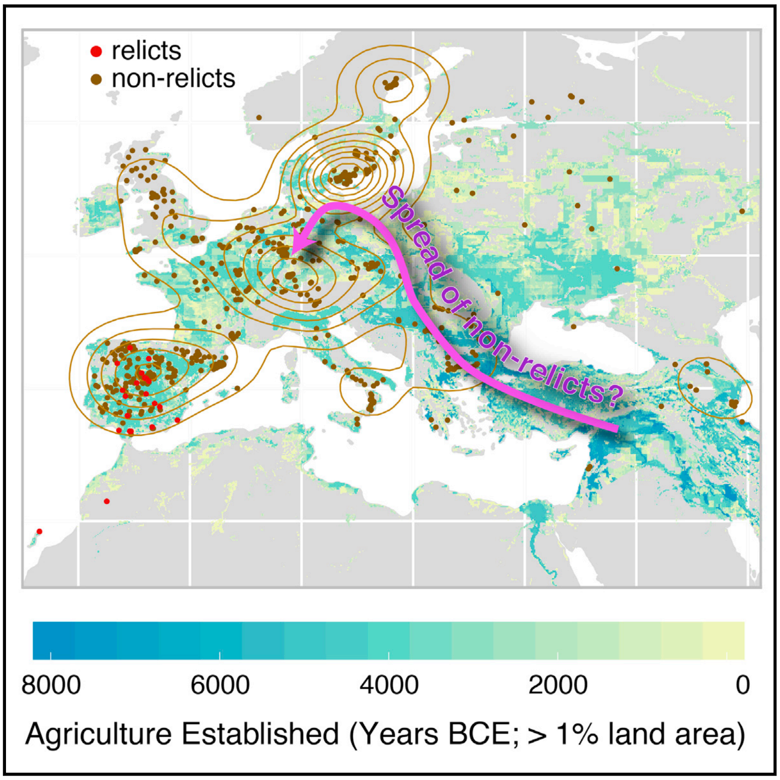
Cell

Resource

1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*

Graphical Abstract



Highlights

- The genomes of 1,135 naturally inbred lines of *Arabidopsis thaliana* are presented
- Relict populations that continue to inhabit ancestral habitats were discovered
- The last glacial maximum was important in structuring the distribution of relicts
- This collection will connect genotypes and phenotypes on a species-wide level

Authors

The 1001 Genomes Consortium

Correspondence

magnus.nordborg@gmi.oeaw.ac.at
(Magnus Nordborg),
weigel@weigelworld.org (Detlef Weigel)

In Brief

Genomic sequencing analysis of over 1,000 natural inbred lines of *Arabidopsis thaliana* reveals its global population structure, migration patterns, and evolutionary history and provides a rich genetic resource for studying phenotypic variation and adaptation.

Accession Numbers

CS78942 (ABRC), SRP056687 (NCBI SRA)

The 1001 Genomes Consortium, 2016, Cell *166*, 481–491
July 14, 2016 © 2016 The Author(s). Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2016.05.063

1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*

The 1001 Genomes Consortium[1,*]
[1]Max Planck Institute for Developmental Biology, Spemannstrasse 35, 72076 Tübingen, Germany
*Correspondence: magnus.nordborg@gmi.oeaw.ac.at (Magnus Nordborg), weigel@weigelworld.org (Detlef Weigel)

http://dx.doi.org/10.1016/j.cell.2016.05.063

SUMMARY

Arabidopsis thaliana serves as a model organism for the study of fundamental physiological, cellular, and molecular processes. It has also greatly advanced our understanding of intraspecific genome variation. We present a detailed map of variation in 1,135 high-quality re-sequenced natural inbred lines representing the native Eurasian and North African range and recently colonized North America. We identify relict populations that continue to inhabit ancestral habitats, primarily in the Iberian Peninsula. They have mixed with a lineage that has spread to northern latitudes from an unknown glacial refugium and is now found in a much broader spectrum of habitats. Insights into the history of the species and the fine-scale distribution of genetic diversity provide the basis for full exploitation of *A. thaliana* natural variation through integration of genomes and epigenomes with molecular and non-molecular phenotypes.

INTRODUCTION

Arabidopsis thaliana remains at the forefront of modern genetics. Decades of work have not only established much of what we know about the physiology and development of plants but also provided insight into how wild populations adapt to biotic and abiotic environments. Few systems share the key advantage of *A. thaliana* for GWAS or complementary forward genetics approaches: the ready availability of a large collection of naturally inbred lines (accessions) that are products of natural selection under diverse ecological conditions. This makes it possible to link genotypes and phenotypes to fitness effects in the laboratory and the field (Aranzana et al., 2005; Atwell et al., 2010; Fournier-Level et al., 2011; Hancock et al., 2011). By adding molecular data for genetically identical individuals—e.g., RNA expression or epigenetic marks— underlying mechanisms can be elucidated much more easily than in other multicellular species.

The fundamental resource for this research program is a set of accessions with complete genome sequences, collected from different locales. Systematic characterization of genome-wide polymorphism in *A. thaliana*, paralleling efforts in humans (Birney and Soranzo, 2015), began with a description of linkage disequilibrium (Nordborg et al., 2002) and population structure in 96 accessions (Nordborg et al., 2005). This was followed by a whole-genome map of deletions and SNPs in 20 global accessions (Clark et al., 2007), which in turn was the basis of a 250k SNP array with multiple markers in each haplotype block (Kim et al., 2007). This array was used to genotype the RegMap collection of 1,307 diverse accessions (Horton et al., 2012). Concurrent with the application of short read sequencing to human genomes, the first *A. thaliana* genomes were resequenced (Ossowski et al., 2008), soon followed by the analysis of larger collections (Cao et al., 2011; Gan et al., 2011; Long et al., 2013; Schmitz et al., 2013). Similar efforts have led to large-scale surveys of sequence diversity in other plants, mostly crops (Chia et al., 2012; Huang et al., 2012; Lin et al., 2014; 100 Tomato Genome Sequencing Consortium, 2014; 3000 Rice Genomes Project, 2014; Zhou et al., 2015).

We extend these efforts with 1,135 *A. thaliana* accessions from a worldwide hierarchical collection. There were several motivations for the current study: to quantify genome variation in a larger and more representative sample of accessions; to investigate the demographic history of the species; to identify features that make specific geographic or genetic subsets particularly well suited for forward genetics, field experiments and selection scans; and to provide a powerful GWAS platform. Previous studies had shown that the ability to detect footprints of selection depended greatly on the sample (e.g., Cao et al., 2011; Long et al., 2013; Huber et al., 2014). Similarly, while GWAS have identified common alleles with major effects from as few as 96 accessions (Aranzana et al., 2005; Atwell et al., 2010), a much larger sample is required for most traits. The SNP-genotyped RegMap panel (Horton et al., 2012) provided such a collection but did not efficiently capture all SNPs and structural variants. Fully sequencing this collection would be of limited benefit, as one could accurately impute the missing data by sequencing a subset. We therefore assembled a set of accessions that sufficiently overlap the RegMap panel for imputation of variants in all lines. The combined collection constitutes a first-rate resource for determining how genetic variation translates into phenotypic variation.

RESULTS AND DISCUSSION

The Sample

We selected accessions for Illumina short read sequencing with several objectives in mind. We sought to cover the global distribution of *A. thaliana* more evenly than the RegMap panel (Horton et al., 2012) while including large regional collections of particular interest from ecological and evolutionary perspectives, notably from Sweden and the Iberian Peninsula (Figure 1A). We also

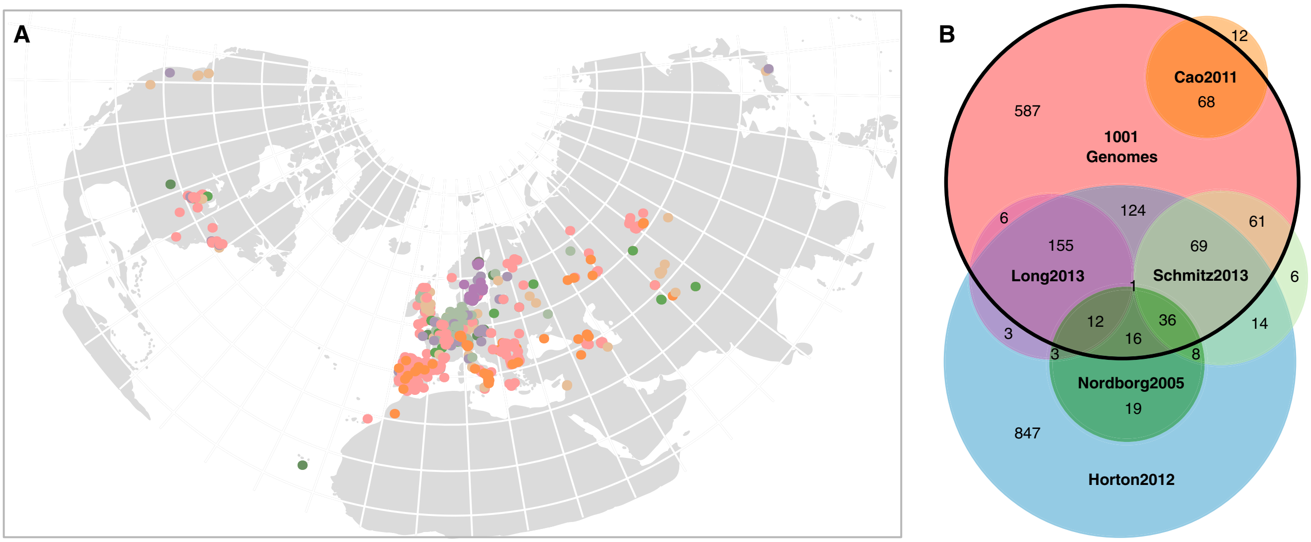


Figure 1. Origins of the 1001 Genomes Accessions

(A) Collection locations of the 1001 Genomes accessions by diversity set (colors correspond to Venn diagram in B).

(B) Relationships between 1001 Genomes accessions and other *A. thaliana* diversity sets (Nordborg et al., 2005; Cao et al., 2011; Horton et al., 2012; Long et al., 2013; Schmitz et al., 2013).

wanted to better sample interesting regions based on prior knowledge of the population structure of the species, such as North America and Central Asia (Sharbel et al., 2000; Nordborg et al., 2005; Schmid et al., 2003; Beck et al., 2008; Platt et al., 2010; Cao et al., 2011; Brennan et al., 2014; Long et al., 2013). Our collection is hierarchical, with a range of geographic distances between nearest neighbors, and a few very densely sampled locales. Most accessions had been genotyped with 149 genome-wide intermediate-frequency SNP markers (Platt et al., 2010) to avoid sequencing identical individuals.

After filtering (described below), we retained sequences of 413 RegMap and 722 new lines, for a total of 1,135 accessions with whole-genome information (see the Data Release section). These 1,135 lines are the focus of this paper; the imputed RegMap set will be described in another paper. Together, the RegMap and 1001 Genomes samples include 2,029 natural *A. thaliana* accessions with high-quality polymorphism data (Figure 1B).

The genomes presented here integrate previously published subsets (Cao et al., 2011; Gan et al., 2011; Horton et al., 2012; Long et al., 2013; Schmitz et al., 2013; Hagmann et al., 2015; Figure 1B). All accessions are available from the stock centers, and we have generated an accession list (see Data Release section) that unifies previous naming schemes and provides provenance information. Our intention is for this collection to remain actively curated as ever more accurate genomes are produced and a wide range of phenotypic data are generated (not only by us, but also by the community— see www.1001genomes.org for information on how to contribute).

The Genomes

A range of Illumina platforms were used across several sequencing centers and over several years, so we instituted stringent quality controls to pare an initial set of over 1,200 sequenced genomes to a final set of 1,135 (see Data Release section). The data are the intersection of the MPI (SHORE) and GMI (GATK) pipelines, independently validated in our pilot studies (Cao et al., 2011; Long et al., 2013). An average of 100 Mb (84%) per line were called against the TAIR10 reference genome (119 Mb). The missing positions differ greatly between lines, such that only 2% of the reference genome lack calls entirely. Based on comparisons with one long read (Pacific Biosciences) and three short read (Illumina) based de novo genome assemblies, we estimate that fewer than 3% of SNP calls are erroneous (i.e., should be reference instead) independently of dataset coverage, with the vast majority being singletons. Over 98% of genotype calls were correct at SNP sites, and only 1.5% of SNPs were mistakenly called as reference (Supplemental Experimental Procedures). We emphasize, however, that this calculation ignores SNPs missed because they are in the vicinity of structural variants, which are difficult to assess with short read technology.

