

Whole-genome sequencing of multiple *Arabidopsis thaliana* populations

Jun Cao^{1,8}, Korbinian Schneeberger^{1,2,8}, Stephan Ossowski^{1,3,4,8}, Torsten Günther^{5,8}, Sebastian Bender¹, Joffrey Fitz¹, Daniel Koenig¹, Christa Lanz¹, Oliver Stegle⁶, Christoph Lippert⁶, Xi Wang¹, Felix Ott¹, Jonas Müller¹, Carlos Alonso-Blanco⁷, Karsten Borgwardt⁶, Karl J Schmid⁵ & Detlef Weigel¹

The plant *Arabidopsis thaliana* occurs naturally in many different habitats throughout Eurasia. As a foundation for identifying genetic variation contributing to adaptation to diverse environments, a 1001 Genomes Project to sequence geographically diverse *A. thaliana* strains has been initiated. Here we present the first phase of this project, based on population-scale sequencing of 80 strains drawn from eight regions throughout the species' native range. We describe the majority of common small-scale polymorphisms as well as many larger insertions and deletions in the *A. thaliana* pan-genome, their effects on gene function, and the patterns of local and global linkage among these variants. The action of processes other than spontaneous mutation is identified by comparing the spectrum of mutations that have accumulated since *A. thaliana* diverged from its closest relative 10 million years ago with the spectrum observed in the laboratory. Recent species-wide selective sweeps are rare, and potentially deleterious mutations are more common in marginal populations.

Genomes are shaped by the interaction of diverse processes and evolutionary forces: recombination, gene conversion, mutation, selection and demography, as well as recurrent cycles of polyploidization and subsequent diploidization, along with hybridization and the associated processes of admixture and introgression. Disentangling the effects of these processes on sequence variation is essential not only for understanding how genetic diversity is generated and maintained but also for tracking down allelic variants responsible for phenotypic variation. *A. thaliana* and its close relatives have been at the forefront of investigations of these processes in plants^{1,2}. For example, both the local and global population structures of *A. thaliana*, which reflect the species' migration history since the Ice Age as well as the surprisingly frequent outcrossing events between the inbred strains, have been studied in considerable detail^{3,4}. The first genome-wide haplotype map of a plant was produced for this species⁵, and the information from this endeavor has already been successfully used for genome-wide association studies (GWAS)^{6–9}. Despite the rapid progress in linking genotype to phenotype, a major gap remains in the ability to identify alleles that are directly responsible for variation in adaptive traits. As in humans, the complete sequencing of genomes provides an essential stepping stone toward this goal. Moreover, the recent completion of a reference genome sequence for the species' closest relative, *Arabidopsis lyrata*, is informing the interpretation of polymorphism patterns in *A. thaliana*¹⁰.

Exploratory efforts with a small number of strains suggested early on that short-read sequencing is an efficient means of describing whole-genome sequence variation in *A. thaliana*^{11,12}, and on the basis of early successes, a 1001 Genomes Project for the species has been advocated¹³ (see URLs for project website). Here we present results from the first major phase of the 1001 Genomes Project, an analysis of 80 strains that were chosen to represent the genetic diversity present in eight populations across the entire native range of the species. The study design supports systematic investigation of the effects of geography and demography on whole-genome sequence variation.

RESULTS

Sequencing of 80 *A. thaliana* accessions

The native range of *A. thaliana* is in Eurasia, spanning varied climates and elevations, from the high mountains of Central Asia to the European Atlantic Coast, and from North Africa to the Arctic Circle. To enable the discovery of both global and local effects on sequence diversity, we focused on six larger geographic regions: the Iberian Peninsula with North Africa; Southern Italy; Eastern Europe; the Caucasus; Southern Russia; and Central Asia. In addition, we sampled two much smaller regions, Swabia, in the southwest of Germany, and South Tyrol, in the north of Italy (Fig. 1). From each region, we selected 7–14 naturally inbred strains, or accessions, that we had identified as genetically diverse on the basis of limited genome-wide genotyping (Fig. 1a and Supplementary Table 1). From a single individual

¹Max Planck Institute for Developmental Biology, Tübingen, Germany. ²Max Planck Institute of Plant Breeding Research, Cologne, Germany. ³Center for Genomic Regulation, Barcelona, Spain. ⁴Universitat Pompeu Fabra, Barcelona, Spain. ⁵Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart, Germany. ⁶Machine Learning and Computational Biology Research Group, Max Planck Institute for Intelligent Systems and Max Planck Institute for Developmental Biology, Tübingen, Germany. ⁷Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas, Madrid, Spain. ⁸These authors contributed equally to this work. Correspondence should be addressed to D.W. (weigel@weigelworld.org).

Received 8 March; accepted 26 July; published online 28 August 2011; doi:10.1038/ng.911

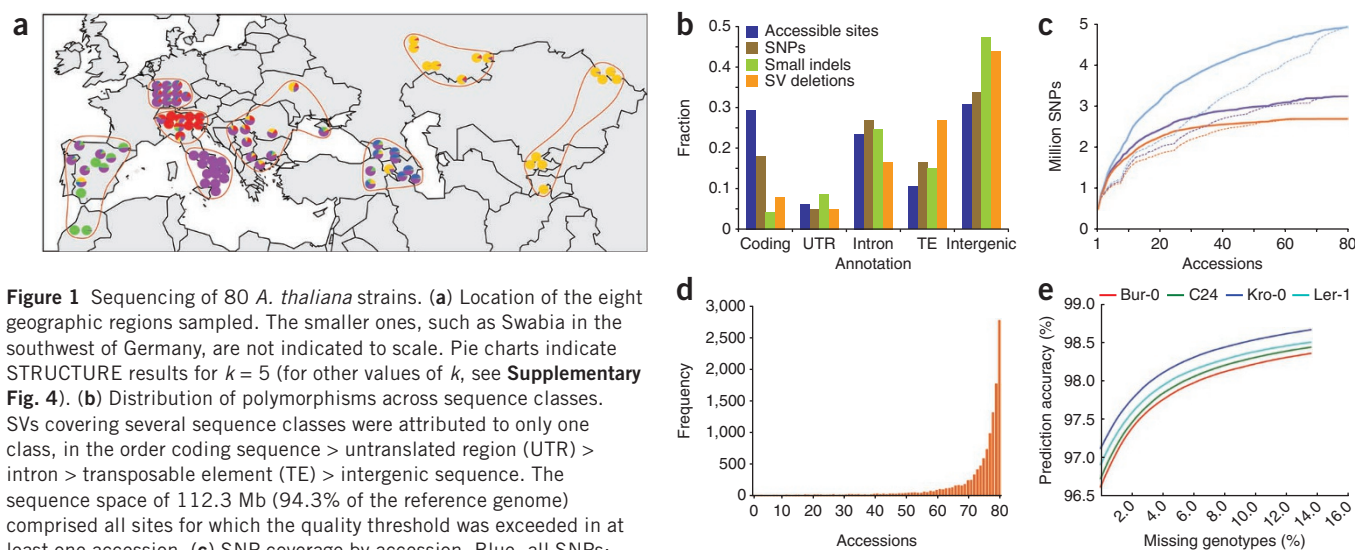


Figure 1 Sequencing of 80 *A. thaliana* strains. (a) Location of the eight geographic regions sampled. The smaller ones, such as Swabia in the southwest of Germany, are not indicated to scale. Pie charts indicate STRUCTURE results for $k = 5$ (for other values of k , see **Supplementary Fig. 4**). (b) Distribution of polymorphisms across sequence classes. SVs covering several sequence classes were attributed to only one class, in the order coding sequence > untranslated region (UTR) > intron > transposable element (TE) > intergenic sequence. The sequence space of 112.3 Mb (94.3% of the reference genome) comprised all sites for which the quality threshold was exceeded in at least one accession. (c) SNP coverage by accession. Blue, all SNPs; purple, private SNPs excluded; orange, private and region-specific SNPs excluded. Solid lines, optimized order of strains; dashed lines, arrangement of strains in reverse order. (d) Allele frequency of transposable elements present in reference but missing in at least one of the 80 strains. (e) Accuracy and percentage of missing genotypes for SNP imputation over a range of imputation thresholds, considering the 80 genomes as a reference panel (3,184 SNPs with <10% missing data) and 216,000 tag SNPs in test accessions.

