral populations (**Supplementary Fig. 1**), supporting the hypothesis that the populations from these marginal areas experienced population bottlenecks[10,11].

We evaluated the historical pattern of recombination in each of the samples by estimating the population-scaled recombination rate (ρ) across the genome[12]. We then averaged these estimates to make a fine-scale genetic map for each of the five chromosomes. We found strong statistical support for recombination hotspots ($P < 0.01$, $n = 4{,}427$; $P < 0.0001$, $n = 1{,}606$) throughout the genome (**Fig. 2** and **Supplementary Fig. 2**).

An earlier study showed a deficit of genic DNA in the recombination hotspots compared to the genomic background in *A. thaliana*, but the authors of that study did not observe an association of hotspots with specific gene classes or repeat content[5]. We found that recombination rates tended to be higher within transposable elements than in the neighboring DNA, and we found the opposite to be true around genes (**Supplementary Fig. 3**). We also found an enrichment of hotspots in pseudogenes (**Fig. 3**). Among the genes showing the highest historical recombination rates were members of repeat families, including loci involved in self-incompatibility (SI) and disease resistance (R-genes). Recombination-mediated amplification of genes can facilitate adaptive evolution by increasing

[1]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, USA. [2]Molecular and Computational Biology, University of Southern California, Los Angeles, California, USA. [3]Department of Plant Pathology, Kansas State University, Manhattan, Kansas, USA. [4]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [5]Gregor Mendel Institute, Austrian Academy of Sciences, Vienna, Austria. [6]These authors contributed equally to this work. Correspondence should be addressed to J.B. (jbergels@uchicago.edu).

Received 2 September 2011; accepted 29 November 2011; published online 8 January 2012; doi:10.1038/ng.1042

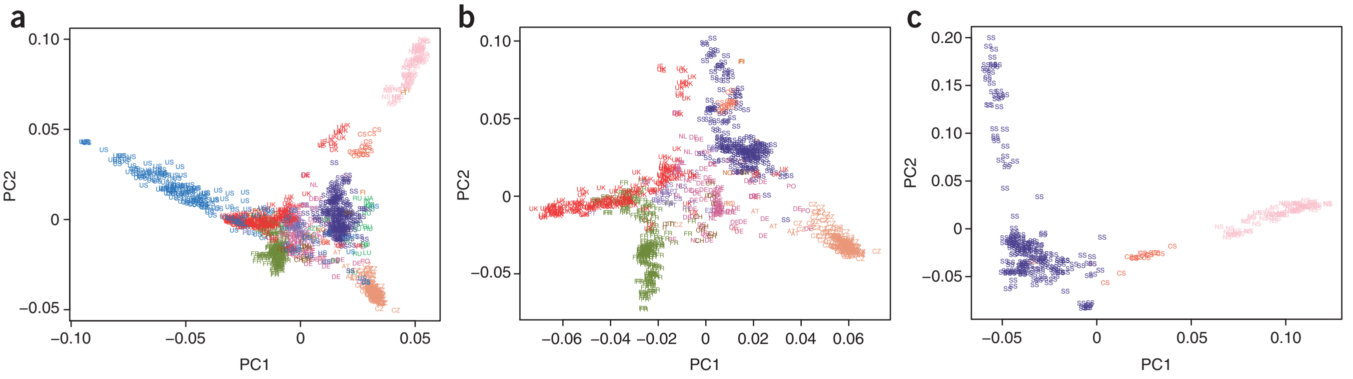

Figure 1 PCA of the study samples. (**a**) A plot of principal component 1 (PC1) and principal component 2 (PC2) distinguishes between the largest samples, although some overlap is evident in the center of the distribution. (**b**) A plot similar to that shown in **a** depicting the core collection area, excluding the Americas and northern Sweden. (**c**) The top two components from a PCA of Fennoscandia, which was the largest regional sample. PT, Portugal; ES, Spain; IT, Italy; CH, Switzerland; BE, Belgium; NL, Netherlands; DK, Denmark; DE, Germany; PO, Poland; NO, Norway; FI, Finland; AT, Austria; CZ, Czech Republic; RO, Romania; EE, Estonia; LT, Lithuania; BY, Belarus; UA, Ukraine; GE, Georgia; AZE, Azerbaijan; RU, Russia; TJ, Tajikistan; KS, Kashmir. Sweden is separated into southern (SS), central (CS) and northern Sweden (NS).