After filtering, the nuclear genomes contained 10,707,430 biallelic SNPs and 1,424,879 small-scale indels (up to 40 bp). This represents one variant on average every 10 bp of the single copy genome, which is the densest variant map for any organism, including the most recent release of the 1000 Genomes Project for humans (1000 Genomes Project Consortium, 2015). 2,842 biallelic SNPs were called in chloroplast genomes and 824 in mitochondrial genomes. The complete data are available as VCF files and as FASTA pseudogenomes (see Data Release section). We also developed web applications to subset the full genome VCF or pseudogenome files and extract data on a selection of genomes and/or specific loci as well as a “Strain ID” application, with which users can identify the genomes in our sample that are most closely related to a newly sequenced

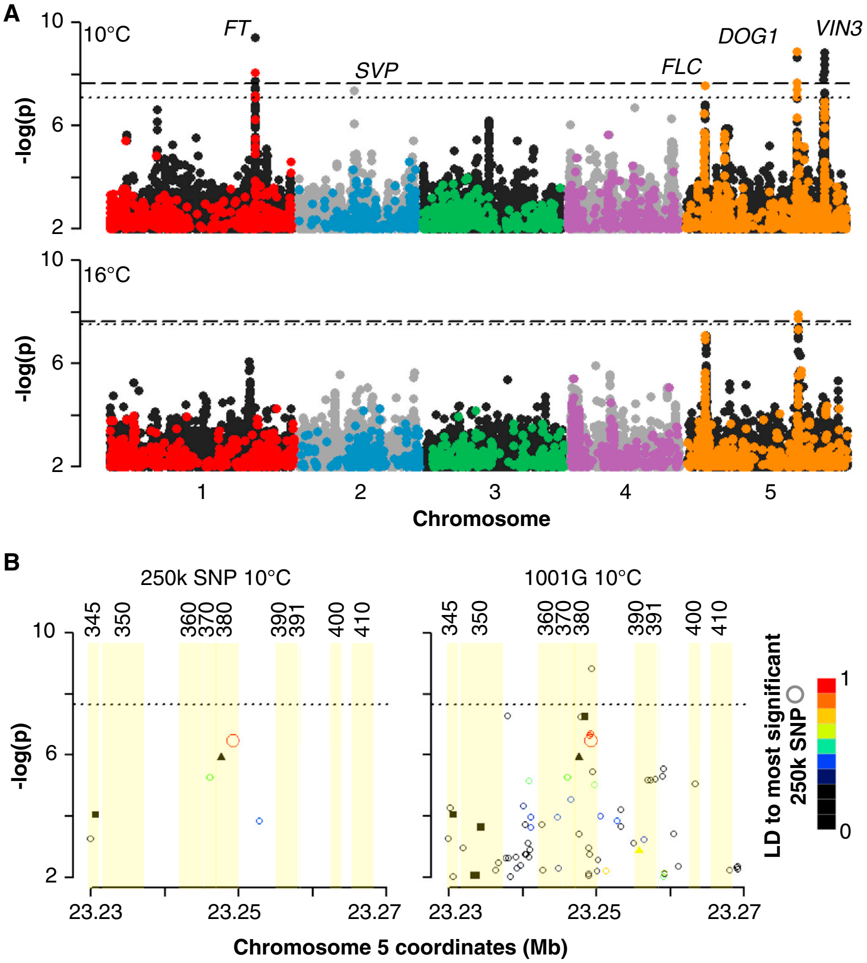


Figure 2. Comparison of GWAS for Flowering Time Using Full Genome Variants and RegMap SNPs

(A) Long day flowering time GWAS with four replicates in 1,003 (10°C) and 971 (16°C) lines. Horizontal lines represent 5% significance thresholds corrected for multiple testing using Bonferroni (dashed) and permutations (dotted). Black and gray dots are all 1001G variants, colored dots the subset also found on the RegMap 250k array.

(B) Comparison of GWAS results near flowering time regulator *VIN3* (At5g57380) with the 180k biallelic SNPs (MAF > 0.03) from the 1001 Genomes full-genome set present on the RegMap 250k array. Numbers above are regional gene identifiers, e.g., "345" = "At5g57345." Shapes denote SNP annotation: circles are non-coding; squares are synonymous; triangles are non-synonymous. Colors represent linkage disequilibrium to the top-ranked SNP in the 250k data.

genome (see Data Release section). As with all short read data, we advise caution in using our pseudogenomes with applications in which the contiguity of DNA sequence is critical, for example, in the generation of PCR primers. Finally, we are committed to supporting the community in developing additional applications that make use of these data.

Genome-wide Association Studies

A major motivation for sequencing a large collection of accessions is to enable GWAS with nearly complete genotype information. For comparison with the RegMap data, we measured flowering time under different environments (10°C and 16°C) in our collection and performed GWAS. We note first that there is little reason not to use full genome data, as permutation-based multiple-comparison thresholds (He et al., 2014; Abney, 2015) can be used to minimize the statistical cost of additional markers (Figure 2A). The chromosome 5 peak at ~23.25 Mb nicely illustrates the advantage of the full genome data (Figure 2B). Although clearly visible using the 192,498 biallelic variants from the 250k SNP array (Horton et al., 2012), not a single SNP reaches genome-wide significance, and the peak might well have been ignored, were it not for the fact that the most significant SNP lies in an intron of the flowering time regulator *VIN3* (Sung and Amasino 2004). In contrast, the full data clearly reveal a significant peak. Notably, the lead SNP in this peak is not in linkage disequilibrium with the tag SNP from the 250k SNP array, even though the two variants are only 60 bp apart.

The remaining peaks (Figure 2A) contain the flowering regulators *FT*, *SVP*, *FLC*, all previously linked to flowering time variation (Schwartz et al., 2009; Méndez-Vigo et al., 2013; Li et al., 2014), and the dormancy regulator *DOG1*, recently shown to affect also flowering time (Huo et al., 2016). As previously noted (Atwell et al., 2010), linkage disequilibrium is normally too extensive to directly pinpoint the causative genes or variants with GWAS alone. For example, the peak on chromosome 2 that contains *SVP* also includes At2g22590, which codes for an UDP-glucosyltransferase, a family of proteins linked to the control of *FLC* expression (Wang et al., 2012).

Population Structure

GWAS provide insights into the genetic basis of natural variation. To interpret such variation, it is essential that we understand the evolutionary history of a species. For an organism such as *A. thaliana*, the simplest population genetics model is strict isolation by distance (IBD), under which the genetic distance between individuals reflects only geographic distance. This model does not fit, as the peaks of pairwise differences do not reflect geography (Figure 3A). One extreme encompasses groups of (nearly) identical individuals corresponding to inbred lineages, a result of selfing. This includes 78 North American accessions, with several smaller clusters of three to seven members, and 40 pairs of accessions that differ by fewer than 1k SNPs (Figure 3B). 60 additional pairs differ by fewer than 50k SNPs, much less than the median of 439,145 for all comparisons. Excluding North

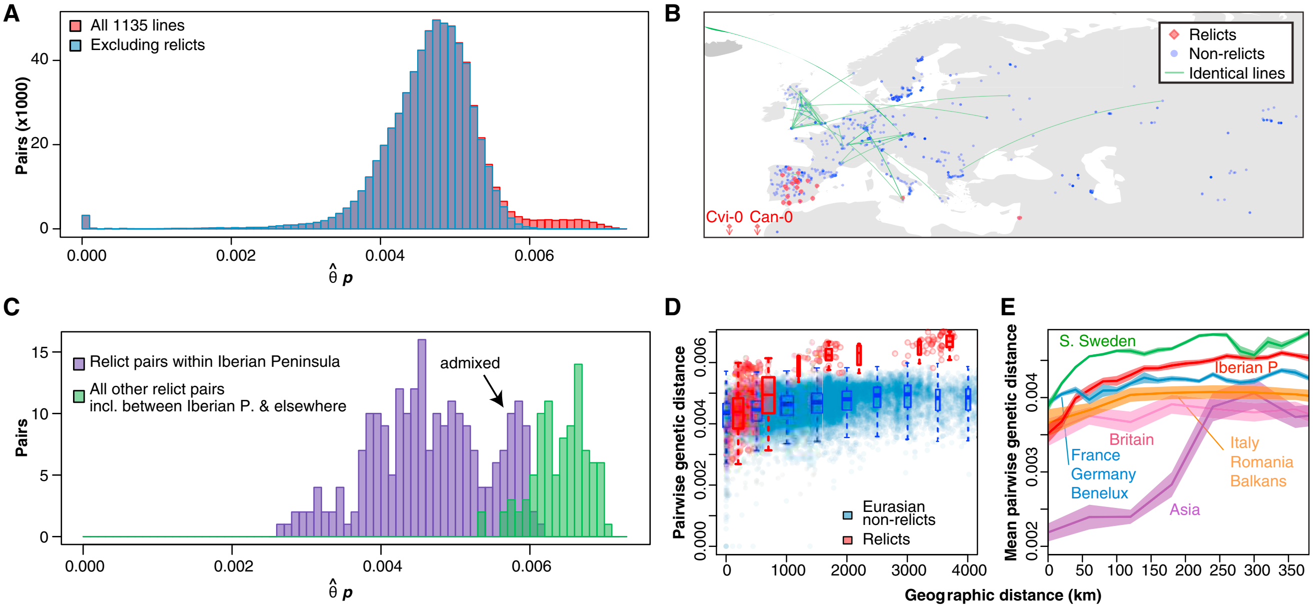


Figure 3. Genetic and Geographic Distances between Accessions

(A) The trimodal distribution of pairwise genetic distances among accessions. The mode near zero reflects very close relationships of nearly identical accessions. The mode near 0.007 includes comparisons between relicts and non-relicts.

(B) Geographic locations of relicts (red) and non-relicts (blue) in Eurasia and North Africa, with pairs of nearly identical accessions at least 1 km apart connected by green lines.

(C) Genetic distances of relict pairs. Pairwise distances between Iberian relicts are of similar magnitude as distances between global non-relicts (see Figure 3A), while the distances between relict groups from different geographic regions are higher. The second mode of high divergence for Iberian relicts is due to accessions admixed with non-relicts.

(D) Genetic distance increases globally with geographic distance for relicts but for non-relicts only over short distances. Horizontal lines indicate median, boxes include second and third quartiles, and whiskers indicate 1.5 times interquartile range.

(E) At regional scales, the rate at which genetic distance scales with geographic distance varies greatly among geographic regions for non-relicts. For each geographic region, the plot shows the genetic distance in bins of increasing geographic distance (a bin-distance of 20 km was used for S. Sweden, Iberian Peninsula, France/Germany/Benelux and 60 km bins were used for Asia, Italy/Romania/Balkans, and Britain because of uneven sampling). The shaded areas show 95% confidence intervals calculated using the ciMean function from the R package lsr.

See also Figures S1, S2, and S3.

American and British accessions, ~80% of these nearly identical pairs were collected within 1 km of each other, most within a few meters. It remains unclear whether the remaining pairs represent true long distance migration, rather than mis-assignment or mix-ups after collection (see Supplemental Experimental Procedures).

Both North America and the British Isles show evidence of recent long-range dispersal (Platt et al., 2010; Horton et al., 2012). While North America harbors a single lineage due to very recent colonization (Hagmann et al., 2015), the British Isles contains numerous widely spread genotypes, suggestive of a more ancient and gradual colonization. The median geographic distance between nearly identical British pairs is 303 km, and only 1 of 40 nearly identical pairs was collected from the same site. While some pairs may reflect labeling errors after collection, close genetic relationships are also observed among more diverged but still rather similar pairs of British accessions, supporting that they are the product of recent gene flow.

At the other end, extreme pair-wise divergences (Figure 3A) are seen with 26 accessions, including 22 from the Iberian Peninsula, and one line each from the Cape Verde Islands, Canary Islands, Sicily, and Lebanon (see Data Release section). We refer to these accessions as "relicts." The 22 Iberian relicts are no more different from each other than are pairs of non-relicts (Figure 3C). The remaining four relicts stand apart from each other and from all other accessions.

By genetic distance, our 1,135 accessions thus comprise six diverged groups: four relict groups with a single line each; one relict group of 22 Iberian accessions; and the majority group of 1,109 accessions. It should be noted that accessions Mr-0 from Italy and Tnz-1 from Tanzania also were extremely diverged, but their sequences failed quality controls and were not included in the final 1,135 accessions. Re-sequencing confirmed that Mr-0 (closely related to Sicilian relict Etna-2) and two further Tanzanian accessions, Tanz-1 and Tanz-2, are relicts. Their sequences will be available in the next data release.