for each strain, we produced 200-bp insert libraries and generated 10- to 20-fold coverage (average 17) with 42- to 64-bp paired-end reads on the Illumina Genome Analyzer platform (**Supplementary Table 1**). We first predicted single-nucleotide polymorphisms (SNPs) and 1- to 20-bp insertions and deletions (hereafter called small indels) in unique regions on the basis of single-read alignments against the 119.2-Mb reference genome, excluding 6.8 Mb of highly repetitive regions^{11,12,14}. For 67.3% of reference positions, we had information from at least 75 strains, and for 43.6% we had complete information. We randomly chose 273 non-overlapping fragments to validate 325 predicted SNPs and small indels, including 93 low-quality singleton SNPs. We could PCR-amplify 304 sequence variants; dideoxy sequencing revealed that all carried the predicted mutation (**Supplementary Table 2**). We also used discordant mapping of paired ends to reveal structural variants (SVs) of at least 20 bp in length (**Supplementary Note**). Of 40 randomly chosen SV deletions with a predicted size of 22 bp to 14.3 kb, we obtained PCR products for 39, all of which were validated. However, this also revealed a bias in our estimates of SV deletion size, with the true size being on average 5 bp smaller (**Supplementary Table 2**).

Across all 80 strains, we identified 4,902,039 SNPs and 810,467 small indels (**Fig. 1b**), most of which were found in at least two strains. Variants present in at least two of the eight geographic regions included 2,684,863 SNPs and 370,258 small indels, which together covered 5.7 Mb of the reference genome (**Supplementary Fig. 1**). As few as 56 accessions were sufficient to detect 98% of all SNPs shared between regions (**Fig. 1c**), and as few as 67 to identify 98% of all non-private SNPs (**Fig. 1c**). Although a large number of rare SNPs remain to be discovered, we have captured a substantial majority of common SNPs (**Supplementary Fig. 1**). Taking into account the experimentally determined mutation rate¹⁵, the SNP data indicate that the effective population size of *A. thaliana* is between 250,000 and 300,000, which is close to previous estimates^{16,17}.

Because of uncertainty associated with the precise locations of SV ends, the number of unique SV deletions was more difficult to determine, but conservatively was at least 174,789, of which 49% were detected in more than one strain (**Supplementary Fig. 1**). In addition

to SVs in unique regions, we identified variation in transposable elements. Of 31,189 transposable element insertions annotated in the reference genome, 24,917 (80%) showed evidence of being partially or completely absent from the genome of at least one of the 80 sequenced strains (**Fig. 1d**).

To assess sequences that were too divergent for alignment to the reference, we assembled *de novo* unmapped read pairs along with pairs where only one of the two reads had been mapped. We recovered 43,003 contigs of minimum length 200 bp that could be anchored to the reference genome through 50-bp overlap of at least one end, indicating that they indeed came from *A. thaliana* genomic DNA. The largest such fragment was 9,059 bp (**Fig. 2a**). A substantial fraction of fragments could be aligned to the genome of a different *A. thaliana* strain, *Ler-1* (ref. 18), or that of the closest sequenced relative, *A. lyrata*¹⁰, suggesting that many of these sequences reflect the ancestral state of *A. thaliana* (**Supplementary Fig. 2** and **Supplementary Table 1**). We also inferred copy-number variants (CNVs) of minimum length 1,000 bp on the basis of coverage and the presence of pseudo-heterozygous SNPs¹¹. We detected 1,059 CNVs covering 2.2 Mb of the reference genome, of which 393 overlapped with coding sequences. Many were found in more than one accession (**Fig. 2b**), and for more than half, differences in relative coverage suggested that accessions belonged to distinct groups with varying degrees of copy-number variation (**Fig. 2c–e** and **Supplementary Table 3**).

On the basis of the first-generation haplotype map of *A. thaliana*⁵, over 1,000 accessions are being genotyped with a 216,000-SNP array⁶ (the 216k SNP array; see URLs). Although the assayed SNPs are in many cases sufficient to detect significant genome-wide associations with a wide range of phenotypes⁶, they contain only a small subset of functionally relevant polymorphisms. In contrast, with the much richer information from whole-genome sequencing, it should be possible to directly discover many of the variants causal for phenotypic differences. To increase the value of the information from the 80 strains for GWAS, we therefore investigated their utility as a resource for imputing genotypes in accessions analyzed with the 216k SNP array. As a case study, we considered four accessions that were not among the 80 strains sequenced in the current study. On the

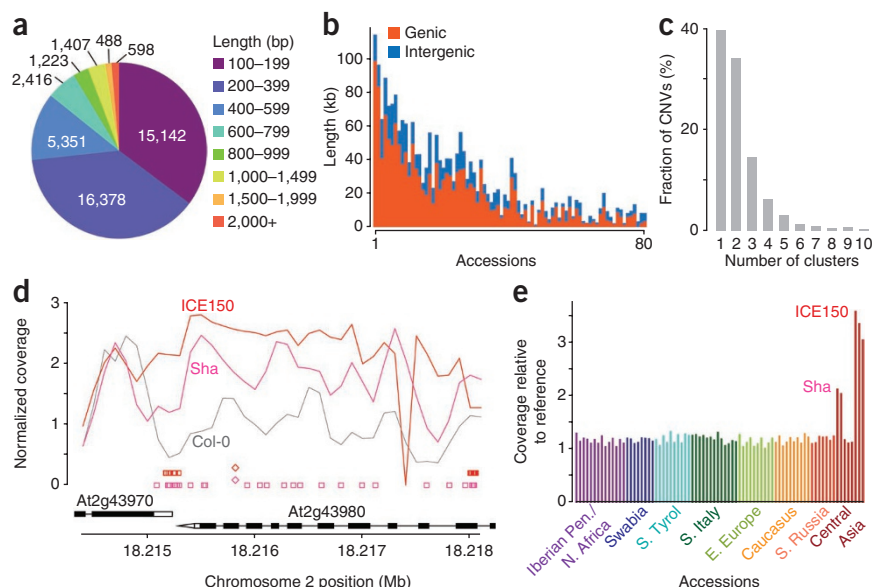


Figure 2 Detection of sequences not present in the reference. **(a)** Length distribution of *de novo*-assembled contigs. Note that this is almost certainly an underestimate, because several accessions yielded few such contigs (**Supplementary Table 1**). **(b)** Allele frequency of CNVs. **(c)** Clustering of accessions for each CNV based on coverage relative to the reference; different clusters had to have at least twofold differences in average coverage. **(d)** Example of a CNV region in two Central Asian accessions. Open diamonds denote pseudo-heterozygous SNPs indicating the presence of different sequences in the inbred genomes; open squares indicate positions of reads for which the corresponding read mate mapped to the reference genome. Although the pattern in ICE150 is consistent with a simple duplication, the situation in Sha appears to be more complex. **(e)** Relative coverage of CNV shown in **d** in 80 accessions, indicating different copy numbers in two groups of Central Asian accessions.

basis of very high sequence coverage with different types of Illumina libraries, the genomes of these strains have been assembled using a reference-guided approach¹⁸. We first called all 216,000 array SNPs in the four accessions and used these data to impute all SNP positions for which we had at least 90% information in the 80 strains. When allowing for no more than 5% missing predictions, accuracy for the remaining sites ranged from 97.8% to 98.2% (**Fig. 1e** and **Supplementary Fig. 3a–d**), which is similar to what has been reported for the human 1000 Genomes Project¹⁹.