© 2012 Nature America, Inc. All rights reserved.

the mutational target size[13], by permitting neofunctionalization or subfunctionalization or by generating chimerical genes[14,15].

To further investigate the genetic determinants of recombination in *A. thaliana*, we matched hotspots with comparable ‘coldspots’ (Online Methods) and asked whether particular sequences or DNA classes were overrepresented in hotspot regions. Simple sequence repeats enriched in hotspots include $((A)_xT)_n$ $(3 \le x \le 4)$, $(AATT)_n$, $(AAATT)_n$ and $(ACG)_n$ ($n > 1$; **Supplementary Table 1**). MuDR transposons, which increase the frequency of meiotic recombination in maize[16], are overrepresented in hotspots in *A. thaliana* (relative risk (RR) = 3.2), as are copia elements (RR = 11.4) and helitrons (RR = 2.1). There is some evidence to suggest that helitrons undergo non-allelic homologous recombination[17], which is one of the processes believed to be responsible for widespread deletions and the apparent shrinkage of the genome of *A. thaliana*[18]. We note, however, that repetitive regions may contain unseen structural variants (sometimes arising from non-allelic homologous recombination itself), which would lead to imperfect estimates of recombination with these classes of DNA.

A common sequence motif (CCNCCNTNNCCNC) is associated with ~40% of recombination hotspots in humans[19]. To determine whether hotspots in *A. thaliana* are also associated with a highly recombinogenic motif, we counted the frequency of all the nucleotide motifs, ranging in lengths from 5–9 bp, in hotspots and coldspots that do not overlap transposable elements or pseudogenes (Online Methods). The strongest scoring motifs included the 9-mer AAAAAAAAA ($(A)_9$), motifs related to $(A)_9$ and several other adenine-rich microsatellites (**Supplementary Table 2**). The $(A)_9$ motif is also among the top (9-bp) candidate motifs identified in humans[20].

In other plant species, including wheat and maize, recombination seems to occur predominantly in gene-rich[21] and intragenic regions[22]. In comparison, the pattern of recombination in *A. thaliana* is more similar to that seen in humans[20], in which recombination is enriched in intergenic regions. Among other factors, *A. thaliana* has a higher proportion of microsatellites per megabase compared to maize and wheat[23], and the recombination landscape may, in part, reflect this. Notably, microsatellites tend to be located outside of genes, with the fraction of microsatellite DNA in *A. thaliana* (and humans) estimated to be ~2–3 times that in maize[23]. Perhaps more importantly, the genome of *A. thaliana* has a higher proportion of single-copy DNA than maize. There seem to have been several thousand deletions in the genome of *A. thaliana* since its divergence from *A. lyrata* (~10 million years ago (MYA)), mostly in the repetitive DNA[18], and our results implicate recombination as a contributing mechanism of these deletions. The maize genome, in contrast, has recently expanded in size through a tetraploidy event (5–12 MYA)[24,25] and a more recent transposon ‘bloom’ (~3 MYA)[26]; maize transposons are often hypermethylated and, consequently, recombinationally inert[27]. Patterns of recombination differ widely even among closely related species[28], and studies in additional taxa will help clarify the roles that individual genomic features have in shaping crossover events.

A. thaliana is found in a wide variety of habitats, and it is likely that adaptation to the environment has been a crucial part of its evolutionary past[29,30]. We investigated the molecular basis of adaptation in *A. thaliana* by scanning its genome for signatures of selection using three complementary approaches. Two of these methods are designed to identify signatures of classic ‘hard’ selective sweeps[31]. The first method, the pairwise haplotype sharing (PHS) test, identifies regions in which there is evidence of extended haplotype homozygosity; PHS analyses have power to detect partial or ongoing sweeps[32]. The second method, the composite likelihood ratio test of the allele frequency spectrum (CLR), has the power to detect complete or nearly complete sweeps[33]. In addition, we calculated Wright’s fixation index (F_{ST}) for all SNPs; F_{ST} distinguishes genomic regions based on broad-scale population differentiation[34] and makes