The geographic distribution of relicts and non-relicts (Figure 3B) confirms that a naive IBD model cannot hold. For example, Iberian non-relicts are more closely related to accessions from Kazakhstan than to Iberian relicts. Moreover, while relicts show strong IBD on all geographic scales, non-relicts have a similarly clear pattern only over short distances (Figure 3D), as expected if they had spread rapidly to occupy their

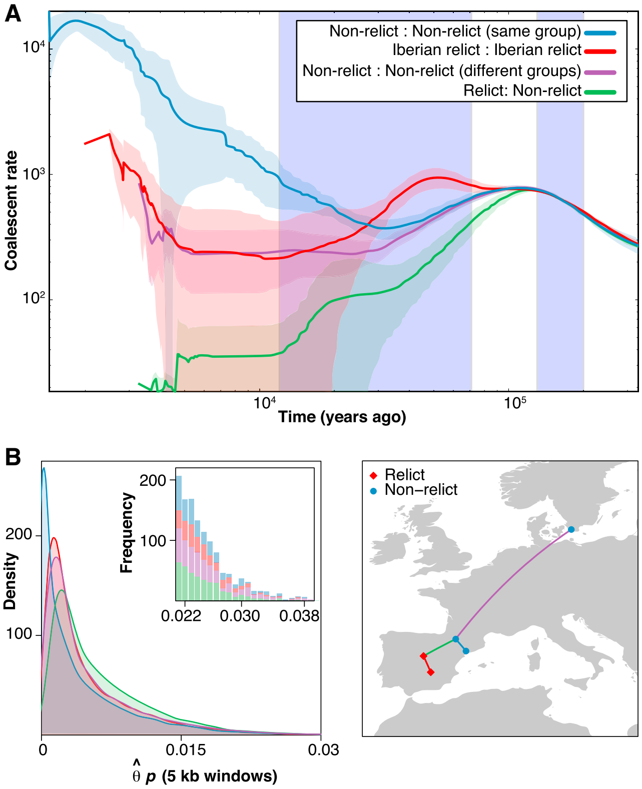


Figure 4. Evidence for the Importance of the Last Glacial Maximum in Structuring Historic and Modern Distribution of Relict and Non-relict Groups

(A) Coalescence rates over time for pairs of individuals from different ADMIXTURE groups, inferred using MSMC. Comparisons are between non-relicts from the same group (blue), Iberian relicts (red), non-relicts from different groups (purple), and relicts and non-relicts (green). The latter also includes comparisons of relicts from different geographic regions, which look similar to relict—non-relict comparisons. Solid lines indicate means, shading standard deviations. Between 49 and 62 random pairs were used. Light blue vertical bars show the last four glacial periods.

(B) Left, distributions of pairwise nucleotide diversity in 5-kb windows for four selected pairs of accessions. Colors indicate provenance of accessions, shown on right. Inset, counts in the extreme tail of the distributions.

See also Table S1.

current, global range, while the relicts had largely stayed put. On a regional scale, there is considerable geographic variation in the strength of IBD among the non-relicts, indicating that the history of colonization is complex (Figure 3E). The existence of outlier accessions, such as Cvi-0 and Mr-0, has been noted before (Nordborg et al., 2005); it is now clear that there are many such accessions, and that they can be common locally. Our data also confirm that the colonization of North America was recent and rapid (Platt et al., 2010; Horton et al., 2012; Hagmann et al., 2015). In addition to groups of nearly identical individuals, 47% of North American pairs exhibit extensive haplotype sharing (total identity-by-descent length over 85 Mb, as inferred using Beagle and GERMLINE, Figure S1) (Browning and Browning, 2009; Gusev et al., 2009), indicating recent mixing among a limited number of initial immigrants. Conversely, European accessions have low genetic relatedness, and the extent of haplotype sharing generally decays with geographic distance.

To examine population structure in greater detail, we used ADMIXTURE (Alexander et al., 2009) to cluster the accessions. In addition to identifying most of the relicts as a genetically distinct group, this analysis breaks non-relicts into eight clusters that broadly correspond to geography (see Data Release section). We defined nine groups based on these clusters and assigned each individual to a group if more than 60% of its genome derived from the corresponding cluster. The 135 individuals not matching this criterion were labeled “Admixed.” There is evidence for admixture between the relict and non-relict groups, as two accessions initially identified as relicts, from Sicily and Lebanon, were found to be admixed. These ADMIXTURE classifications were used in all subsequent analyses.

The ADMIXTURE groups do not correspond to idealized randomly mating populations. There is a regional and variable pattern of IBD (Figure 3E). Similarly, geographic locality prediction using SPA (Yang et al., 2012) demonstrates the existence of population structure both within and between groups (Figure S2) and highlights the variability in IBD (Figure S3).

To elucidate the historical processes that have shaped extant diversity, we estimated the distribution of coalescence times for the different populations using MSMC (Schiffels and Durbin, 2014). The results suggest that glacial refugia are largely responsible for present population structure (Figure 4A). Coalescence rates are an indication of relatedness, with higher rates indicating closer average relatedness (and smaller effective population size). Since the last glaciation, coalescence rates within non-relict ADMIXTURE groups were much higher than for Iberian relicts, or between members of different non-relict groups, and coalescence rates between relicts and non-relicts were essentially zero. The rate of coalescence between relicts and non-relicts was also lower than the other rates during the last glaciation, indicating that they were isolated from each other during this period. At the same time, the rate of coalescence among Iberian relicts was high, indicating a local bottleneck, with only slight differences in coalescence rate within and between non-relict groups, consistent with these groups being the product of post-glacial expansion.

In contrast, current population structure is not reflected in the rate of coalescence before the last glaciation; there has since been sufficient migration and time to erase all traces of earlier population structure. The distribution of highly diverged haplotypes at individual loci in the genome is thus independent of present population structure. The first polymorphism study in *A. thaliana*, with *ADH*, noted already the presence of surprisingly diverged haplotypes, and interpreted it as evidence for balancing selection (Hanfstingl et al., 1994). Many realized that the phenomenon was common and that deep population structure must be responsible (Aguadé, 2001; Nordborg et al., 2005; Wright and Gaut, 2005). However, the population structure required to account for pairwise sequence divergence of several percent at individual loci, compared to a genome-wide average of 0.5%, must be far older than the most recent glaciation. Thus, while recent coalescence times, as reflected in low pairwise sequence divergence, are more common in within-group comparisons (Figure 4B), the tails of extreme values look very similar for within- and between-group comparisons (Figure 4B, inset).

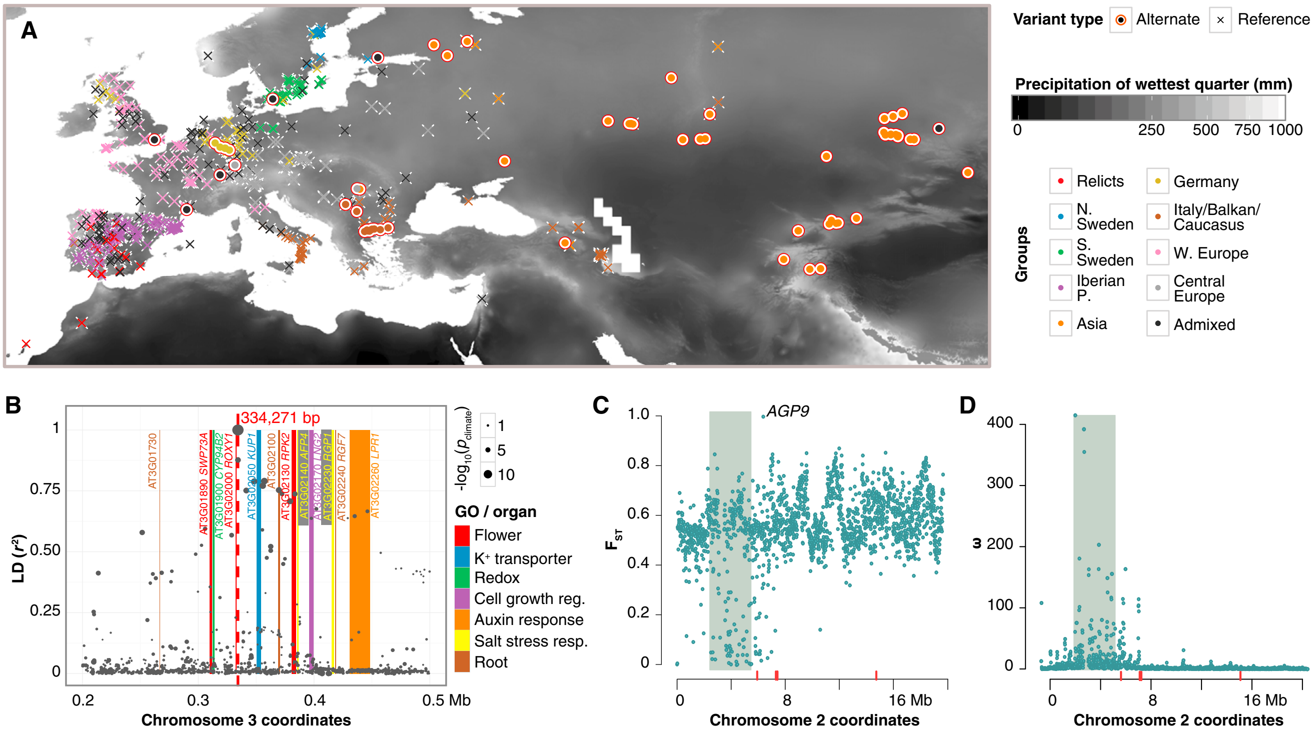


Figure 5. Footprints of Selection

(A) Distribution of accessions containing the reference or alternate variant for a locus strongly associated with precipitation in the wettest quarter. The alternate allele is most frequent in the Asian group, but it is also present in other groups.

(B) A climate associated and spatially disjunct SNP (red dashed line), located in a region densely populated with genes affecting traits such as root growth, salt tolerance, flowering, and detoxification.

(C) The distribution of maximum F_{ST} scores in 10-kb windows along chromosome 2. The centromere is shaded, and the locations of NLR-containing disease resistance genes are in red.

(D) The distribution of ω statistics in 10-kb windows along chromosome 2. Labels as in Figure 5C.

See also Figures S4 and S5 and Tables S2, S3, and S4.

It is important not to exaggerate the divergence between relicts and non-relicts. In a survey of four-sample gene genealogies between the Col-0, Ler-0, and two random Iberian relicts, 26% place Col-0 closer to one relict than to Ler-0, rather than the two non-relicts together (Table S1). Two additional four-sample analyses, one with the outgroup *A. lyrata*, Col-0, and two relicts, and the other with *A. lyrata*, one relict, Col-0 and another non-relict, produced similar patterns. While genes with the expected topology were significantly more common (43% and 53%, $p < 0.001$), many genes supported the alternative topology of a non-relict being closest to a relict.

Footprints of Selection in the Genome

The last glacial maximum was followed by pronounced expansion of the global *A. thaliana* population. It is therefore natural to search for footprints of selection related to adaptation to new environments, especially to climate, which varies considerably across the species range (Hancock et al., 2011). We examined correlations of genetic variants with six climate variables that capture variability in current temperature and precipitation using a mixed model that controls for genome-wide relatedness across samples (Zhou and Stephens, 2012). Twenty SNPs are significantly correlated with precipitation-related variables, at a False Discovery Rate of 5% (Table S2). Three associations are characterized by a much higher derived allele frequency in the Iberian relicts than in the general population, possibly indicative of local adaptation from new mutations. One affects the *ERF1* drought response regulator (Cheng et al., 2013). *ERF1* is also involved in resistance to several fungal pathogens (Berrocal-Lobo and Molina, 2004), as is *MLO11* (Acevedo-Garcia et al., 2014), which is located near two of the other variants. The connection to drought response and fungal defense suggests that selection could be due to tradeoffs between abiotic and biotic stress.

Strong local adaptation may create abrupt geographical changes in allele frequency. We used SPA (Yang et al., 2012) to search for SNPs showing this pattern, and intersected these SNPs with climate GWAS hits to identify variants with local adaptation to climate. Spatial and climatic distributions of genetic variants are intertwined with population structure (Figure S3), so almost no significant variants remain once population structure is taken into account. A single variant associated with precipitation in the wettest quarter also shows a significant geographic gradient (Figure 5A). This variant is in a genomic region densely populated with genes that have been implicated in root development and metabolism, flowering time and flower development, salt tolerance, and detoxification (Figure 5B).