Global patterns of polymorphisms

Linkage disequilibrium (LD) in the global population of *A. thaliana* is limited, reflecting relatively frequent outcrossing as well as rapid migration⁵, yet isolation by distance over several scales is easily detected in Eurasia. The species exists in metapopulations, with genetically identical plants in the native range being generally restricted to individual stands^{3,4,20–22}. Within each region, the proportion of SNPs not shared by other regions ranged from 1.7% to 12.7%, and the number of private SNPs per accession was between 322 and 93,199 (**Fig. 3a** and **Supplementary Fig. 1a**). The population from the Iberian Peninsula and North Africa had the highest fraction of region- and accession-specific SNPs, whereas Central Asia had the lowest. F_{ST} , an indicator of allelic differentiation, ranged from 0.06 to 0.21 when the eight regions were compared. Caucasus and Eastern European accessions were the least distinct, with a pattern of admixture between them revealed by STRUCTURE (**Fig. 1a** and **Supplementary Fig. 4**). The Central Asian and Southern Russian accessions were much more highly differentiated from those in the accessions of other regions than from each other, consistent with a principal-component analysis and STRUCTURE results (**Fig. 1a** and **Supplementary Figs. 4** and **5**).

The differences in regional diversity were also seen in the LD patterns. The background level of LD is an indication of relatedness throughout the genome: the more related two strains are, the more often they have the same

alleles at physically unlinked loci^{23,24}. Accordingly, we observed large regional differences in LD. For example, the Swabian sample, from a small area of only about 1,000 km², had a background level of LD that was much lower than that in Central Asia, an area that is several 100 times larger (**Fig. 3b**).

The observations on regional patterns of diversity and LD are consistent with previous proposals for the demographic history of *A. thaliana*, which postulated the existence of refugia around the Mediterranean and the Black Sea during the last glaciation. Apart from South Tyrol on the southern face of the Alps, Southern Russia and Central Asia were the most uniform regions, indicating that they were colonized relatively recently^{21,22,25–27}.

In a genome-wide analysis of 20 genetically diverse accessions²⁸, we had previously found that candidates for strong selective sweeps were rare, with only one very obvious region of about 0.5 Mb on chromosome 1. In our collection of 80 strains, this was again the only region that stood out in a genome-wide analysis of variation in haplotype diversity²⁹ and LD (**Fig. 4a** and **Supplementary Fig. 6**). One other notable region was on chromosome 2, where we identified a block of elevated LD in the Caucasus (**Fig. 4a**). If the chromosome 1 region corresponds to a relatively recent species-wide sweep, there should be an excess of low-frequency alleles; this is indeed what we found. Although the frequency of derived alleles (as deduced from a comparison with

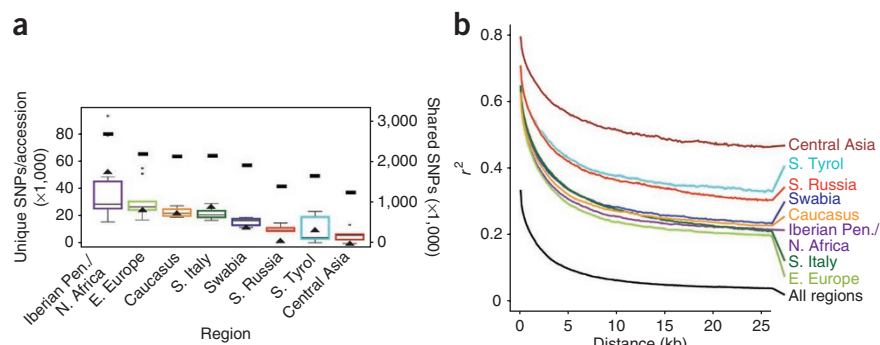


Figure 3 Patterns of allele sharing among populations. **(a)** SNPs that are region specific (triangles) or shared between regions (horizontal bars). Box plots indicate distribution of accession-specific SNPs. **(b)** Decay of LD, expressed as r^2 , in seven individuals from each region or from all individuals.

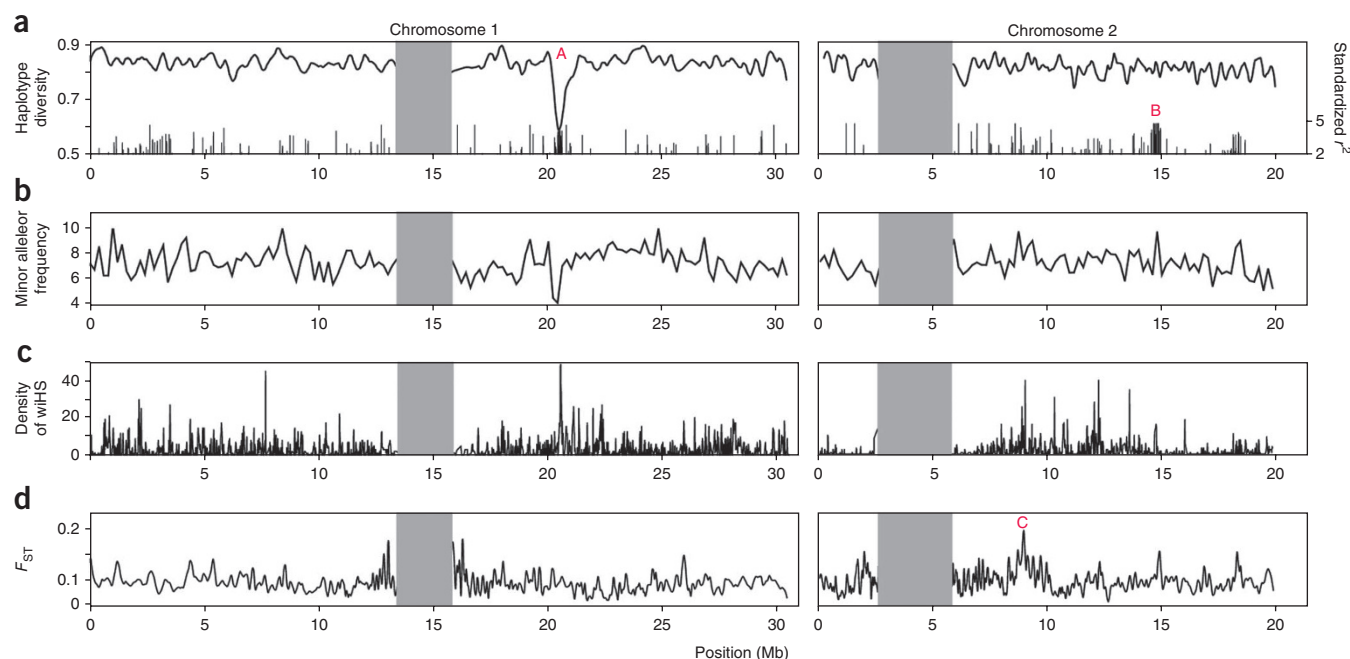


Figure 4 Footprints of selection. (a) Haplotype diversity (continuous lines; window size 10 kb, scale on the left) and LD in Caucasus accessions (vertical bars; window size 5 kb, scale on the right) on chromosomes 1 and 2. 'A' marks the sweep candidate on chromosome 1, and 'B' marks the high-LD block on chromosome 2. (b) Minor-allele frequency, window size 200 kb. (c) Genome-wide wiHS statistics³⁰, window size 10 kb. (d) F_{ST} , calculated for all SNPs, with window size of 50 SNPs and a local polynomial regression performed. 'C' marks the highest non-centromere peak. See **Supplementary Figure 6** for other chromosomes and regions.