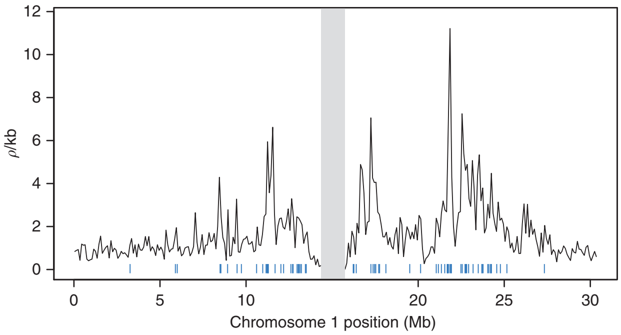

Figure 2 Recombination rate variation for chromosome 1. Shown are estimates of recombination rates in 100-kb windows (black) and the location of hotspots (blue; the centromere, which was excluded, is shown in gray). The hotspots shown were identified in at least eight of the nine regional samples (ρ/kb > 3).

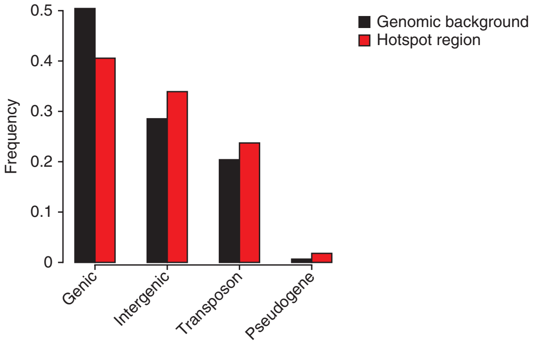

Figure 3 The proportion of DNA within and outside of inferred hotspots. After separating hotspot regions from the genomic background, a deficit of genic DNA in hotspot regions and a strong enrichment for DNA classified as either intergenic, transposon or pseudogene can be seen.

no assumptions as to whether selection happened on new or on standing (previously existing) genetic variation.

Several of the most differentiated SNPs are in or near flowering-related genes such as *SHORT VEGETATIVE PHASE* (*SVP*), a MADS box gene that negatively regulates the transition to flowering[35]. *SVP* was previously identified in GWAS for several flowering-related phenotypes[1], and the geographic distribution for the most differentiated SNP in *SVP* differentiates accessions collected in Fennoscandia, eastern Europe and Russia from accessions collected across the rest of the species distribution. Other loci differentiating accessions collected in Fennoscandia from those in northwest Europe include the flowering-related loci *COP1-interacting protein 4.1* (*CIP4.1*), *FRIGIDA* (*FRI*) and *FLOWERING LOCUS C* (*FLC*), as well as a locus that has a major role in increasing seed dormancy in accessions collected from low latitudes, *DELAY OF GERMINATION 1* (*DOG1*)[36,37].

In previous scans for selection in *A. thaliana*, alleles of the flowering-related locus *FRI*[32] and a region on chromosome 1 (20.34–20.49 Mb)[6,10] were identified as putative targets of selection. We find additional evidence of selection in these regions and identify several new candidates for selection. We found the strongest signal for a partial sweep on chromosome 4 (15.48–15.93 Mb), at a haplotype that occurs throughout the species range. Follow-up studies will be required to localize and confirm the target of selection, but the signal peaked on a SNP at ~15.66 Mb in a gene of unknown function (*AT4G32440*).

Because PHS, CLR and F_{ST} analyses identify loci at different stages in the selection process or loci that are experiencing different modes of adaptation[38], one might not expect the results from these scans to overlap. In fact, they rarely do (**Supplementary Fig. 4**). As an example, **Figure 4a** shows the overlap for the most extreme signals (the top 1% of scores) on chromosome 2. Among the most likely targets of selection is a region identified by both the PHS and F_{ST} analyses (chromosome 2, 13.44–13.86 Mb at a 1% cutoff). This partial sweep is differentiated among populations (**Fig. 4b**) and overlaps with a previously identified genomic duplication[39].