Because the strategies above attempt to eliminate false positives from population structure, it is difficult to detect variants under population-specific selection. To identify such genes, we calculated F_{ST} between admixture groups for all SNPs (Weir and Cockerham, 1984). The most diverged region is on chromosome 2 at 6.401 Mb, overlapping the gene encoding *AGP9*, which has not been linked to adaptive processes before (Figures 5C and S4A). Regions adjacent to centromeres exhibit the lowest F_{ST} values. In agreement with previous results (Long et al., 2013), these regions contain excessive linkage disequilibrium (ω, Figures 5D and S4B), which suggests that they have been shaped by selective sweeps or background selection.

In addition to these global patterns, we identified loci that may contribute to adaptive differences between Iberian relicts and non-relicts. We paired each relict with the geographically closest non-relict (Table S3). Over 100 variants have diverged between the two groups, including several in or near *EIN2*, a development and stress regulator (Alonso et al., 1999), and *AP2*, which is important for flower and seed development (Licausi et al., 2013). Additional genes with differentially fixed variants are *LUG* and *SLK1*, which encode transcriptional co-repressors that interact biochemically and genetically with each other and with *AP2* (Sridhar et al., 2004; Bao et al., 2010). Finally, a deeply diverged region around 18.796 Mb on chromosome 2 includes two flowering time regulators, *AGL6* and *SOC1* (Samach et al., 2000; Huang et al., 2013). As expected from these candidates of selection, the top biological processes (GO terms) strongly overrepresented in these results are ‘‘flower development’’ and ‘‘ABA-related activities’’ (Table S4). Consistent with differentiation in flowering time, relicts flower in 10°C long days on average 21 days later than their nearest non-relicts ($t = 4.69$, $df = 41$; $p = 3 \times 10^{-5}$), suggesting that life-history differences contributed to the spread of non-relicts.

Demographic history can affect the efficacy of selection, and mutations that are likely to be deleterious are common in *A. thaliana*, especially in marginal populations (Cao et al., 2011). We therefore predicted the impacts of coding sequence variants in different genetic groups using SNPeff (Cingolani et al., 2012). Most genes, 27,525, contained at least one variant likely to change protein function, with 17,692 having at least one high-impact variant. On average, 440 genes per accession, for a total of 15,060 genes, had at least one variant predicted to inactivate the gene, although this is likely an overestimate, as it does not account for compensatory mutations or different transcript isoforms (Gan et al., 2011; Schneeberger et al., 2011; Long et al., 2013). Relicts have proportionally the most genes with potentially deleterious mutations, consistent with a reduced efficiency of selection in the relicts due to small effective population size, with the caveat that mapping to a non-relict reference may again lead us to overestimate such variants (Figure S5).

Conclusions

The Natural History of *A. thaliana*

The exquisite detail with which we have characterized the spatial pattern of polymorphism in *A. thaliana* has clarified prior hypotheses and revealed surprising aspects of the species' history. In particular, the crucial importance of the last ice age has come into much sharper relief (Sharbel et al., 2000; Nordborg et al., 2005; Schmid et al., 2003; Beck et al., 2008; François et al., 2008; Picó et al., 2008). The picture that emerges is that modern *A. thaliana* is a complex mixture of survivors from multiple glacial refugia, with population expansion having strongly favored the descendants of a specific refugium, possibly as a result of human activity. Under this model, the ‘‘relict’’ accessions are simply those that survived this expansion/invasion. Several lines of evidence support this interpretation.

The pattern of isolation-by-distance suggests that relict populations have been relatively stationary while the non-relicts' range rapidly expanded (Figure 3D). Consistent with this model, the climate at the relict locations has changed much less since the last glacial maximum than where modern non-relicts are found (Figures 6A and S6). The Iberian Peninsula is especially interesting, given the presence of a large number of relicts interspersed with non-relicts (Figure 3C). Although relicts are widely distributed there, they are restricted to a very specific environment characterized by old oak and pine forests, high climate seasonality, high temperatures, and low rainfall (Figure 6A). Iberian relicts correspond to a genetic lineage that has been previously identified in the southwestern Mediterranean region, supporting the idea that they survived in a glacial refugium in North Africa (Brennan et al., 2014). In contrast, non-relicts are found more often in agricultural and urban areas, consistent with expansion of non-relicts having been associated with human activity, and with the relative rarity of relicts reflecting destruction of undisturbed habitats.

The source of the non-relicts, which comprise most modern *A. thaliana* individuals, remains obscure. The Iberian Peninsula has the largest regional diversity, and Mediterranean regions tend to be more diverse than other regions (Figure 6B). There is a gradient of decreasing diversity from south to north (Figure 6C), as expected after a range expansion from southern glacial refugia (Petit et al., 2003). However, this pattern is likely due to admixture between relicts and invading non-relicts in these regions (high diversity in the Iberian Peninsula almost certainly is) and does not reveal the origin of the invaders. Indeed, omitting relict and admixed accessions, it would be easy to come to the conclusion that the center of diversity is southern Sweden and that diversity decreases from north to south across the entire range. The relatively high diversity seen in southern Sweden (Figure 6D), and also in Russia (Figure 6B) may similarly be the result of admixture, in this case between the original post-glacial colonizers and more recent weedy varieties that accompanied the spread of agriculture, perhaps giving rise to the higher values of Tajima's D in these regions. Resolving this issue using only contemporary collections will be difficult.

One pattern that does seem clear is that longitudinal gradients of regional diversity are much weaker than latitudinal ones (Figure 6C), most likely reflecting the relative ease with which organisms in Eurasia can move along the east-west axis. The spread of *A. thaliana* and other weedy species may have been further enhanced by the rapid expansion of agriculture along this axis (François et al., 2008). Of particular interest in this respect are unusual populations, such as those in North America, which was colonized only a few centuries ago, and in Central Asia, where

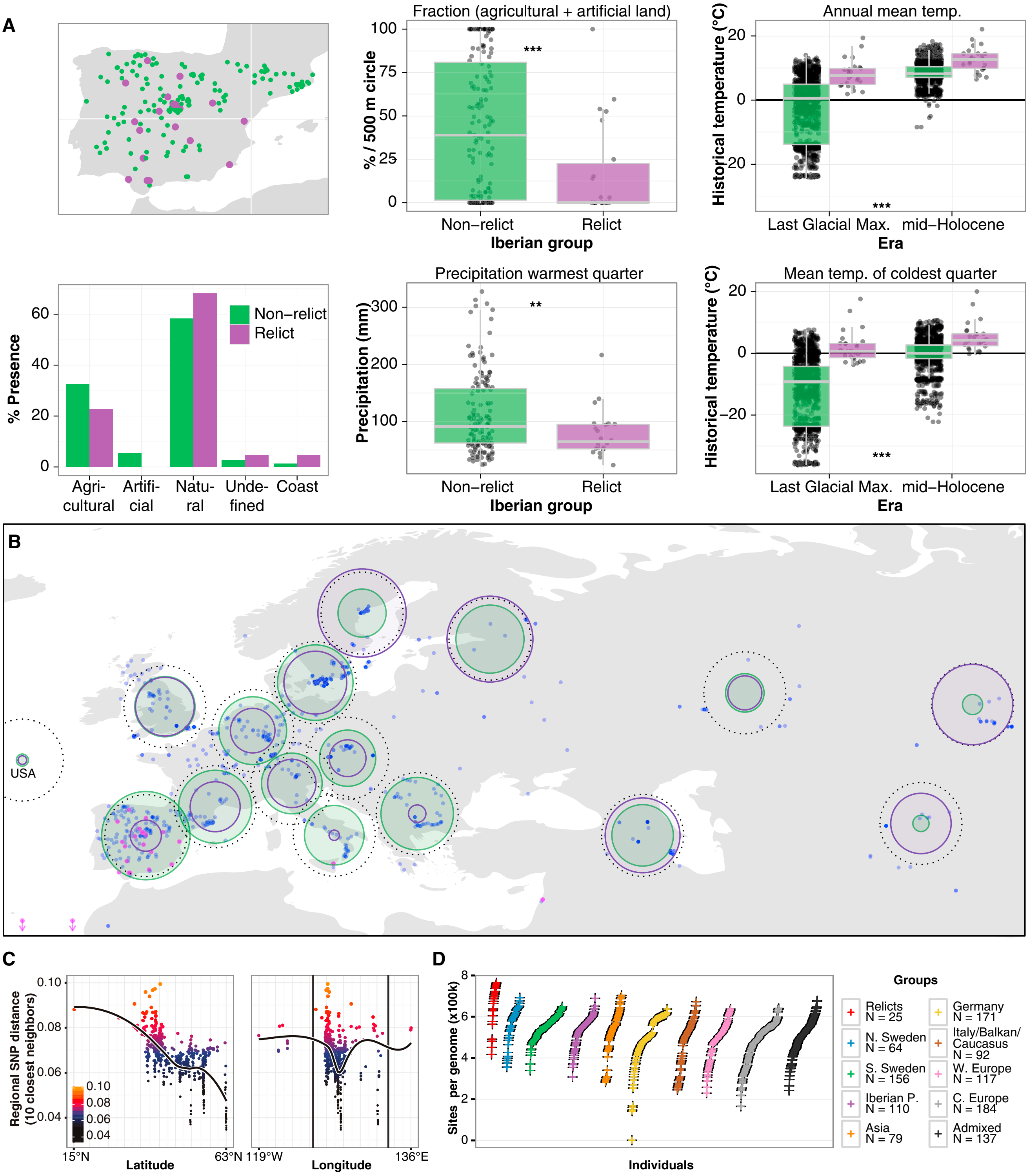


Figure 6. Local Genetic Diversity in Different Regions and Groups

(A) Current land use, current and paleoclimate for relicts and non-relicts. Relicts are purple (**p < 0.01; ***p < 0.001). Horizontal lines indicate median, boxes include second and third quartiles, and whiskers indicate 1.5 times the inter-quartile-range.

(legend continued on next page)

reduced genetic differentiation suggests rapid expansion over large geographic distances from a few very small and remote glacial refugia.

A High-Quality Community Resource

Questions one can address with natural accessions of *A. thaliana* include how patterns of genetic and epigenetic diversity arose and which forces drive adaptation to the environment. In addition, our knowledge of fundamental molecular processes can be greatly improved through the study of natural change-of-function alleles (Weigel and Nordborg, 2015). Crucial for these purposes is a well characterized, curated, and publicly available collection of accessions. We provide such a collection. Using it as a starting point, increasingly detailed information about (epi)genomes and molecular and non-molecular phenotypes can now be generated. This is a sharp distinction from similar efforts in outcrossing organisms, in which immortalized genotypes are either only available as cell lines, or do not represent adapted genotypes sampled from nature.

The selection of accessions for ecological field studies and laboratory experiments should take into account their full genetic backgrounds. No subset will be optimal for all purposes. A more diverse sample will contain more genetic heterogeneity, which reduces genetic mapping power but captures more variants. The other extreme is represented by the lineage that has recently colonized North America; while it is phenotypically quite uniform, its low diversity provides an opportunity to study the role of de novo variation in adaptation. One should also take into account the natural history of accessions, including local ecology and climate, which may enable informed decisions about phenotypic variation that is likely to reflect adaptation. For example, temperature and precipitation vary greatly across the species' range and between groups (Figure S6), and one would expect differences in physiological and developmental responses of Spanish and Swedish accessions.

Few, if any, systems offer the benefits of *A. thaliana*: a myriad of sequenced, clonal lineages from a range of ecologically diverse habitats, with patterns of linkage disequilibrium favorable for GWAS, all in an experimentally tractable organism. Another dimension can now be added to traditional functional genomics databases: adaptive variation. The 1001 Genomes collection provides an outstanding opportunity to decipher how genetic variation translates into phenotypic variation and to study the many ways in which plants respond—and have responded—to environmental challenges.

EXPERIMENTAL PROCEDURES

Sequencing and Primary Analysis

We initially selected 1,227 worldwide accessions based on genotyping (Platt et al., 2010; Horton et al., 2012) and geographic diversity (Beck et al., 2008; Brennan et al., 2014; Hagmann et al., 2015). They were sequenced by Weigel (MPI), Nordborg (GMI), Ecker (Salk), Mott (Oxford), and Monsanto. Bergelson (University of Chicago) generated the bulk of the seed and tissue used. Paired-end (PE) sequencing employed several generations of the Illumina platform: 1.3+ (80 accessions), 1.5+ (396 accessions), and 1.8+ (751 accessions).

Variants were called with MPI-SHORE (Ossowski et al., 2008) and GMI-GATK (v1.6-5, DePristo et al., 2011) pipelines, validated in our pilot studies (Cao et al., 2011; Long et al., 2013). We generated intersection VCF files with high quality in both pipelines. A series of quality checks resulted in a final set of 1,135 accessions, used for further analyses unless mentioned differently. Variant calls were benchmarked using whole-genome alignments of one long read (Pacific Biosciences) and three short read (Illumina) de novo assemblies against the TAIR10 reference. The average true positive rate (TPR) was 98%, the average false negative rate (FNR) 1.5%, the false discovery rate (FDR) 3%, independent of coverage depth used for the variant calls (Table S5). Pseudogenomes were generated by combining reference and variant calls, including indels.