*A. lyrata*¹⁰) was not elevated, this sweep candidate was supported by the weighted integrated haplotype score (wiHS), a measure that we developed (see Online Methods)³⁰ (Fig. 4b,c and **Supplementary Fig. 6**). Additionally, wiHS was high near *FRIGIDA* (*FRI*), a locus previously demonstrated to evolve under positive selection^{31,32}.

Differentiation between populations was not uniform across the genome. Both F_{ST} and fluctuations of F_{ST} were elevated closer to the centromeres (**Supplementary Fig. 6**). The highest peak distant from any centromere was found on chromosome 2 (Fig. 4d). It spanned over 600 kb and differentiated the European populations from the Central Asian and Southern Russian populations. Two-thirds of all nonsynonymous SNPs that were not shared between the two geographic regions ($F_{ST} = 1$) were located under this peak, with the extreme value suggesting that this part of the genome includes causal factors for population differentiation. This region also contained clusters of SNPs with high wiHS values (Fig. 4c).

Comparing the mutational and polymorphism spectra

The spectrum of polymorphisms in a wild individual results from mutation and selection as well as from other processes such as biased gene conversion, which favors the transmission of G/C over A/T alleles^{33,34}. We exploited the reference genome sequence of the *A. thaliana* relative *A. lyrata*¹⁰ to determine which alleles have arisen since the two species diverged about 10⁷ years ago^{35,36}. Given a generation time of about 1 year and an experimentally determined mutation rate of less than 10⁻⁸ per generation per site (ref. 15), only a small minority of sites will have experienced multiple mutations, a fact that in turn simplifies the interpretation of the observed patterns.

The mutation spectrum in *A. thaliana* plants grown over a few years in the greenhouse is strongly biased. Six possible types of single-nucleotide changes can be distinguished, and G:C→A:T transitions account for about half of the observed mutations¹⁵. In the 80 accessions,

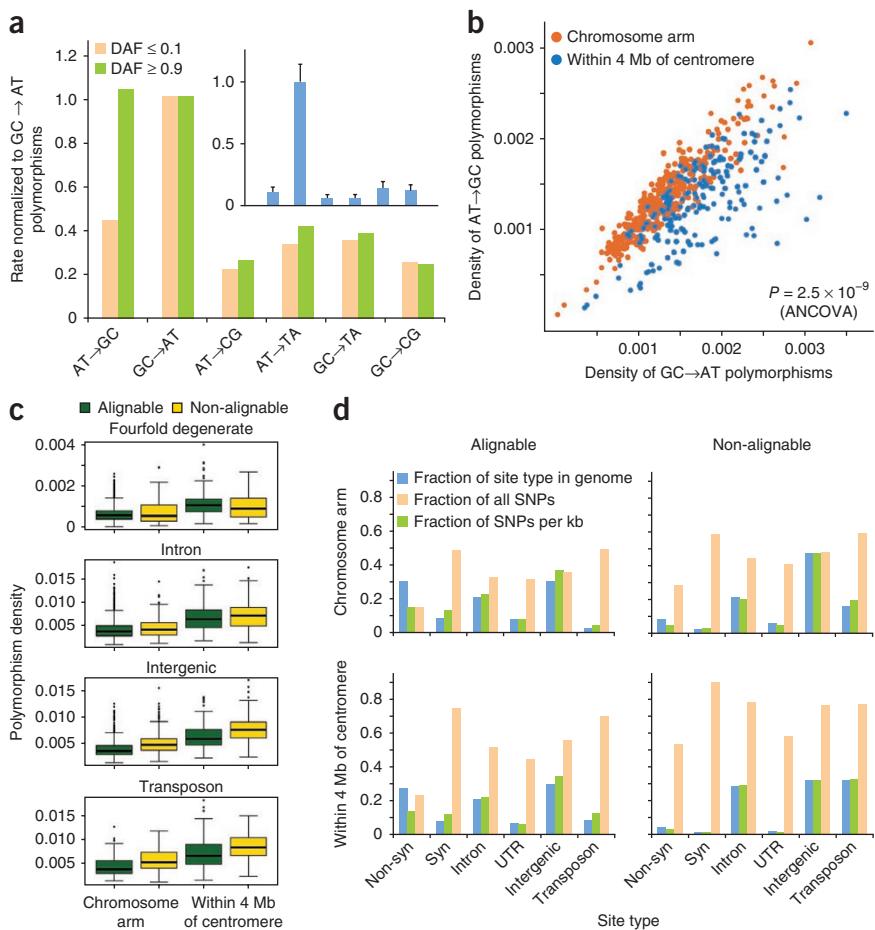
the distribution of the different classes was much less skewed, and an excess of G:C→A:T polymorphisms was seen only in derived alleles of low frequency, most of which will be the result of relatively recent mutations. In contrast, among high-frequency derived alleles, A:T→G:C and G:C→A:T changes were almost in balance (Fig. 5a). The allele frequency-dependent difference in the ratio of these two polymorphism types is in principle consistent with the effect of biased gene conversion, which can masquerade as selection for G/C over A/T alleles³⁴. Such a phenomenon has previously been observed in *A. thaliana*, and biased gene conversion has been suggested as a possible cause (P. Calabrese, T. Hu and M. Nordborg, personal communication). However, biased gene conversion should affect G:C→A:T and C:G→A:T substitutions equally, which we did not find to be the case.

Deamination of methylated cytosines may account at least in part for the excess of G:C→A:T substitutions observed in the greenhouse¹⁵. In agreement with this possibility, we find that ancestral cytosines reported to be methylated in the *A. thaliana* reference strain by either of two groups^{37,38} had a more than twofold higher polymorphism rate than unmethylated cytosines, with the highest rate found in CHG sites (where H is A, C or T) (Table 1). We furthermore noticed that G:C→A:T transitions were not equally distributed throughout the genome; rather, the genome-wide excess over the opposite class is almost exclusively due to low-frequency SNPs in non-protein coding sequences around the centromeres, extending about 4 Mb into each arm (Fig. 5b and **Supplementary Fig. 7**). This is consistent with a slight over-representation of G:C→A:T polymorphisms in these regions among spontaneous mutations in the greenhouse¹⁵. This effect, in turn, probably contributes to the fact that these regions are more polymorphic than the rest of the genome, as has been noted before in *A. thaliana*^{21,22,28} and rice³⁹.

Polymorphism levels as measured by pairwise sequence differences are positively correlated with population recombination rate ρ



Figure 5 Mutational spectrum. (a) Rates of the six different types of polymorphisms, polarized against *A. lyrata*. Rates of G:C→A:T type polymorphisms were set to a reference level of 1 such that for DAF ≤ 0.1, 1 equals 0.016 per site, and for DAF ≥ 0.9, 1 equals 0.002 per site. For comparison, inset shows spontaneous mutation spectrum in *A. thaliana*¹⁵, where 1 equals 1 × 10⁻⁸ per site per generation. (b) Distribution of intergenic transitions in 200-kb windows along chromosomes. See **Supplementary Figure 7** for other site types. (c,d) Polymorphism density as a function of position on chromosome and alignability to *A. lyrata*.