To confirm that our selection scans are identifying candidate regions of interest, we asked whether genes associated with 107 traits[1] grouped into four phenotypic classes related to flowering time, plant defense (for example, recognition of a pathogen's secreted effectors), ionomics (which measures concentrations of trace mineral elements within plant tissue) and development (for example, seed dormancy, leaf morphology or growth rate) overlap with these selection signals. **Supplementary Figure 4** shows the distribution of the top 1% of GWAS, PHS, CLR and F_{ST} signals for each of the five chromosomes. Next, we conducted an enrichment analysis to ask whether the top signals from these GWAS (Online Methods) were enriched in the tails of the scores from the three selection scans. This test was somewhat underpowered: the sample sizes used in these GWAS were small ($n < 200$) and the phenotypes seen were far from exhaustive. Furthermore, the correction for population structure that was applied in the previous study could lead to a high false-negative rate for some of the geographically distributed traits[2]. Nevertheless, we found striking and statistically significant enrichments for phenotypes related to defense (F_{ST}), development (PHS statistic) and ionomic phenotypes (CLR) in the extreme tail (0.1%) of the selection scores (**Supplementary Fig. 5**).

The observation that the F_{ST} analysis is enriched for defense-associated SNPs without any concomitant increase in enrichment in the PHS or CLR analyses is consistent with the emerging view that defense-related traits show little evidence of repetitive selective sweeps, as has been suggested under an arms race model[40,41]. Flowering is correlated with geography, and, as we expected, we saw an enrichment of SNPs associated with flowering-related phenotypes using scans for population differentiation (F_{ST}). There was also a strong enrichment of flowering with the PHS statistic, suggesting that the SNPs responsible for variation in flowering are experiencing ongoing or partial sweeps. An enrichment analysis of the results from GWAS of all 107 individual phenotypes[1] helped to distinguish the underlying modes of adaptation for a wide variety of traits (see URLs and **Fig. 5**).

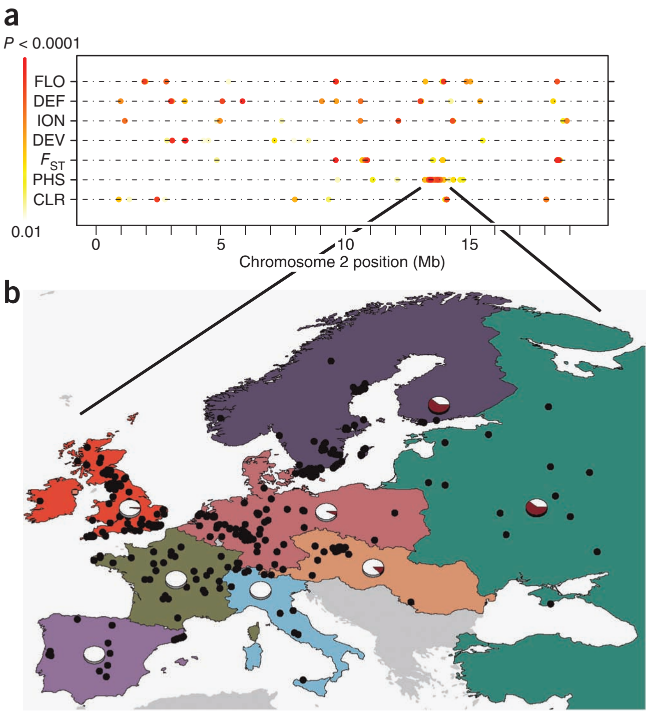

Figure 4 Overlap of selection scans with results from GWAS on chromosome 2. (**a**) The top 1% (genome wide) of scores are shown for three scans of selection (F_{ST}, PHS and CLR); also shown are the top results from GWAS of 107 phenotypes separated into four phenotypic categories: flowering (FLO), defense (DEF), ionomics (ION) and development (DEV). (**b**) The geographic distribution of an unusually long haplotype (with the frequency shown in the pie charts) identified by both the PHS and F_{ST} scans (chr 2: 13.44–13.86 Mb).