Population Genetic Analyses

Please see Supplemental Experimental Procedures for details.

Data Release

Data and tools are available at http://1001genomes.org. We uploaded raw reads in FASTQ format for 1,135 final accessions to NCBI SRA (SRP056687). We are releasing the following files at http://1001genomes.org/data/GMI-MPI/releases/v3.1: full VCF variant files for each accession, VCF files with quality reference calls, a combined Full Genome VCF file for all genomes, a standard merged group VCF file without invariant positions, a variant annotated SnpEff VCF file, and individual pseudogenome files. Several tools to facilitate the use of this data are available under http://tools.1001genomes.org, including a strain ID web application, a viewer pf ADMIXTURE group membership, and a tool to retrieve specific regions of pseudogenomes in FASTA. Accession metadata, including group membership, are available under http://1001genomes.org/tables/1001genomes-accessions.html. See supplemental data release for additional tools and datasets.

ACCESSION NUMBERS

Seeds from direct progeny or siblings of sequenced individuals were deposited with the *Arabidopsis* Biological Resource Center (ABRC), where these were multiplied once. The entire set of accessions is available under accession ID CS78942. The raw sequencing reads for the sequenced individuals have been uploaded in FASTQ format to NCBI SRA (SRP056687).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, and five tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2016.05.063.

CONSORTIA

The members of The 1001 Genomes Consortium for this project are Carlos Alonso-Blanco, Jorge Andrade, Claude Becker, Felix Bemm, Joy Bergelson, Karsten M. Borgwardt, Jun Cao, Eunyoung Chae, Todd M. Dezwaan, Wei Ding, Joseph R. Ecker, Moises Exposito-Alonso, Ashley Farlow, Joffrey Fitz, Xiangchao Gan, Dominik G. Grimm, Angela M. Hancock, Stefan R. Henz, Svante Holm, Matthew Horton, Mike Jarsulic, Randall A. Kerstetter, Arthur Korte, Pamela Korte, Christa Lanz, Cheng-Ruei Lee, Dazhe Meng, Todd P. Michael,

(B) The geographic distribution of average pairwise distance (π) and Tajima's D. Sizes of the green circles indicate regional π (range from 0.002 [USA] to 0.006 [Iberian Peninsula]). Dotted circles indicate the global value, 0.006. Size of purple circles represent the regional values of Tajima's D (range from -1.01 [Northern Sweden] to -2.08 [USA], global value -2.04). Blue dots indicate sampling sites.

(C) Regional diversity as a function of latitude or longitude.

(D) Rank ordered distribution of non-private variants in each accession by ADMIXTURE group, offset to show density.

See also Figure S6.

Richard Mott, Ni Wayan Muliyati, Thomas Nägele, Matthias Nagler, Viktoria Nizhynska, Magnus Nordborg, Polina Yu. Novikova, F. Xavier Picó, Alexander Platzer, Fernando A. Rabanal, Alex Rodriguez, Beth A. Rowan, Patrice A. Salomé, Karl J. Schmid, Robert J. Schmitz, Ümit Seren, Felice Gianluca Sperone, Mitchell Sudkamp, Hannes Svardal, Matt M. Tanzer, Donald Todd, Samuel L. Volchenboum, Congmao Wang, George Wang, Xi Wang, Wolfram Weckwerth, Detlef Weigel, Xuefeng Zhou.

AUTHOR CONTRIBUTIONS

J.B., J.R.E., M.No., M.S., and D.W. coordinated the project. C.A.-B., C.B., J.B., J.C., E.C., T.M.D., J.R.E., A.M.H., S.H., M.H., A.K., P.K., N.W.M., M.Na., T.N., M.No., P.N., F.X.P., B.A.R., K.J.S., F.G.S., M.M.T., D.T., W.W., and D.W. selected and generated the samples. C.B., J.B., J.C., X.G., C.L., B.A.R., M.J., R.A.K., T.P.M., R.M., V.N., R.J.S., F.G.S., M.S., S.L.V., and X.Z. generated and handled sequence data. J.A., F.B., A.F., D.G.G., D.M., P.Y.N., A.P., F.A.R., A.R., C.W., and X.W. performed primary sequence analyses and variant annotation. M.E.-A., W.D., A.F., A.M.H., S.R.H., M.H., A.K., C.-R.L., M.No., H.S., and G.W. performed population genetic analyses. J.F., P.K., A.K., A.P., Ü.S., and C.W. curated the online resources and databases. K.M.B., D.G.G., A.K., M.No., and P.A.S. performed GWAS. M.E.-A., A.F., F.B., A.M.H., M.H., A.K., C.-R.L., M.No., A.P., H.S., C.W., G.W., and D.W. wrote the manuscript.

ACKNOWLEDGMENTS

We thank A. Vijayaraghavan, G. Hightower, and E. Thomas for help with plant growth and data handling and generation and T. Karasov for critical reading of the manuscript. Funded in part by fellowships from EU Marie Curie program (A.M.H.) and Alexander von Humboldt Foundation (B.A.R.), grants BIO2013-45407-P (C.A.-B.) and CGL2012-33220/BOS (F.X.P.) from Ministerio de Economia y Competitividad from Spain, BBSRC BB/F022697/1 and Wellcome Trust WT090532/Z/09/Z (R.M.), DFG SCHM1354-7-1 (K.J.S.), NIH and NSF (J.B., J.R.E.), Austrian Science Fund P 26342 and I 1022 (M.No., T.N., W.W.), the ERC (MAXMAP, M.No.; IMMUNEMESIS, D.W.), a collaborative grant from Austrian Science Fund and DFG (SPP ADAPTOMICS; M.No., D.W.), Austrian Academy of Sciences (M.No.), and Max Planck Society (D.W.). K.M.B. and D.W. are shareholders of Computomics GmbH. J.C. is an employee of Dow AgroSciences LLC. T.M.D., R.A.K., T.P.M., M.S., M.M.T., D.T., X.Z. are current or former employees of Monsanto Company. J.F. is an employee of Tropic IT Ltd. X.W. is an employee of Bayer CropScience AG.

Received: December 18, 2015
Revised: April 20, 2016
Accepted: May 17, 2016
Published: June 9, 2016

REFERENCES

Abney, M. (2015). Permutation testing in the presence of polygenic variation. Genet. Epidemiol. *39*, 249–258.

Acevedo-Garcia, J., Kusch, S., and Panstruga, R. (2014). *Magical mystery tour*: MLO proteins in plant immunity and beyond. New Phytol. *204*, 273–281.

Aguadé, M. (2001). Nucleotide sequence variation at two genes of the phenylpropanoid pathway, the *FAH1* and *F3H* genes, in *Arabidopsis thaliana*. Mol. Biol. Evol. *18*, 1–9.

Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.

Alonso, J.M., Hirayama, T., Roman, G., Nourizadeh, S., and Ecker, J.R. (1999). EIN2, a bifunctional transducer of ethylene and stress responses in *Arabidopsis*. Science *284*, 2148–2152.

Aranzana, M.J., Kim, S., Zhao, K., Bakker, E., Horton, M., Jakob, K., Lister, C., Molitor, J., Shindo, C., Tang, C., et al. (2005). Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. PLoS Genet. *1*, e60.

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A.M., Hu, T.T., et al. (2010). Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature *465*, 627–631.

Bao, F., Azhakanandam, S., and Franks, R.G. (2010). SEUSS and SEUSS-LIKE transcriptional adaptors regulate floral and embryonic development in Arabidopsis. Plant Physiol. *152*, 821–836.

Beck, J.B., Schmuths, H., and Schaal, B.A. (2008). Native range genetic variation in *Arabidopsis thaliana* is strongly geographically structured and reflects Pleistocene glacial dynamics. Mol. Ecol. *17*, 902–915.

Berrocal-Lobo, M., and Molina, A. (2004). Ethylene response factor 1 mediates *Arabidopsis* resistance to the soilborne fungus *Fusarium oxysporum*. Mol. Plant Microbe Interact. *17*, 763–770.

Birney, E., and Soranzo, N. (2015). Human genomics: The end of the start for population sequencing. Nature *526*, 52–53.

Brennan, A.C., Méndez-Vigo, B., Haddioui, A., Martínez-Zapater, J.M., Picó, F.X., and Alonso-Blanco, C. (2014). The genetic structure of *Arabidopsis thaliana* in the south-western Mediterranean range reveals a shared history between North Africa and southern Europe. BMC Plant Biol. *14*, 17.

Browning, B.L., and Browning, S.R. (2009). 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.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., et al. (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat. Genet. *43*, 956–963.

Cheng, M.C., Liao, P.M., Kuo, W.W., and Lin, T.P. (2013). The Arabidopsis ETHYLENE RESPONSE FACTOR1 regulates abiotic stress-responsive gene expression by binding to different cis-acting elements in response to different stress signals. Plant Physiol. *162*, 1566–1582.

Chia, J.M., Song, C., Bradbury, P.J., Costich, D., de Leon, N., Doebley, J., Elshire, R.J., Gaut, B., Geller, L., Glaubitz, J.C., et al. (2012). Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. *44*, 803–807.

Cingolani, P., Platts, A., Wang, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w^{1118}*; iso-2; iso-3*. Fly (Austin) *6*, 80–92.

Clark, R.M., Schweikert, G., Toomajian, C., Ossowski, S., Zeller, G., Shinn, P., Warthmann, N., Hu, T.T., Fu, G., Hinds, D.A., et al. (2007). Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. Science *317*, 338–342.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Fournier-Level, A., Korte, A., Cooper, M.D., Nordborg, M., Schmitt, J., and Wilczek, A.M. (2011). A map of local adaptation in *Arabidopsis thaliana*. Science *334*, 86–89.

François, O., Blum, M.G., Jakobsson, M., and Rosenberg, N.A. (2008). Demographic history of european populations of *Arabidopsis thaliana*. PLoS Genet. *4*, e1000075.

Gan, X., Stegle, O., Behr, J., Steffen, J.G., Drewe, P., Hildebrand, K.L., Lyngsoe, R., Schultheiss, S.J., Osborne, E.J., Sreedharan, V.T., et al. (2011). Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature *477*, 419–423.

Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., and Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. Genome Res. *19*, 318–326.

Hagmann, J., Becker, C., Müller, J., Stegle, O., Meyer, R.C., Wang, G., Schneeberger, K., Fitz, J., Altmann, T., Bergelson, J., et al. (2015). Century-scale methylome stability in a recently diverged *Arabidopsis thaliana* lineage. PLoS Genet. *11*, e1004920.

Hancock, A.M., Brachi, B., Faure, N., Horton, M.W., Jarymowycz, L.B., Sperone, F.G., Toomajian, C., Roux, F., and Bergelson, J. (2011). Adaptation to climate across the *Arabidopsis thaliana* genome. Science *334*, 83–86.

Hanfstingl, U., Berry, A., Kellogg, E.A., Costa, J.T., 3rd, Rüdiger, W., and Ausubel, F.M. (1994). Haplotypic divergence coupled with lack of diversity at the *Arabidopsis thaliana* alcohol dehydrogenase locus: roles for both balancing and directional selection? Genetics *138*, 811–828.

He, B.Z., Ludwig, M.Z., Dickerson, D.A., Barse, L., Arun, B., Vilhjálmsson, B.J., Jiang, P., Park, S.Y., Tamarina, N.A., Selleck, S.B., et al. (2014). Effect of genetic variation in a *Drosophila* model of diabetes-associated misfolded human proinsulin. Genetics *196*, 557–567.

Horton, M.W., Hancock, A.M., Huang, Y.S., Toomajian, C., Atwell, S., Auton, A., Muliyati, N.W., Platt, A., Sperone, F.G., Vilhjálmsson, B.J., et al. (2012). Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. Nat. Genet. *44*, 212–216.

Huang, X., Kurata, N., Wei, X., Wang, Z.X., Wang, A., Zhao, Q., Zhao, Y., Liu, K., Lu, H., Li, W., et al. (2012). A map of rice genome variation reveals the origin of cultivated rice. Nature *490*, 497–501.

Huang, X., Ding, J., Effgen, S., Turck, F., and Koornneef, M. (2013). Multiple loci and genetic interactions involving flowering time genes regulate stem branching among natural variants of Arabidopsis. New Phytol. *199*, 843–857.