(ref. 5). This relationship was supported by our data (**Supplementary Fig. 8**), but we found that the relationship was affected by chromosome location and polymorphism type; the proportion of explained variance, r^2 , could be as high as 0.49, for intergenic sites on the chromosome arms, and as low as 0.22, for synonymous sites in centromere-adjacent regions. As reported before⁵, genome-wide ρ was only weakly correlated with recombination rates directly estimated from F_2 crosses (**Supplementary Fig. 8**). Perhaps most interesting is the finding of increased polymorphism rates in regions that cannot be aligned against the *A. lyrata* genome, with a compound effect of distance to centromeres. The fraction of nonsynonymous substitutions was also disproportionately elevated in these regions, which might result from a combination of increased mutation rate and less efficient selection (**Fig. 5c,d**).

An atlas of high-confidence mutations

A major motivation for complete sequencing of genomes is to identify variants that are likely to have functional consequences based on their annotation. Across the 80 accessions analyzed here, we discovered in 6,197 genes 12,468 SNPs that altered start codons, introduced premature stop codons, extended the open reading frame of the reference sequence, or affected splice donor or acceptor sites (**Supplementary Table 3**). Furthermore, 27,167 small indels have the potential to cause frameshifts in open reading frames. However, we have recently found that multiple indels in *A. thaliana* often combine to restore the correct frame¹⁸, and thus the number of frameshifts that drastically change coding sequences is likely to be much smaller. Finally, 4,525 larger SV deletions overlapped coding sequences of 2,247 genes by at least 50 bp (**Supplementary Fig. 9** and **Supplementary Table 3**).

Table 1 SNPs and cytosine methylation at ancestral sites

Context	Reference all		Reference methylated ^a	
	Number	% SNP	Number	% SNP
CpG	3,375,883	5.1	655,645	8.8
CpHpG	4,013,740	3.4	181,936	12.6
CpHpH	19,356,698	3.6	584,612	7.6
All Cs	26,746,321	3.7	1,422,193	8.7
All bases	75,038,357	2.8	Not applicable	Not applicable

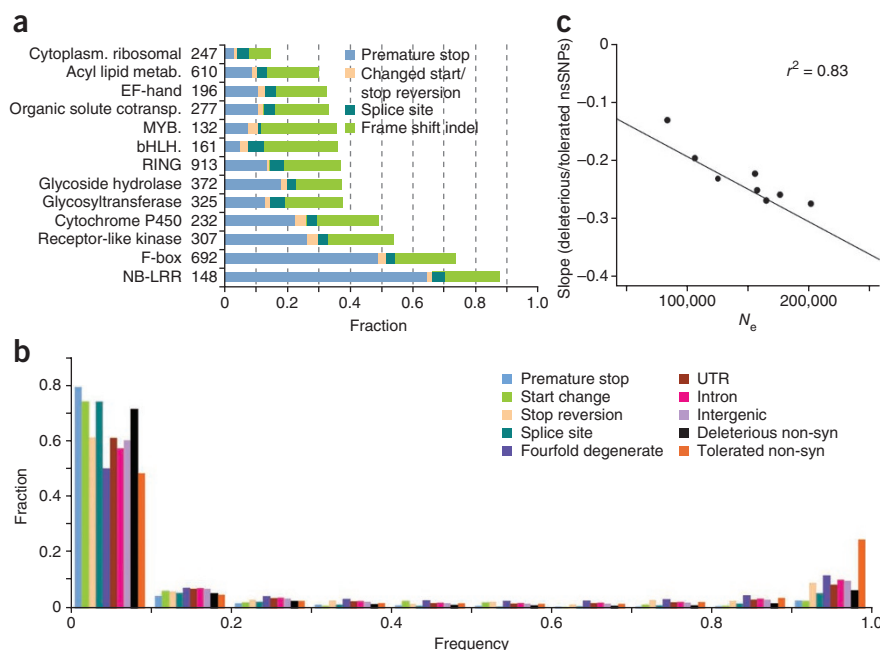
^aMethylated on top or bottom strand according to either reference 36 or reference 37.

Because multiple out-of-phase indels can compensate for each other, because alternative splice sites might be used and because the boundaries of SVs are uncertain, we considered for further analyses only the most dramatic mutations, those that result in premature stop codons. As compared to those in previous studies^{11,28}, the much larger set of accessions analyzed here greatly increased the number of affected genes: 4,263 genes had a premature stop in at least one accession, and 2,793 in two or more accessions. Mutations were nonrandomly distributed: genes that were affected in more than one accession often had sustained independent drastic mutations (**Fig. 6a**). The distribution across gene families was similar to what has been reported before, with the NB-LRR, F-box, RLK and RING families being the most commonly affected ones^{11,28}. Notably, these four families were also the most common ones represented in newly assembled fragments that did not align completely against the reference genome (**Supplementary Fig. 2**).

Leaving the NB-LRR genes aside, it was surprising that for over 100 of the genes with premature stop codons, single or multiple mutant phenotypes have been reported in one of the commonly used laboratory strains (see URLs). We therefore suspected that many, if not most, of these were false positives. Dideoxy sequencing of PCR products, however, confirmed 29 out of 36 such SNPs genes with known single mutant phenotypes (**Supplementary Table 2**). Of the remaining seven, two had the predicted SNP but also additional complex changes, and one had, in addition to the predicted SNP, an adjacent SNP that restored the open reading frame. Another one appeared to be heterozygous, likely because of a recent duplication, consistent with recent findings that different *A. thaliana* strains

Figure 6 Deleterious polymorphisms.

(a) Incidence of drastic mutations in major gene families. (b) Minor- and derived-allele frequencies of polymorphisms with different functional annotation. (c) Relationship between frequency spectrum of apparent deleterious mutations and population size. Shown is the correlation between estimated effective population size and the slope of the regression lines of the ratio between deleterious and tolerated mutations versus their derived allele frequencies for each geographic region. More negative values indicate stronger selection against inferred deleterious mutations⁴¹ (Supplementary Fig. 10).



can carry reciprocal mutations in members of paralogous genes with redundant function^{40,41}. We also analyzed 25 deletions with a predicted size of 20–354 bp that overlapped the open reading frame of genes with known mutant phenotypes by at least 1 bp; a deletion could be confirmed in 19, and an overlap with the coding sequence in 15, of these (Supplementary Table 3). However, as for the SNPs, we cannot exclude the possibility that such deletions are compensated by nearby complex changes.

Although many of the drastic mutations are rare and are thus unlikely to be adaptive (Fig. 6b), there were notable exceptions. These mutations include those in *PHYTOCHROME D* (*PHYD*), for which a small deletion was previously reported to alter the light sensitivity of a single accession⁴² and for which we identified multiple loss-of-function alleles. We also identified mutations that have occurred since *A. thaliana* and *A. lyrata* diverged, assuming that sequences identical in both species normally constitute the ancestral allele. We found 5,282 premature stop codons that had apparently arisen on the *A. thaliana* lineage; of these, 168 segregated at a frequency of 0.9 or higher.

Turning to more subtle mutations, we used MAPP to calculate the fraction of nonsynonymous mutations likely to be deleterious^{43,44}. As expected, deleterious mutations were rarer among high- compared to low-frequency derived alleles. There were, however, important differences between populations: whereas singletons were about four times as common as fixed alleles in most geographic areas, this difference was only about twofold in the Southern Russian and Central Asian groups. We found that the ratio of deleterious to tolerated mutations was highly negatively correlated with the effective population size (Fig. 6c and Supplementary Fig. 10). This can be explained by deleterious mutations accumulating at intermediate frequencies in small populations as a result of less effective purifying selection⁴⁴. These observations are consistent with a recent bottleneck and a reduced effective population size of these marginal populations.