© 2012 Nature America, Inc. All rights reserved.

© 2012 Nature America, Inc. All rights reserved.

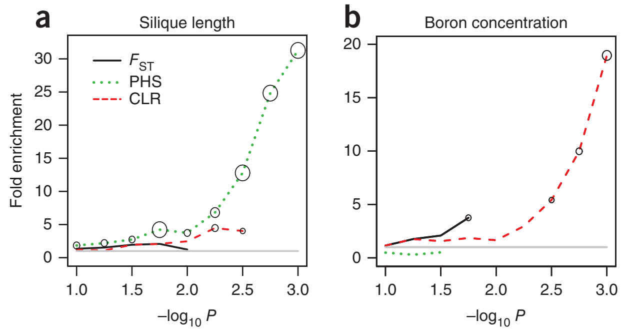

Figure 5 Enrichment of GWAS results with signals of selection. (**a**,**b**) SNPs associated with silique length (**a**) and *in planta* concentration of boron (**b**) are strongly enriched in the extreme tails of scans testing for signatures of selection (for example, a $-\log_{10}$ rank statistic of 1 corresponds to the 10% tail). The sizes of the circles denote significance based on 1,000 permutations (the smallest circle shown corresponds to $P = 0.032$ and the largest circle corresponds to $P = 0.001$). GWAS SNPs were considered if their minor allele frequency was greater than 5% and when $P < 1 \times 10^{-4}$.

This analysis provides insights into the history of recombination and positive selection in a plant species. We provide evidence that our selection scans are enriched for genomic regions that underlie natural variation in ecologically crucial traits and identify several new candidates of selection. Notably, alternative scans of selection have identified disparate traits; together these scans provide a comprehensive picture of how selection has acted on the genome of *A. thaliana*. In addition, we investigated the genetic determinants of recombination at a scale achieved so far only in humans and found that microsatellites have a fundamental role in meiotic recombination in *A. thaliana*.

In *A. thaliana*, recombination and natural selection have largely been studied with population-level genetic data. However, an advantage of using model species is the ability to empirically confirm hotspots and the roles (if any) that candidate loci have in selection or recombination. Our genotyped accessions will be maintained as selfed lineages (see the accession code section), and it will soon be possible to impute the entire genome sequence from data being generated by the *A. thaliana* 1001 Genomes Project[10,42]. Projects are also currently underway using this panel in functional studies and GWAS, both by us and others. Based on previous experience[1–4], strategic selection of members of particular mapping populations will provide an increase in mapping resolution and advance the overlapping aim of both population and ecological genetics: understanding the genetic basis of adaptation in an environmental context.

URLs. The genotype data, the selection results and a browser hosting the selection scores (in a genomic context) are available on the project web site at http://bergelson.uchicago.edu/regmap-data/regmap.html/. The 1001 Genomes Project web site is at http://1001genomes.org/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. The seeds are available through the *Arabidopsis* Biological Resource Center (ABRC) under accession number CS77400.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank M. Hubisz for the code for the CLR statistic and for help in using the software. We also thank E. Leffler for helpful discussions regarding recombination hotspots. M.W.H. was supported by a National Science Foundation Predoctoral Fellowship, a Graduate Assistance in Areas of National Need (GAANN) training grant and an Achievement Rewards for College Scientists (ARCS) Foundation Scholarship. This research was supported by US National Institutes of Health grant GM057994 (J.B.), US National Institutes of Health grant GM083068 (J.B. and M.N.), National Science Foundation 2010 grants (M.N. and J.B.) and a Dropkin Foundation Fellowship (A.M.H.). This is contribution number 11-363-J from the Kansas Agricultural Experiment Station (C.T.).