Huber, C.D., Nordborg, M., Hermisson, J., and Hellmann, I. (2014). Keeping it local: evidence for positive selection in Swedish *Arabidopsis thaliana*. Mol. Biol. Evol. *31*, 3026–3039.

Huo, H., Wei, S., and Bradford, K.J. (2016). *DELAY OF GERMINATION1* (*DOG1*) regulates both seed dormancy and flowering time through microRNA pathways. Proc. Natl. Acad. Sci. USA *113*, E2199–E2206.

Kim, S., Plagnol, V., Hu, T.T., Toomajian, C., Clark, R.M., Ossowski, S., Ecker, J.R., Weigel, D., and Nordborg, M. (2007). Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nat. Genet. *39*, 1151–1155.

Li, P., Filiault, D., Box, M.S., Kerdaffrec, E., van Oosterhout, C., Wilczek, A.M., Schmitt, J., McMullan, M., Bergelson, J., Nordborg, M., and Dean, C. (2014). Multiple *FLC* haplotypes defined by independent *cis*-regulatory variation underpin life history diversity in *Arabidopsis thaliana*. Genes Dev. *28*, 1635–1640.

Licausi, F., Ohme-Takagi, M., and Perata, P. (2013). APETALA2/Ethylene Responsive Factor (AP2/ERF) transcription factors: mediators of stress responses and developmental programs. New Phytol. *199*, 639–649.

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., Zhang, Z., Lun, Y., Li, S., Wang, X., et al. (2014). Genomic analyses provide insights into the history of tomato breeding. Nat. Genet. *46*, 1220–1226.

Long, Q., Rabanal, F.A., Meng, D., Huber, C.D., Farlow, A., Platzer, A., Zhang, Q., Vilhjálmsson, B.J., Korte, A., Nizhynska, V., et al. (2013). Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. Nat. Genet. *45*, 884–890.

Méndez-Vigo, B., Martínez-Zapater, J.M., and Alonso-Blanco, C. (2013). The flowering repressor *SVP* underlies a novel *Arabidopsis thaliana* QTL interacting with the genetic background. PLoS Genet. *9*, e1003289.

Nordborg, M., Borevitz, J.O., Bergelson, J., Berry, C.C., Chory, J., Hagenblad, J., Kreitman, M., Maloof, J.N., Noyes, T., Oefner, P.J., et al. (2002). The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat. Genet. *30*, 190–193.

Nordborg, M., Hu, T.T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., et al. (2005). The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol. *3*, e196.

Ossowski, S., Schneeberger, K., Clark, R.M., Lanz, C., Warthmann, N., and Weigel, D. (2008). Sequencing of natural strains of *Arabidopsis thaliana* with short reads. Genome Res. *18*, 2024–2033.

Petit, R., Aguinagalde, I., de Beaulieu, J.L., Bittkau, C., Brewer, S., Cheddadi, R., Ennos, R., Fineschi, S., Grivet, D., Lascoux, M., et al. (2003). Glacial refugia: hotspots but not melting pots of genetic diversity. Science *300*, 1563–1565.

Picó, F.X., Méndez-Vigo, B., Martínez-Zapater, J.M., and Alonso-Blanco, C. (2008). Natural genetic variation of *Arabidopsis thaliana* is geographically structured in the Iberian peninsula. Genetics *180*, 1009–1021.

Platt, A., Horton, M., Huang, Y.S., Li, Y., Anastasio, A.E., Mulyati, N.W., Agren, J., Bossdorf, O., Byers, D., Donohue, K., et al. (2010). The scale of population structure in *Arabidopsis thaliana*. PLoS Genet. *6*, e1000843.

Samach, A., Onouchi, H., Gold, S.E., Ditta, G.S., Schwarz-Sommer, Z., Yanofsky, M.F., and Coupland, G. (2000). Distinct roles of CONSTANS target genes in reproductive development of *Arabidopsis*. Science *288*, 1613–1616.

Schiffels, S., and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. Nat. Genet. *46*, 919–925.

Schmid, K.J., Sorensen, T.R., Stracke, R., Törjék, O., Altmann, T., Mitchell-Olds, T., and Weisshaar, B. (2003). Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. Genome Res. *13* (6A), 1250–1257.

Schmitz, R.J., Schultz, M.D., Urich, M.A., Nery, J.R., Pelizzola, M., Libiger, O., Alix, A., McCosh, R.B., Chen, H., Schork, N.J., and Ecker, J.R. (2013). Patterns of population epigenomic diversity. Nature *495*, 193–198.

Schneeberger, K., Ossowski, S., Ott, F., Klein, J.D., Wang, X., Lanz, C., Smith, L.M., Cao, J., Fitz, J., Warthmann, N., et al. (2011). Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. Proc. Natl. Acad. Sci. USA *108*, 10249–10254.

Schwartz, C., Balasubramanian, S., Warthmann, N., Michael, T.P., Lempe, J., Sureshkumar, S., Kobayashi, Y., Maloof, J.N., Borevitz, J.O., Chory, J., and Weigel, D. (2009). Cis-regulatory changes at *FLOWERING LOCUS T* mediate natural variation in flowering responses of *Arabidopsis thaliana*. Genetics *183*, 723–732.

Sharbel, T.F., Haubold, B., and Mitchell-Olds, T. (2000). Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. Mol. Ecol. *9*, 2109–2118.

Sridhar, V.V., Surendrarao, A., Gonzalez, D., Conlan, R.S., and Liu, Z. (2004). Transcriptional repression of target genes by LEUNIG and SEUSS, two interacting regulatory proteins for Arabidopsis flower development. Proc. Natl. Acad. Sci. USA *101*, 11494–11499.

Sung, S., and Amasino, R.M. (2004). Vernalization in *Arabidopsis thaliana* is mediated by the PHD finger protein VIN3. Nature *427*, 159–164.

100 Tomato Genome Sequencing Consortium, Aflitos, S., Schijlen, E., de Jong, H., de Ridder, D., Smit, S., Finkers, R., Wang, J., Zhang, G., Li, N., et al. (2014). Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. Plant J. *80*, 136–148.

1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

3000 Rice Genomes Project (2014). The 3,000 rice genomes project. Gigascience *3*, 7.

Wang, B., Jin, S.H., Hu, H.Q., Sun, Y.G., Wang, Y.W., Han, P., and Hou, B.K. (2012). UGT87A2, an Arabidopsis glycosyltransferase, regulates flowering time via *FLOWERING LOCUS C*. New Phytol. *194*, 666–675.

Weigel, D., and Nordborg, M. (2015). Population genomics for understanding adaptation in wild plant species. Annu. Rev. Genet. *49*, 315–338.

Weir, B.S., and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. Evolution *38*, 1358–1370.

Wright, S.I., and Gaut, B.S. (2005). Molecular population genetics and the search for adaptive evolution in plants. Mol. Biol. Evol. *22*, 506–519.

Yang, W.Y., Novembre, J., Eskin, E., and Halperin, E. (2012). A model-based approach for analysis of spatial structure in genetic data. Nat. Genet. *44*, 725–731.

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. *44*, 821–824.

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y., Ma, Y., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat. Biotechnol. *33*, 408–414.

Supplemental Figures

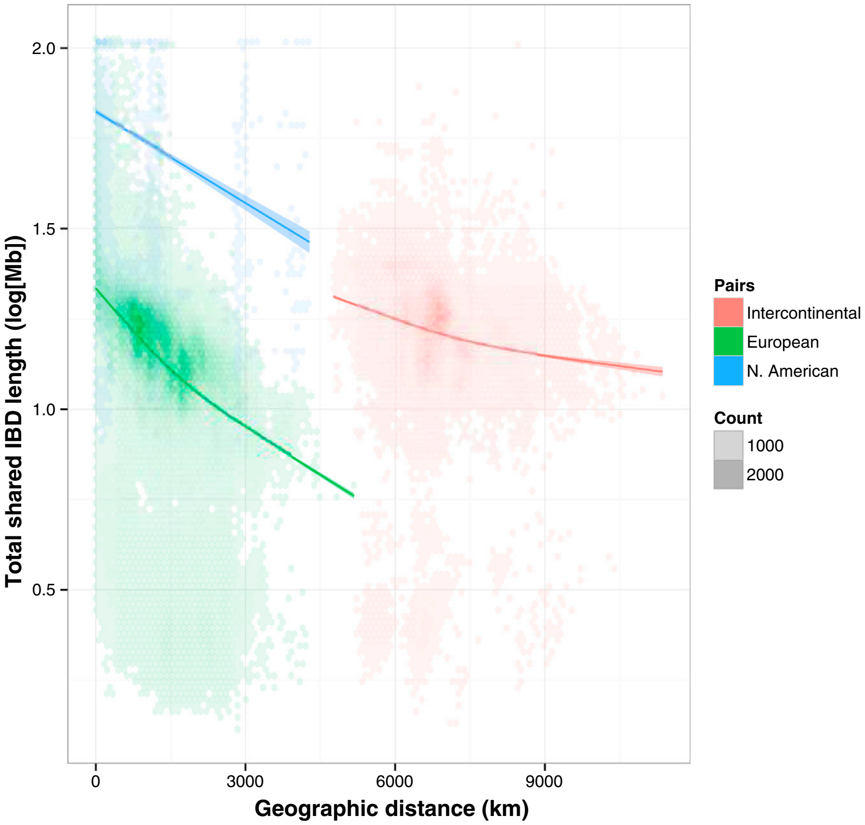


Figure S1. Overall Relationship between Total Length of Identity-by-Descent Segments and Geographical Distance in Kilometers for Each Pair in Three Groups, Related to Figure 3

In the US, many pairs share IBD segments that total over 85 Mb; in Europe, this is the case for only 0.05% of pairs, with the vast majority of pairs having IBD sharing in the range of 15-25 Mb. The minimal IBD length threshold was 10 kb. Intermediate values of intercontinental comparisons are consistent with a recent colonization of North America from European ancestors. Hexagonal bins of density, with General Additive Model predictions (k = 3, solid lines) and the 95% CI of these predictions (shaded regions).