Finally, of 173 *MIRNA* loci, 17 completely lacked sequence coverage in at least one strain. These included, however, only two of the deeply conserved miRNAs, and both of these were in families with multiple copies. Of the remaining 15, 8 are not found in *A. lyrata* and 1 produces a different miRNA in *A. lyrata*^{45,46}. Similarly conspicuous was that the nonconserved loci were missing in as many as 33 strains, providing further evidence that these miRNAs are not essential (Supplementary Table 4).

DISCUSSION

Our analysis of 80 genomes provides a survey of the majority of common SNPs, along with many small and large deletions as well

as duplicated regions and nonreference sequences segregating in *A. thaliana*. These will include many variants that support adaptation to the wide range of environments encountered by the species across its native range. GWAS can be a powerful tool for identifying such alleles, as long as they are sufficiently frequent in the population, and several recent reports have demonstrated the potential of this approach for *A. thaliana*^{6–9,47}. Much as in human studies, a first step in bringing GWAS to *A. thaliana* has been to genotype over 1,000 strains with about 216,000 tag SNPs that should predict over 90% of all common variants (see URLs). A pilot experiment has shown that as few as 96 lines can be sufficient for high-precision GWAS mapping of alleles underlying genetically simple traits⁶. The SNPs interrogated on the 216k SNP array have all been called in the 80 strains we have sequenced; thus, the collection of 80 lines, which is available from the *Arabidopsis* Biological Resources Center (CS76427), can be seamlessly integrated into any GWAS effort that makes use of the array data. We have shown that the polymorphisms identified in the 80 accessions can be used for highly accurate imputation of polymorphisms in strains that have only been analyzed with the 216k SNP array. GWAS scans with the imputed SNPs perform as well as with the directly observed SNPs, confirming the quality of the imputation approach (Supplementary Fig. 2e,f; the fact that, unlike in humans, imputed SNPs in *A. thaliana* do not outperform the observed SNPs is likely due to the much higher density of observed SNPs, around 2 per kb). Furthermore, the near-complete sequence information in our 80 strains will often enable the direct identification of candidates for the causal variant underlying an adaptive trait. To increase the utility of our dataset for the community, we developed the web application POLYMORPH (see URLs) for query, analysis and visualization of polymorphism data. Arbitrary subsets of the 80 strains can be defined, and sequence data for selected genomic regions can be easily extracted for more detailed analyses.

Apart from population structure, a major challenge for GWAS is genetic heterogeneity, whereby different alleles at the same locus may have similar phenotypic effects^{48,49}. The power of combining functional information across alleles of different origin has been well known since the discovery of recurrent mutations in genes for Mendelian disorders in humans and in oncogenes. We have found that the pattern of

mutations that are likely to severely impair gene function because they truncate open reading frames or delete portions of the coding region is not random, suggesting that genetic heterogeneity will be relevant in *A. thaliana* as it is in humans. Current GWAS methods, however, generally do not consider the nature of a variant allele because polymorphism information is incomplete and assayed variants are used only as proxies for all linked polymorphisms. Indeed, variation in one of the best-studied traits in *A. thaliana*, the requirement of extended cold for flowering, is to a large extent controlled by two genes, *FRI* and *FLOWERING LOCUS C (FLC)*, with numerous alleles in the global *A. thaliana* population. As expected, allelic heterogeneity greatly complicates the detection of these loci in GWAS analyses⁶. With improved sequencing methods, the next phase of the *A. thaliana* 1001 Genomes Project should lead to the complete reconstruction of alleles, starting with reference guided assemblies^{18,50}. Use of the near-complete information afforded by studies such as the 1001 Genomes Project will in future support the joint consideration of all alleles predicted to decrease gene activity.

URLs. 1001 Genomes Project web site, <http://1001genomes.org/>; download site for data from this paper, <http://1001genomes.org/data/MPI/MPICao2010/>; 216k SNP array and GWAS data, <http://regmap.uchicago.edu> and <http://arabidopsis.usc.edu>; *A. thaliana* mutant phenotypes, <http://arabidopsis.org/>; POLYMORPH, <http://polymorph.weigelworld.org>.

METHODS

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

Data availability. Flat files of polymorphisms as well as a complete genome matrix based on the reference genome for all 80 strains can be downloaded from the project website (see URLs). A matrix of calls for 216,000 SNPs representing the tag SNPs used for GWAS⁶ is available at the above website and is included in POLYMORPH (see URLs).

Accession codes. Seeds of all 80 strains are available from the *Arabidopsis* Biological Resources Center stock center under accession number CS76427. The sequencing reads have been deposited in the NCBI Short Read Archive under accession number SRA029270.

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

ACKNOWLEDGMENTS

We thank M. Koch for discussion of *Arabidopsis* phylogeography; P. Salomé for *F₂* recombination data; E. Chae, T. Hu, M. Nordborg and P. Salomé for improving the manuscript with their comments; and P. Calabrese, T. Hu and M. Nordborg for communicating their observations regarding the potential effects of biased gene conversion. This work has been supported by grants from the German Federal Ministry of Education and Research (BMBF)-Bioenergy 2021 Initiative (BioÖl, 0315429C; K.J.S.), the Volkswagen Foundation (I/84 225; T.G.), ERA-PG grant ARABRAS and GABI-FUTURE grant GABI-GNADE from the BMBF (D.W.), a Gottfried Wilhelm Leibniz Award of the German Research Foundation (DFG) (D.W.) and the Max Planck Society (K.B., D.W.).

AUTHOR CONTRIBUTIONS

J.C., K.S., S.O., K.J.S. and D.W. designed the study. C.A.-B. and K.J.S. provided biological material. J.C., D.K. and C. Lanz generated the sequencing data. J.C., K.S., S.O., X.W. and F.O. performed primary analysis of the sequencing data. J.C., K.S., S.O. and S.B. identified and annotated polymorphisms. J.C., J.M. and T.G. performed population genomic analyses. O.S., C. Lippert and K.B. performed imputations. J.F. developed web tools. J.C., K.J.S. and D.W. wrote the manuscript with contributions from all authors.