AUTHOR CONTRIBUTIONS

M.W.H., M.N., J.O.B. and J.B. conceived of and designed the experiments. M.W.H., A.M.H. and C.T. carried out all population genetic analyses. A.A. developed the method used to identify hotspots of recombination. S.A., N.W.M. and A.P. were responsible for the experimental aspects of choosing and genotyping the selected lines. Y.S.H. and B.J.V. analyzed the raw array data. F.G.S. designed the maps shown in the manuscript and on the project website. M.W.H. and J.B. wrote the paper. All other authors commented on the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**, 627–631 (2010).
2. Brachi, B. *et al.* Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature. *PLoS Genet.* **6**, e1000940 (2010).
3. Li, Y., Huang, Y., Bergelson, J., Nordborg, M. & Borevitz, J.O. Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **107**, 21199–21204 (2010).
4. Baxter, I. *et al.* A coastal cline in sodium accumulation in *Arabidopsis thaliana* is driven by natural variation of the sodium transporter AtHKT1;1. *PLoS Genet.* **6**, e1001193 (2010).
5. Kim, S. *et al.* Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.* **39**, 1151–1155 (2007).
6. Clark, R.M. *et al.* Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* **317**, 338–342 (2007).
7. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
8. Nordborg, M. *et al.* The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3**, e196 (2005).
9. Platt, A. *et al.* The scale of population structure in *Arabidopsis thaliana*. *PLoS Genet.* **6**, e1000843 (2010).
10. Cao, J. *et al.* Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* **43**, 956–963 (2011).
11. Lewandowska-Sabat, A.M., Fjellheim, S. & Rognli, O.A. Extremely low genetic variability and highly structured local populations of *Arabidopsis thaliana* at higher latitudes. *Mol. Ecol.* **19**, 4753–4764 (2010).
12. McVean, G.A. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584 (2004).
13. Bergthorsson, U., Andersson, D.I. & Roth, J.R. Ohno's dilemma: evolution of new genes under continuous selection. *Proc. Natl. Acad. Sci. USA* **104**, 17004–17009 (2007).
14. Yang, S. *et al.* Repetitive element–mediated recombination as a mechanism for new gene origination in *Drosophila*. *PLoS Genet.* **4**, e3 (2008).
15. McDowell, J.M. *et al.* Intragenic recombination and diversifying selection contribute to the evolution of downy mildew resistance at the *RPP8* locus of *Arabidopsis*. *Plant Cell* **10**, 1861–1874 (1998).
16. Yandeau-Nelson, M.D. *et al.* MuDR transposase increases the frequency of meiotic crossovers in the vicinity of a Mu insertion in the maize *a1* gene. *Genetics* **169**, 917–929 (2005).
17. Hollister, J.D. & Gaut, B.S. Population and evolutionary dynamics of Helitron transposable elements in *Arabidopsis thaliana*. *Mol. Biol. Evol.* **24**, 2515–2524 (2007).
18. Hu, T.T. *et al.* The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476–481 (2011).
19. Myers, S., Freeman, C., Auton, A., Donnelly, P. & McVean, G. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat. Genet.* **40**, 1124–1129 (2008).
20. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321–324 (2005).
21. Gill, K.S., Gill, B.S., Endo, T.R. & Taylor, T. Identification and high-density mapping of gene-rich regions in chromosome group 1 of wheat. *Genetics* **144**, 1883–1891 (1996).
22. Lichten, M. & Goldman, A.S. Meiotic recombination hotspots. *Annu. Rev. Genet.* **29**, 423–444 (1995).
23. Morgante, M., Hanafey, M. & Powell, W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.* **30**, 194–200 (2002).