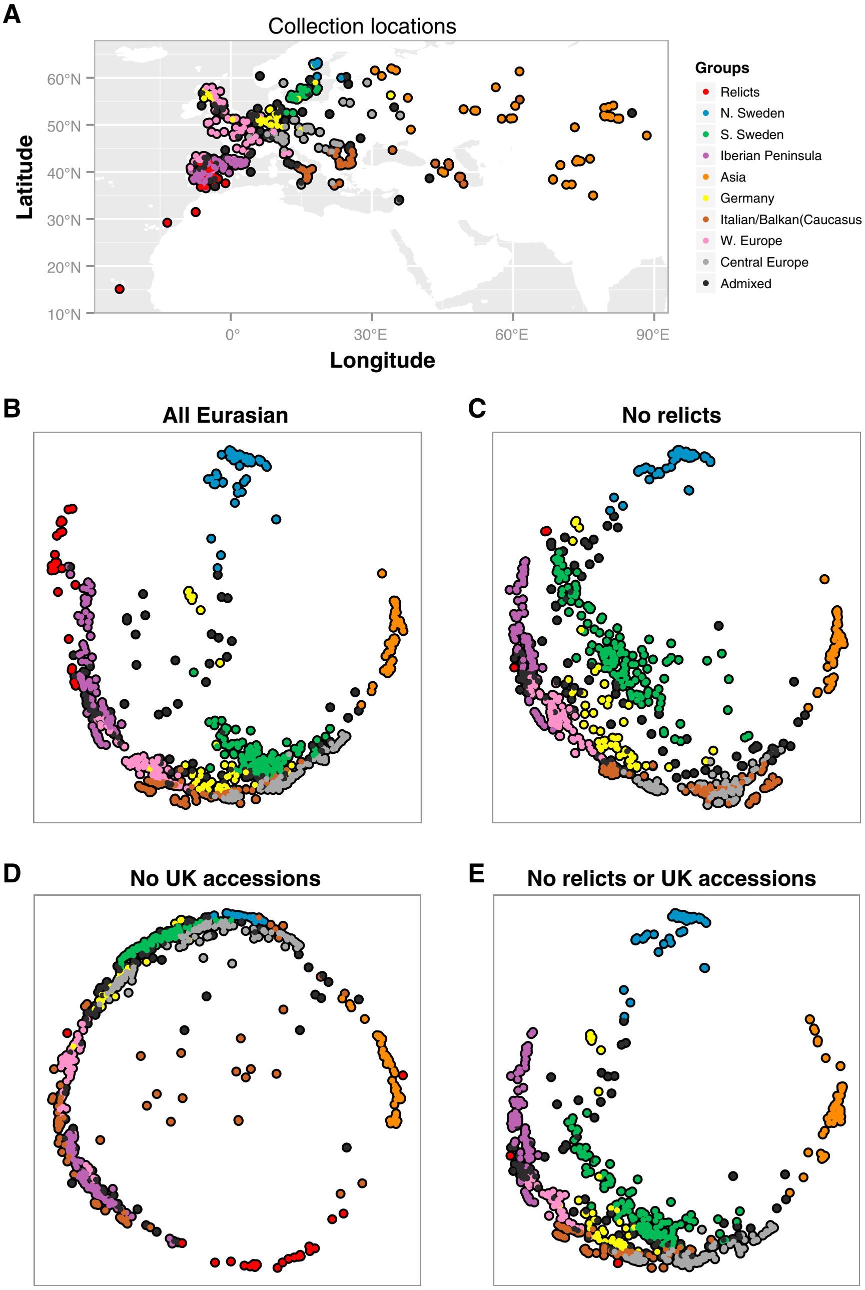
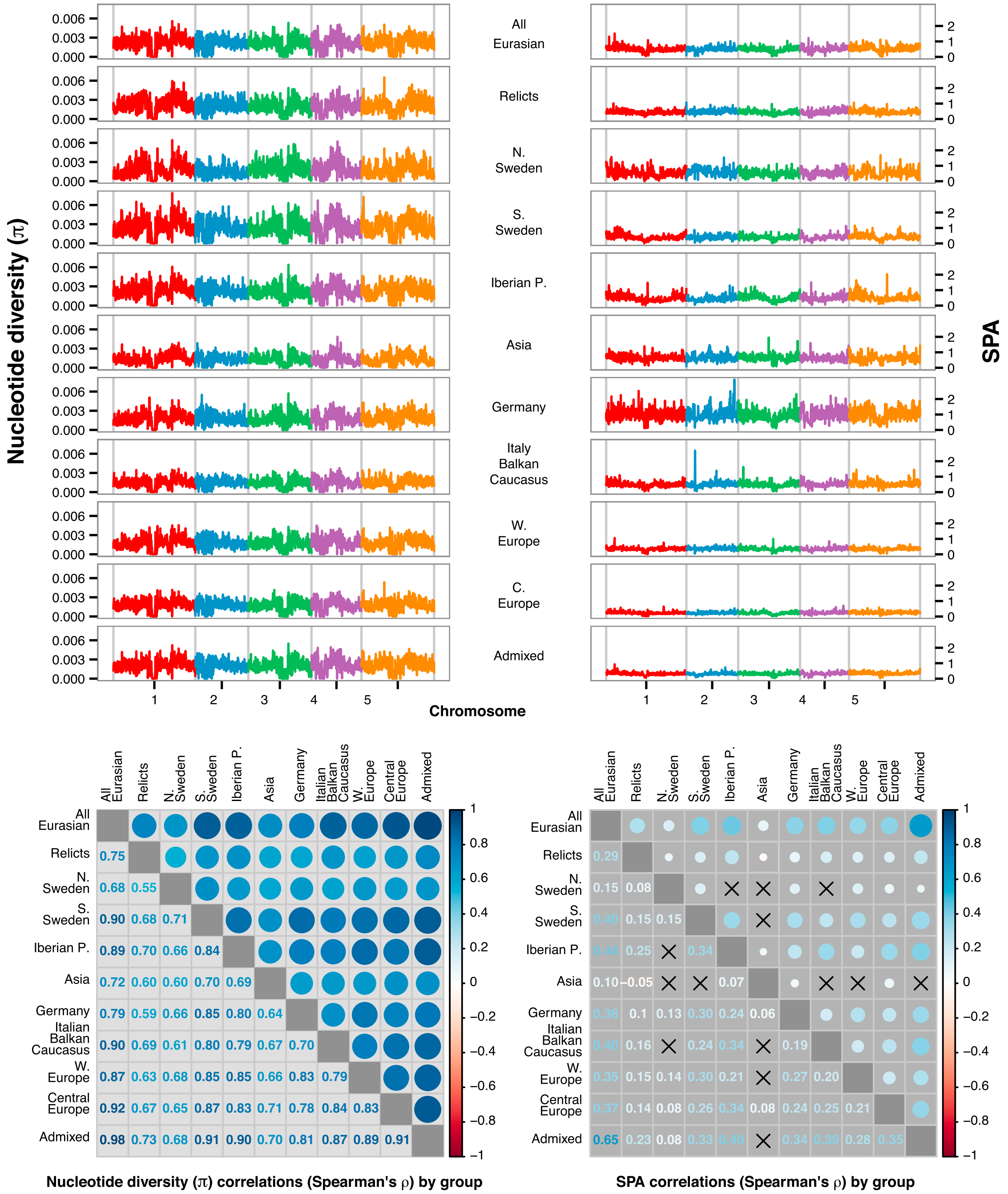


Figure S2. Geographic Prediction from SPA Suggests that a Simple Isolation-by-Distance Model Does Not Hold, Related to Figure 3

(A and B) Under this model, if geographic gradients of SNPs were smooth, the collection locations of the accessions should be recovered in the SPA analysis in B. They are not. Instead, we find strong spatial gradients between, and gentle gradients within groups.

(C and D) The strength of these gradients is disproportionately influenced by the relict and UK accessions.

(E) Finer geographic gradients of SNP variation, especially in Southern Swedish populations, are observable when relict and UK accessions are excluded. Note that unsupervised predictions from SPA analysis are translation, scale, and rotation invariant (dimensionless).



	All Eurasian	Relicts	N. Sweden	S. Sweden	Iberian P.	Asia	Germany	Italian Balkan Caucasus	W. Europe	Central Europe	Admixed
All Eurasian		●	●	●	●	●	●	●	●	●	●
Relicts	0.75		●	●	●	●	●	●	●	●	●
N. Sweden	0.68	0.55		●	●	●	●	●	●	●	●
S. Sweden	0.90	0.68	0.71		●	●	●	●	●	●	●
Iberian P.	0.89	0.70	0.66	0.84		●	●	●	●	●	●
Asia	0.72	0.60	0.60	0.70	0.69		●	●	●	●	●
Germany	0.79	0.59	0.66	0.85	0.80	0.64		●	●	●	●
Italian Balkan Caucasus	0.90	0.69	0.61	0.80	0.79	0.67	0.70		●	●	●
W. Europe	0.87	0.63	0.68	0.85	0.85	0.66	0.83	0.79		●	●
Central Europe	0.92	0.67	0.65	0.87	0.83	0.71	0.78	0.84	0.83		●
Admixed	0.98	0.73	0.68	0.91	0.90	0.70	0.81	0.87	0.89	0.91	

Nucleotide diversity (π) correlations (Spearman's ρ) by group

	All Eurasian	Relicts	N. Sweden	S. Sweden	Iberian P.	Asia	Germany	Italian Balkan Caucasus	W. Europe	Central Europe	Admixed
All Eurasian		●	●	●	●	●	●	●	●	●	●
Relicts	0.29		●	●	●	●	●	●	●	●	●
N. Sweden	0.15	0.08		●	×	×	●	×	●	●	●
S. Sweden	0.40	0.15	0.15		●	×	●	●	●	●	●
Iberian P.	0.44	0.25	×	0.34		●	●	●	●	●	●
Asia	0.10	−0.05	×	×	0.07		●	×	×	●	×
Germany	0.38	0.1	0.13	0.30	0.24	0.06		●	●	●	●
Italian Balkan Caucasus	0.40	0.16	×	0.24	0.34	×	0.19		●	●	●
W. Europe	0.35	0.15	0.14	0.30	0.21	×	0.27	0.20		●	●
Central Europe	0.37	0.14	0.08	0.26	0.34	0.08	0.24	0.25	0.21		●
Admixed	0.65	0.23	0.08	0.33	0.40	×	0.34	0.39	0.28	0.35	

SPA correlations (Spearman's ρ) by group

Figure S3. Nucleotide Diversity (π) and Spatial Gradients of Genetic Diversity (SPA) by Group in 50 kb Windows, Related to Figure 3

While overall genetic diversity across the genome is similar across genetic groups, the geographic distribution of that variability differs across groups, with the lowest associations between the North Swedish and Asian groups. Correlations that did not reach the significance threshold of p = 0.01 are marked with an “X.”

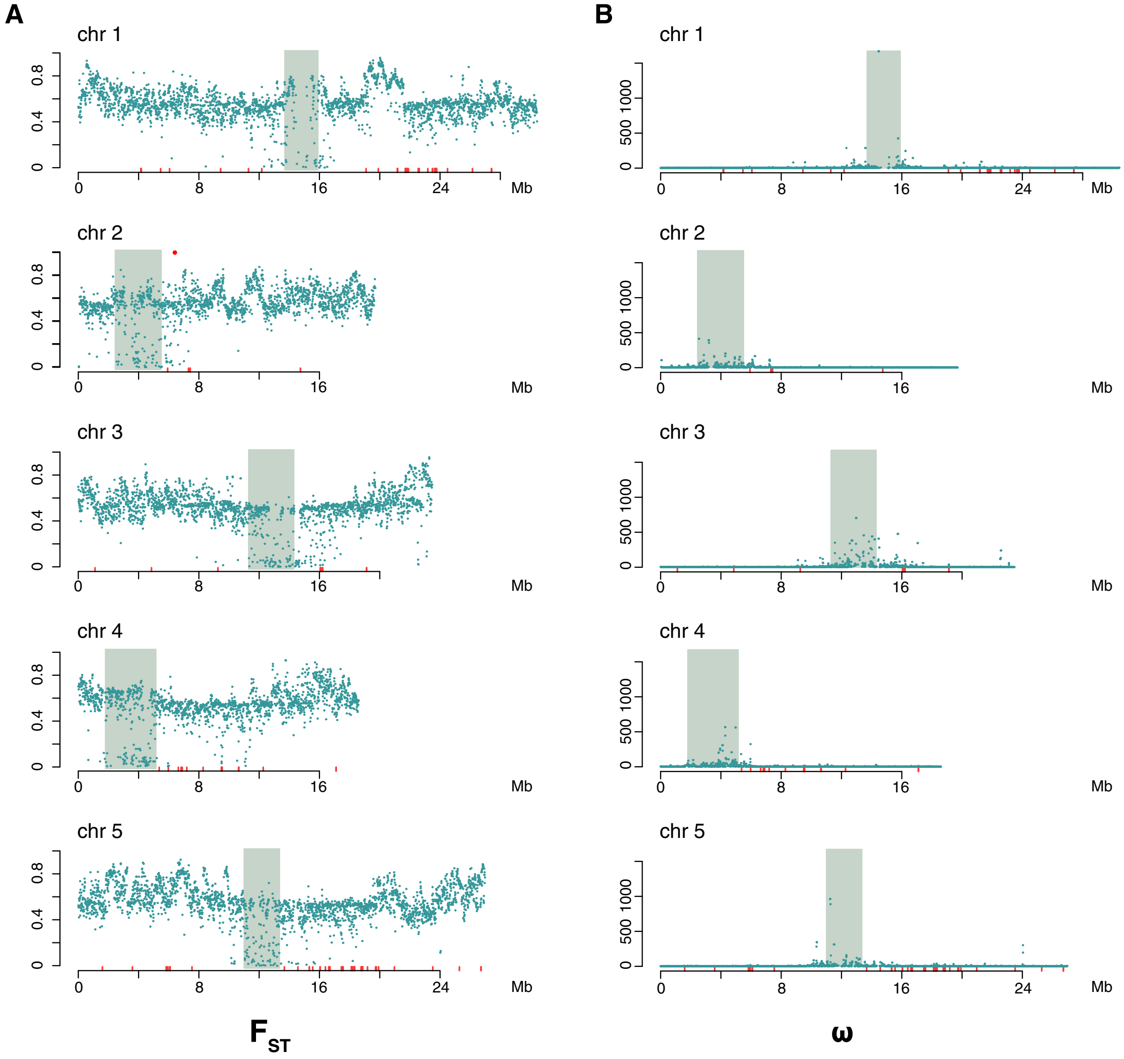


Figure S4. Genome-wide F_{ST} and ω for Each Chromosome, Related to Figure 5

(A) Distribution of maximum F_{ST} scores in 10-kb windows. The centromere is shaded in each figure, and the locations of NB-LRR genes (“resistance” or *R* genes) are shown in red.

(B) Distribution of ω-statistic in 10-kb windows. Labels as in A. On each chromosome, the lowest genome-wide F_{ST} values, and largest estimates of the ω-statistic, are near the centromeres, which suggests that selective sweeps or background selection are common in these regions of the genome.