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>

- Clauss, M.J. & Koch, M.A. Poorly known relatives of *Arabidopsis thaliana*. *Trends Plant Sci.* **11**, 449–459 (2006).
- Mitchell-Olds, T. & Schmitt, J. Genetic mechanisms and evolutionary significance of natural variation in *Arabidopsis*. *Nature* **441**, 947–952 (2006).
- Platt, A. *et al.* The scale of population structure in *Arabidopsis thaliana*. *PLoS Genet.* **6**, e1000843 (2010).
- Bomblies, K. *et al.* Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of *Arabidopsis thaliana*. *PLoS Genet.* **6**, e1000890 (2010).
- Kim, S. *et al.* Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.* **39**, 1151–1155 (2007).
- Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**, 627–631 (2010).
- Brachi, B. *et al.* Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature. *PLoS Genet.* **6**, e1000940 (2010).
- 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).
- 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).
- Hu, T.T. *et al.* The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476–481 (2011).
- Ossowski, S. *et al.* Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* **18**, 2024–2033 (2008).
- Schneeberger, K. *et al.* Simultaneous alignment of short reads against multiple genomes. *Genome Biol.* **10**, R98 (2009).
- Weigel, D. & Mott, R. The 1001 Genomes Project for *Arabidopsis thaliana*. *Genome Biol.* **10**, 107 (2009).
- The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
- Ossowski, S. *et al.* The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**, 92–94 (2010).
- Moore, R.C. & Purugganan, M.D. The early stages of duplicate gene evolution. *Proc. Natl. Acad. Sci. USA* **100**, 15682–15687 (2003).
- Gossmann, T.I. *et al.* Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol. Biol. Evol.* **27**, 1822–1832 (2010).
- Schneeberger, K. *et al.* Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. *Proc. Natl. Acad. Sci. USA* **108**, 10249–10254 (2011).
- The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
- Sharbel, T.F., Haubold, B. & Mitchell-Olds, T. Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. *Mol. Ecol.* **9**, 2109–2118 (2000).
- Schmid, K.J., Ramos-Onsins, S., Ringvs-Beckstein, H., Weisshaar, B. & Mitchell-Olds, T. A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics* **169**, 1601–1615 (2005).
- Nordborg, M. *et al.* The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3**, e196 (2005).
- Nordborg, M. & Tavaré, S. Linkage disequilibrium: what history has to tell us. *Trends Genet.* **18**, 83–90 (2002).
- Flint-Garcia, S.A., Thornsberry, J.M. & Buckler, E.S. Structure of linkage disequilibrium in plants. *Annu. Rev. Plant Biol.* **54**, 357–374 (2003).
- Schmid, K.J. *et al.* Evidence for a large-scale population structure of *Arabidopsis thaliana* from genome-wide single nucleotide polymorphism markers. *Theor. Appl. Genet.* **112**, 1104–1114 (2006).
- Beck, J.B., Schmuths, H. & Schaal, B.A. Native range genetic variation in *Arabidopsis thaliana* is strongly geographically structured and reflects Pleistocene glacial dynamics. *Mol. Ecol.* **17**, 902–915 (2008).
- Picó, F.X., Mendez-Vigo, B., Martínez-Zapater, J.M. & Alonso-Blanco, C. Natural genetic variation of *Arabidopsis thaliana* is geographically structured in the Iberian peninsula. *Genetics* **180**, 1009–1021 (2008).
- Clark, R.M. *et al.* Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* **317**, 338–342 (2007).
- Nei, M. *Molecular Evolutionary Genetics* (Columbia University Press, New York, 1987).
- Günther, T. & Schmid, K.J. Improved haplotype-based detection of ongoing selective sweeps towards an application in *Arabidopsis thaliana*. *BMC Res. Notes* **4**, 232 (2011).
- Le Corre, V., Roux, F. & Reboud, X. DNA polymorphism at the *FRIGIDA* gene in *Arabidopsis thaliana*: extensive nonsynonymous variation is consistent with local selection for flowering time. *Mol. Biol. Evol.* **19**, 1261–1271 (2002).
- Toomajian, C. *et al.* A nonparametric test reveals selection for rapid flowering in the *Arabidopsis* genome. *PLoS Biol.* **4**, e137 (2006).
- Lynch, M. *The Origins of Genome Architecture* (Sinauer Associates, Sunderland, Massachusetts, USA, 2007).
- Galtier, N. & Duret, L. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet.* **23**, 273–277 (2007).

35. Wright, S.I., Lauga, B. & Charlesworth, D. Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. *Mol. Biol. Evol.* **19**, 1407–1420 (2002).
36. Beilstein, M.A., Nagalingum, N.S., Clements, M.D., Manchester, S.R. & Mathews, S. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **107**, 18724–18728 (2010).
37. Cokus, S.J. *et al.* Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**, 215–219 (2008).
38. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536 (2008).
39. Huang, X. *et al.* Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42**, 961–967 (2010).
40. Bikard, D. *et al.* Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. *Science* **323**, 623–626 (2009).
41. Vlad, D., Rappaport, F., Simon, M. & Loudet, O. Gene transposition causing natural variation for growth in *Arabidopsis thaliana*. *PLoS Genet.* **6**, e1000945 (2010).
42. Aukerman, M.J. *et al.* A deletion in the *PHYD* gene of the *Arabidopsis* Wassilewskija ecotype defines a role for phytochrome D in red/far-red light sensing. *Plant Cell* **9**, 1317–1326 (1997).
43. Stone, E.A. & Sidow, A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.* **15**, 978–986 (2005).
44. Günther, T. & Schmid, K.J. Deleterious amino acid polymorphisms in *Arabidopsis thaliana* and rice. *Theor. Appl. Genet.* **121**, 157–168 (2010).
45. Fahlgren, N. *et al.* MicroRNA gene evolution in *Arabidopsis lyrata* and *Arabidopsis thaliana*. *Plant Cell* **22**, 1074–1089 (2010).
46. Ma, Z., Coruh, C. & Axtell, M.J. *Arabidopsis lyrata* small RNAs: transient MIRNA and small interfering RNA loci within the *Arabidopsis* genus. *Plant Cell* **22**, 1090–1103 (2010).
47. Todesco, M. *et al.* Natural allelic variation underlying a major fitness trade-off in *Arabidopsis thaliana*. *Nature* **465**, 632–636 (2010).
48. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
49. Dermitzakis, E.T. & Clark, A.G. Genetics. Life after GWA studies. *Science* **326**, 239–240 (2009).
50. Gan, X. *et al.* Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature*, in the press (2011).

ONLINE METHODS

Plant material. Strains were selected from collections representing Central Asia and Southern Russia²⁵, the Caucasus²⁶, the Balkans²⁵, Swabia in southwestern Germany⁴, Southern Italy, the Iberian Peninsula with North Africa, the Italian Alps and South Tyrol (this work). Within each region, we maximized genetic and geographic diversity on the basis of at least 100 genome-wide SNP markers^{3,4,25}. Usually a single accession per local site was chosen, but if the number of sampling localities in a geographic region was limited, two or three accessions per local site were selected.

Library preparation and sequencing. Genomic DNA from a single individual representing each strain was used for library construction and sequencing⁵¹; progeny of the sequenced individuals were submitted to the stock center.

Prediction of SNPs and 1- to 20-bp indels. SNPs and indels up to 3 bp were called with SHORE¹¹. From a matrix with empirically determined penalties for various alignment features, we calculated a quality score for consensus calls at each reference position. Features included alignment repetitiveness, absolute and observed-to-expected coverage, allele frequency, base qualities, sequence complexity, GC content, probability of misaligned indels and local coverage uniformity upstream (**Supplementary Table 5**). The 'core' of an alignment from positions 5 to $n - 5$ was weighted more highly¹¹. Expected coverage was computed from genome-wide correlation of local coverage and GC content. To further flag difficult alignments in homo- and dipolymer runs as well as short simple sequence repeats, sequence complexity was defined by different triplets around the focal position. Individual variants were accepted with quality above Q24, or above Q14 for additional alleles at the same position.

To identify indels up to 20 bp, we combined BWA⁵² (allowing gaps up to 20 bp) and BLAT⁵³ with a custom split-read indel prediction algorithm. All reads unmapped with BWA were aligned to the genome using BLAT. Non-repetitive BLAT alignments with one gap in either query or target were merged with BWA alignments. A simple realignment algorithm was used to adjust gaps in homopolymers and simple repeats such that the last instance of the repeat was annotated as deleted or inserted. In a process similar to the consensus-calling algorithm of SHORE¹¹, we implemented an indel prediction method based on a position-wise majority vote that takes into account multiple sequence- and alignment-based features as well as quality features describing the mapping quality upstream and downstream of the gap. To measure the likelihood of a misalignment, we computed the probability P that any heptamer of an anchor occurred by chance within the length of the indel. We further measured the sequence complexity of an anchor as the number of different triplets it contained. Gapped or split-read alignments with a length of greater than 6, a P value of < 0.005 and sequence complexity of at least 5 were accepted as high quality. Homozygous indels were accepted if they were supported by at least two high-quality alignments out of three alignments, had an allele frequency of greater than 0.51 and did not show coverage biases or repetitiveness.