24. Blanc, G. & Wolfe, K.H. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**, 1667–1678 (2004).
25. Swigonová, Z. *et al.* Close split of sorghum and maize genome progenitors. *Genome Res.* **14**, 1916–1923 (2004).
26. SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. & Bennetzen, J.L. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**, 43–45 (1998).
27. He, L. & Dooner, H.K. Haplotype structure strongly affects recombination in a maize genetic interval polymorphic for Helitron and retrotransposon insertions. *Proc. Natl. Acad. Sci. USA* **106**, 8410–8416 (2009).
28. Myers, S. *et al.* Drive against hotspot motifs in primates implicates the *PRDM9* gene in meiotic recombination. *Science* **327**, 876–879 (2010).
29. Hancock, A.M. *et al.* Adaptation to climate across the *Arabidopsis thaliana* genome. *Science* **334**, 83–86 (2011).
30. Fournier-Level, A. *et al.* A map of local adaptation in *Arabidopsis thaliana*. *Science* **334**, 86–89 (2011).
31. Smith, J.M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35 (1974).
32. Toomajian, C. *et al.* A nonparametric test reveals selection for rapid flowering in the *Arabidopsis* genome. *PLoS Biol.* **4**, e137 (2006).
33. Nielsen, R. *et al.* Genomic scans for selective sweeps using SNP data. *Genome Res.* **15**, 1566–1575 (2005).
34. Lewontin, R.C. & Krakauer, J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**, 175–195 (1973).
35. Hartmann, U. *et al.* Molecular cloning of SVP: a negative regulator of the floral transition in *Arabidopsis*. *Plant J.* **21**, 351–360 (2000).
36. Alonso-Blanco, C., Bentsink, L., Hanhart, C.J., Blankestijn-de Vries, H. & Koornneef, M. Analysis of natural allelic variation at seed dormancy loci of *Arabidopsis thaliana*. *Genetics* **164**, 711–729 (2003).
37. Chiang, G.C. *et al. DOG1* expression is predicted by the seed-maturation environment and contributes to geographical variation in germination in *Arabidopsis thaliana*. *Mol. Ecol.* **20**, 3336–3349 (2011).
38. Pritchard, J.K., Pickrell, J.K. & Coop, G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.* **20**, R208–R215 (2010).
39. Lin, X. *et al.* Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* **402**, 761–768 (1999).
40. Bakker, E.G., Traw, M.B., Toomajian, C., Kreitman, M. & Bergelson, J. Low levels of polymorphism in genes that control the activation of defense response in *Arabidopsis thaliana*. *Genetics* **178**, 2031–2043 (2008).
41. Bakker, E.G., Toomajian, C., Kreitman, M. & Bergelson, J. A genome-wide survey of R gene polymorphisms in *Arabidopsis*. *Plant Cell* **18**, 1803–1818 (2006).
42. Gan, X. *et al.* Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* **477**, 419–423 (2011).

© 2012 Nature America, Inc. All rights reserved.

npg

ONLINE METHODS

Plant samples. Researchers from a previous study[9] identified 1,799 unique haplogroups in a worldwide collection (>5,700 samples) of *A. thaliana*. We genotyped 837 accessions and combined them with 473 samples genotyped previously[1,3,4]. In total, we genotyped 1,310 accessions using a custom Affymetrix SNP tiling array (AtSNPtile1), which surveys 248,584 SNPs. We followed the same DNA extraction and hybridization protocols as we used previously[1]. After quality control, we identified 214,051 SNPs in each sample. We discarded three accessions (Uod-2, Blh-1 and Santa Clara) that we suspected were contaminants either because their genotypes conflicted with previously collected SNP data[1] or because they differed phenotypically from what has been observed previously. Because our analyses relied on high-quality geographic coordinates, we removed samples whose geographic origins were suspect[43]. In addition, we omitted accessions that would form small sample sizes (including accessions from New Zealand, Cape Verde and Libya); both the full dataset ($n = 1{,}307$) and the geographically referenced dataset ($n = 1{,}193$) are available on the project website (see URLs).

Estimates of linkage disequilibrium. To estimate ρ, we followed a previous approach[20]. Briefly, we used the interval program of LDhat[12], which estimates recombination rates for each regional sample after splitting the data into regions of 2,500 SNPs each (with an overlap of 200 SNPs between regions to allow burn-in). We used a block penalty of 5 and discarded the first one third of 10,000,000 total iterations. Contiguous estimates remained after removing the first (at the 5′ end) and last (at the 3′ end) 100 SNPs from each region.

Identifying hotspots of recombination. To search for recombination hotspots, we used recombination rate estimates obtained from LDhat[12]. To assess the significance of local peaks in recombination, we used a method similar to that used by LDhot[12]. Specifically, for a 2-kb window, a composite likelihood ratio test statistic[12] is calculated for the model in which the recombination rate within the hotspot is equal to the background rate and for a model in which it is allowed to be greater. The significance of the test statistic was assessed using coalescent simulations to reject the null hypothesis of a background recombination rate within the hotspot. We then filtered putative hotspots whose estimated recombination rate across the 2-kb window was ρ (per kb) < 3. The test was repeated for all 2-kb windows across the genome, shifting 1-kb between each window. To combine hotspots across populations, we followed an earlier approach[20].

Searching for the genetic determinants of recombination. To search for sequence features associated with elevated rates of recombination, we matched hotspots with regions for which there was no evidence of a hotspot ($P = 1.0$). These coldspots were matched to hotspots (to within 10%) based on GC content, SNP density and physical length. To assess the enrichment of a particular sequence feature, we followed a previous approach[19]. We counted each simple sequence repeat or transposable element in both hotspots and coldspots and determined its significance using a binomial test; we assessed the significance of motifs (of size 5–9 bp) using a Fisher's exact test to account for the different numbers of hotspots and coldspots in DNA not overlapping transposable elements or pseudogenes. All *P* values were Bonferroni corrected to account for multiple testing.

Identifying signals of selection. The CLR was calculated according to the methods from a previous study[5] with the grid size equal to the number of SNPs on each chromosome. Because of the size of our panel, we took five random samples of 1,025 individuals and analyzed each dataset using the CLR test. We then averaged these replicates to calculate each SNP's CLR score. We were able to determine the ancestral state of 121,624 SNPs (57% of the SNPs typed in *A. thaliana*)[18].

F_{ST} scores were calculated using the methods from a previous study[44]. The PHS statistic is based on the average length of a (pairwise) shared haplotype compared to the genomic average for this pair of individuals[32]. PHS scores were calculated with 1,144 accessions. To take into account genotyping error, haplotype sharing ended when a mismatch occured within 5-kb of another mismatch. Haplotypes shorter than 20 kb were ignored. The haplotype length for any given pair of accessions around a specific SNP is normalized by the distribution of shared haplotypes between that pair of accessions. These values are then averaged over all pairwise comparisons within the allele class and contrasted to the same average of normalized values for all pairwise comparisons of the opposite allele at the same SNP. The difference of values for alleles at the same SNP is then normalized by the distribution of values for all SNPs with the same allele frequency. Because the demographic history of *A. thaliana* is unknown, we considered the genomic pattern of scores from these tests, focusing on those loci that were extreme relative to the rest of the genome.

To generate **Figures 4**, **5** and **Supplementary Figures 4, 5**, we split the genome into 10-kb windows and took the maximum score from the PHS, CLR and F_{ST} scans for each window as the test statistics. To visualize overlap with GWAS, we used results from previous GWAS of 107 phenotypes[1], which were undertaken to account for population structure. We removed SNPs with either low minor-allele frequency (<0.05) or significance ($-\log_{10} P < 4.0$) and then split these results into 10-kb windows. We considered scores per phenotypic class (**Fig. 4** and **Supplementary Figs. 4,5**) or individual phenotype (**Fig. 5** and listed on the project website) as the test statistics.

Enrichment analysis of GWAS SNPs with tests of selection. The enrichment analyses were conducted across 10-kb windows (as described above). For each test of selection, we asked whether there was an enrichment of GWAS-associated windows (after accounting for population structure[1]) in the tail of the distributions of PHS, CLR and F_{ST} scores.

Windows were ranked and for each window, a rank-based score, sometimes referred to as an empirical *P* value, was calculated. Then, for each of the three scans for selection, we asked whether, within the set of windows in the lower tail of the distribution, the proportion of GWAS-associated windows was greater than the proportion of non-GWAS–associated windows. To assess the significance of the observed enrichments of overlap between selection and GWAS signals, we compared our observed results to a null distribution created based on 1,000 permutations. For each permutation, we re-sampled a genome-wide set of windows, preserving the relative positions of the windows, but shifting them by a randomly chosen, uniformly drawn number of windows for each permutation.

43. Anastasio, A.E. *et al.* Source verification of misidentified *Arabidopsis thaliana* accessions. *Plant J.* **67**, 554–566 (2011).
44. Weir, B.S. & Cockerham, C.C. Estimating *F*-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).

© 2012 Nature America, Inc. All rights reserved.

doi:10.1038/ng.1042