SV deletion predictions. Distance and orientation of read mates were obtained with SHORE. Read mates were hierarchically clustered by insert type, orientation and size variation. Clusters were assigned to the different SV types. A statistical framework⁵⁴ was used to assign probability values (**Supplementary Note**), and read coverage was used to flag false positives. Complex rearrangements were excluded from further analysis. Transposable element deletions had to overlap reference transposable elements by at least 50 bp.

Assembly of sequences absent from the reference and CNV analysis. Leftover reads and their mapped mates, if any, were assembled *de novo* using Velvet 1.1 (ref. 55). Contigs longer than 199 bp were aligned against the reference genome using MuMmer 3.2 (ref. 56).

Copy-number variants were detected with three steps, by comparison with actual coverage of the resequenced *A. thaliana* reference strain Col-0. First, seed regions predicted in each accession were combined into a single set of CNV candidates, which were again evaluated in each accession. Seeds were identified on the basis of copy-variable positions (CVPs) and deviation from sequencing depth and expected coverage¹¹. A seed had to satisfy the following criteria: at least one CVP: per-position coverage greater than $1.5 \times$ expected coverage; average coverage within the entire seed greater than

$1.8 \times$ expected coverage and greater than $1.3 \times$ sequencing depth; and minimum length 50 bp. Seeds within 100 bp from each other were merged, retaining only regions greater than 999 bp. For evaluation of relative coverage of CNVs, read per kilobase per million (RPKM) values were recorded in each accession and Col-0, using the total number of aligned reads for each accession as the denominator.

Population-genetic analyses. GENEPOP 4.0 (ref. 57) was used to calculate Weir and Cockerham's F_{ST} estimator⁵⁸ for all SNPs. Pairwise F_{ST} between geographic regions and F_{ST} across all regions were computed in sliding windows. For population structure and detection of admixture, STRUCTURE 2.3 (ref. 59) was applied to 10,000 SNPs, randomly selected at 10-kb minimum intervals from all nonsingleton SNPs for which complete information was available. STRUCTURE was run in the admixture model and repeated ten times per k with a burn-in of 50,000 iterations followed by 50,000 MCMC iterations. To estimate the most likely value of k , we averaged the clustered-ness statistic⁶⁰ over all repeats per k . We merged the different Q matrices with CLUMPP⁶¹ and plotted population assignment with *distrupt*⁶². Principal-component analysis of all SNPs was performed with EIGENSOFT⁶³. The number of SNPs to be discovered by sequencing additional strains was estimated as described⁶⁴.

Polymorphism statistics and tests for selection. Haplotype diversity²⁹ was calculated using all SNPs with a minor allele frequency of at least 5% and complete information. Haplotypes were defined as clusters of identical genotypes in a 10-kb window. Tajima's D ⁶⁵ and Fay and Wu's H ^{66,67} were calculated from all SNPs for which there was complete information and for which the likely ancestral state was available from comparison with *A. lyrata*. We have also developed a modified version of the integrated haplotype score (iHS⁶⁸), the weighted iHS (wiHS³⁰). We modified the original extended haplotype homozygosity statistic (EHH)⁶⁹ to account for unequally related strains or population structure in the sample with a matrix of pairwise distances between all strains. The contribution of each to the test statistic is weighted based on its uniqueness. We measure the uniqueness as average distance to all other strains with the same allele divided by the sum of all pairwise distances. In the classical EHH, all strains contribute equally. For the weighted EHH, the contribution of each strain is multiplied by its uniqueness value, which should reduce the bias caused by unequal relatedness in the sample, and then an integral over the weighted EHH scores around each focal SNP is calculated. All subsequent steps are identical to those in the original iHS approach⁶⁸. The positions of SNPs on a genetic map were estimated using LDhat⁷⁰.

LD and recombination rates. The ANALYSIS package⁷¹ was used to calculate LD, using SNPs with information in at least 78 strains. A custom R script was used to plot LD decay. The population recombination rate ρ was estimated using a penalized likelihood within a Bayesian reversible-jump Markov-chain Monte Carlo scheme (rjMCMC) implemented in the Interval program of LDhat⁷⁰, using SNPs with information in at least 72 strains.

Deleterious substitutions. All amino acid polymorphisms were mapped onto their orthologous protein sequences from *A. lyrata* to polarize the alleles. The *A. lyrata* protein sequence was then used as query for a PSI-BLAST⁷² search of the TrEMBL database⁷³ to construct a multiple alignment of homologous proteins. A neighbor-joining tree of these sequences was calculated with SEMPHY⁷⁴. The multiple alignments and trees were used as input for MAPP⁴³.

MIRNA locus coverage analysis. Genes were called absent if all positions were completely within SV deletions, had no calls or had zero coverage.

SNP imputation. To construct the reference panel, we excluded structural variants and considered only SNPs with at least 90% information. The imputation system used is based on BEAGLE⁷⁵. To improve local adaptation of the model, the genotype prediction was carried out in 5-Mb windows. A padding overlap of 1 Mb was added to the left and right of each such window to avoid boundary effects. Imputation accuracy was assessed on imputed genotypes, requiring different confidence levels of the imputation system.



51. Mirouze, M. *et al.* Selective epigenetic control of retrotransposition in *Arabidopsis*. *Nature* **461**, 427–430 (2009).
52. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
53. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
54. Lee, S., Cheran, E. & Brudno, M. A robust framework for detecting structural variations in a genome. *Bioinformatics* **24**, i59–i67 (2008).
55. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
56. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
57. Rousset, F. Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux. *Mol. Ecol. Resources* **8**, 103–106 (2008).
58. Cockerham, C.C. & Weir, B.S. Covariances of relatives stemming from a population undergoing mixed self and random mating. *Biometrics* **40**, 157–164 (1984).
59. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
60. Rosenberg, N.A. *et al.* Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet.* **1**, e70 (2005).
61. Jakobsson, M. & Rosenberg, N.A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801–1806 (2007).
62. Rosenberg, N.A. Distruct: a program for the graphical display of population structure. *Mol. Ecol. Notes* **4**, 137–138 (2004).
63. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
64. Ionita-Laza, I., Lange, C. & Laird, N.M. Estimating the number of unseen variants in the human genome. *Proc. Natl. Acad. Sci. USA* **106**, 5008–5013 (2009).
65. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
66. Fay, J.C. & Wu, C.I. Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405–1413 (2000).
67. Zeng, K., Fu, Y.X., Shi, S. & Wu, C.I. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* **174**, 1431–1439 (2006).
68. Voight, B.F., Kudaravalli, S., Wen, X. & Pritchard, J.K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
69. Sabeti, P.C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
70. Auton, A. & McVean, G. Recombination rate estimation in the presence of hotspots. *Genome Res.* **17**, 1219–1227 (2007).
71. Thornton, K. Libsequence: a C. class library for evolutionary genetic analysis. *Bioinformatics* **19**, 2325–2327 (2003).
72. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
73. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
74. Friedman, N., Ninio, M., Pe'er, I. & Pupko, T. A structural EM algorithm for phylogenetic inference. *J. Comput. Biol.* **9**, 331–353 (2002).
75. Browning, B.L. & Browning, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009).