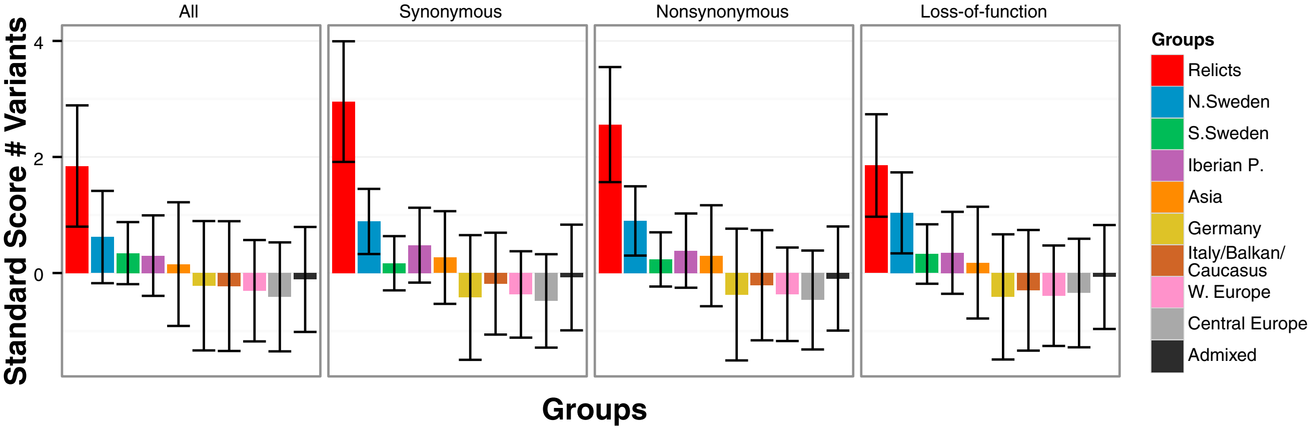


Figure S5. Variants by Type and Group, Related to Figure 5

Mean and standard deviation of the standard score (Z-score) of the number of variants of each type, by group. Relicts show the greatest normalized number of variants, especially of synonymous variants. Bars indicate means, whiskers represent one standard deviation.

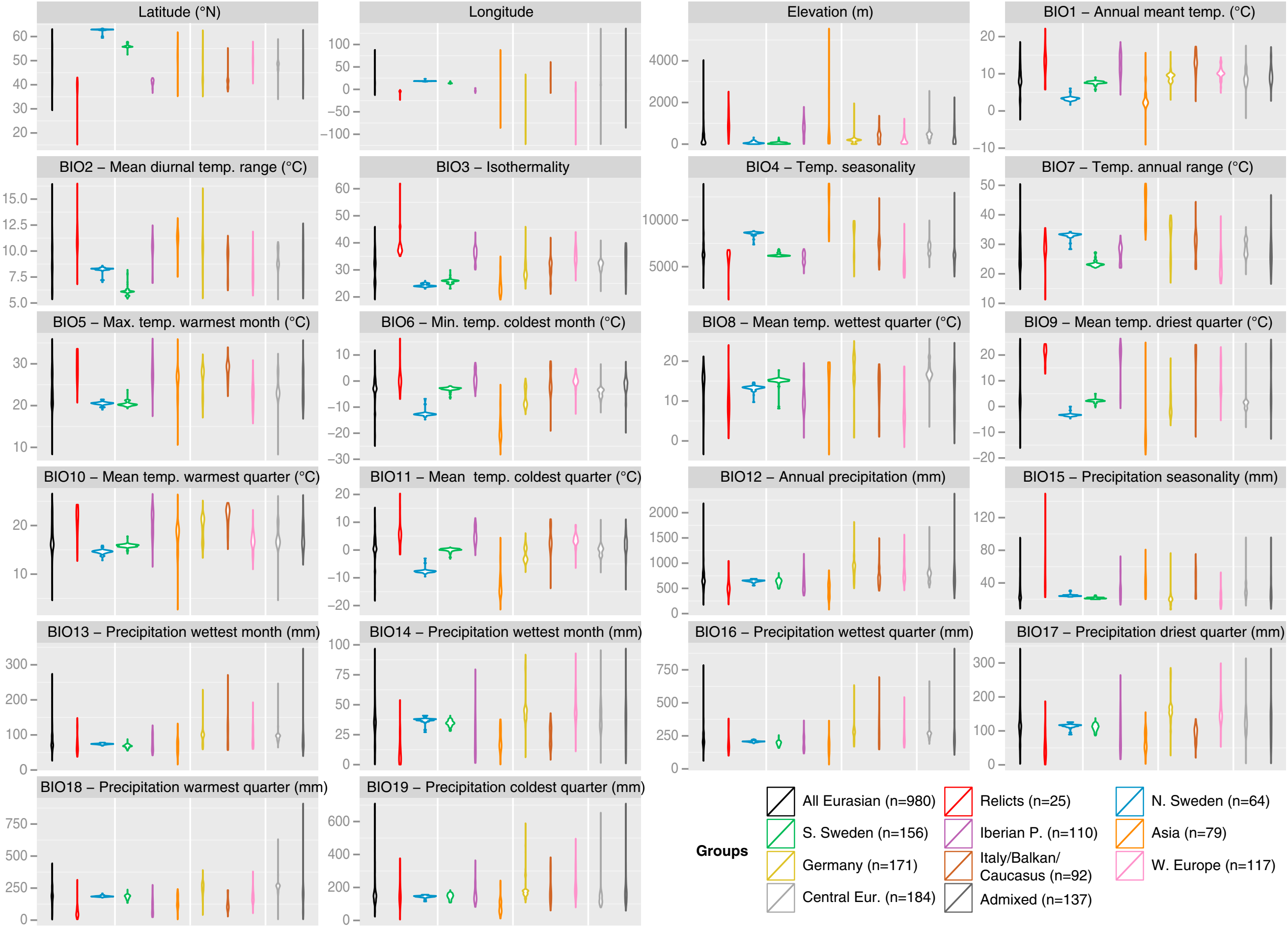


Figure S6. Climatic and Geographic Representation of Samples within Groups, Related to Figure 6

While the distribution of climate within each group is generally in proportion to the distribution of geography (latitude, longitude, and elevation), some are not. For example, although the Asian accessions are widely distributed, the range of precipitation experienced by these accessions is surprisingly narrow. Note that a few non-Eurasian accessions were nevertheless assigned to admixture groups, and are included in these distributions. Violin plots show probability densities within admixture groups for each geoclimatic variable.

Cell, Volume *166*

Supplemental Information

1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*

The 1001 Genomes Consortium

EXTENDED EXPERIMENTAL PROCEDURES

Variant calling

We used the MPI-SHORE and GMI-GATK pipelines, validated in our pilot studies (Cao et al., 2011; Long et al., 2013). Briefly, for the GMI-GATK pipeline, PE reads were mapped to the *Arabidopsis thaliana* TAIR10 reference genome with BWA (v0.5.9-r16) (Li and Durbin, 2009). Format conversions and removal of duplicated reads were done with Samtools (v0.1.6 (r453)) (Li et al., 2009). Local realignment around indels was performed with GATK (v1.6-5) (DePristo et al., 2011) as follows: a first round of indels, called with the UnifiedGenotyper function, was provided to the RealignerTargetCreator function to generate the set of intervals required by the IndelRealigner function. SNPs and indels were called separately, and individually for each accession with the UnifiedGenotyper function, and later merged with the function CombineVariants. Transposons were called from the source data with TE-locate with a resolution of 1000 bp and a minimal read-pair support of 5 (parameters 'minimal Distance to count' = 1000 and 'minimal supporting reads' = 5) (Platzer et al., 2012).

For the MPI-SHORE pipeline (Ossowski et al., 2008), each accession was analyzed separately, using BWA sample (v0.6.2) with option "-n 0.1" to map the reads to the TAIR10 *Arabidopsis thaliana* reference genome sequence. SNPs and short indels were called with SHORE consensus. The matrix with empirically determined penalties for various alignment features from (Cao et al., 2011) was used to calculate the quality score for consensus calls. 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 (Cao et al., 2011). Positions with SHORE quality >= 25 and minimum allele frequency of alternative base call >= 0.9 were retained as variant SNP/indel or reference calls. A SHORE2VCF script was developed to convert the SHORE specific quality values into the standard Variant Call Format (VCF) v4.1 (Danecek et al., 2011). In addition to the standard VCF file from SHORE, we also generated a quality reference VCF file with information only from SHORE quality reference calls (Ossowski et al., 2008).

Both the GMI-GATK and the MPI-SHORE pipeline produced standards VCF file for SNPs and short indels for each accessions. For the intersection VCF files, only calls that were in agreement between the two pipelines and that had a quality value >= 25 were accepted. Each intersected VCF file was integrated with the corresponding SHORE quality reference VCF file, which was then used as input for a Full Genome VCF file with the VCF merge tool (Danecek et al., 2011). A standard merged groups VCF file was generated after removing reference calls. SnpEff was used to annotate the standard merged group VCF file (Cingolani et al., 2012), which yielded the variant annotated SnpEff VCF file. All VCF files meet the standard VCF v4.1 (Danecek et al., 2011).

Chloroplast and mitochondrial variants were called using UnifiedGenotyper in GATK with ploidy set to 1, from the same BAM files generated by the GMI-GATK pipeline. Only biallelic SNPs were retained with a quality score >= 100, coverage >= 10x and with >= 75% of reads supporting the alternate allele. Mitochondrial and chloroplastic DNA copy number was estimated as the median coverage normalized by the median coverage of chr1:1-10Mb.

Quality control and validation

Sequence coverage, as determined by read depth at mapped positions, ranged between 2-118X, with 1209 lines having at least 5x and 1058 lines having at least 10x coverage. Read lengths varied from 36 bp to 143 bp, with 1097 lines having reads at least 75 bp long, and 894 lines at least 100 bp. We excluded all non-reference accessions that did not meet the following criteria from the final set: at least 5x coverage; at least 100,000 SNP calls relative to the reference; at least 80% overlap in SNPs between the GMI-GATK and MPI-SHORE pipelines; at least 50% of SNPs called homozygous in both pipelines; contradictions between the two pipelines less than 0.1%; at least 95% concordance with RegMap 250k SNP array data, unless there were reasons to believe that the 250k array data were problematic (Horton et al., 2012). Where more than one lab had sequenced the same accession, the highest quality accession was kept. Finally, accessions with doubtful geographic origin, based on clustering of whole-genome data, were removed. This resulted in a final set of 1135 accessions. Unless mentioned specifically, all analyses were based on this set.

For quality control, we *de novo* assembled three genomes using Illumina reads, from accessions L*er*-1, Ws-2 and Sha. We produced three sequence data sets for each strain: Illumina MiSeq overlapping paired-end reads (400 bp inserts, 250 bp reads, 80x-114x coverage); Illumina HiSeq 2000 mate-pair reads (7 kb inserts, 101 bp reads, 36x-158x coverage); Illumina HiSeq 2000 fosmid-end reads (40 kb inserts, 101 bp reads, 18x-86x coverage). We ran *ALLPATHS-LG* (Gnerre et al., 2011) on a combination of these datasets. We further upgraded the L*er*-1 assembly by filling and reducing scaffold gaps with Ler-1 PacBio reads (PacBio P2C2; 3.4 kb mean/2.2 kb median reads, 240x coverage). We also exploited a more recent Pacific Biosciences L*er*-1 assembly (http://datasets.pacb.com.s3.amazonaws.com/2014/Arabidopsis/reads/list.html). All *de novo* assembled genomes were aligned against the TAIR10 reference using dnadiff (Kurtz et al., 2004). Variants were called directly from whole genome alignments (WGA) using show-snps. Only one-to-one alignments with an identity of at least 90% were taken into account. WGA-based variant calls were compared with variant calls from our final 1001G dataset, bsed on combining the results from the MPI-SHORE and GMI-GATK pipelines. Indels were left-aligned and normalized prior to the comparison using the norm function from bcftools. True Positives (TP) are positions where 1001G and WGA variants are concordant, False Positives (FP) where 1001G variants were not supported by WGA (which may either show a reference call or a different variant), False Negatives (FN) where a 1001G reference call is not supported by WGA (which has a variant). The False Discovery Rate (FDR) is defined as FP/(TP+FP), the False Negative Rate is defined as FN/(FN+TP) and the True Positive Rate is defined as TP/(TP+FN). Comparing the combined variant calls with the WGA-based variant calls resulted in an average TPR of 98%, an average FNR of 1.5% and an average FDR of 3%. TPR and FDR were further monitored over a range of different coverage depths (5-45x coverage) by randomly subsetting a L*er*-1 dataset to call variants with the two pipelines against TAIR10 as described above. FDR (3±0.4%), FNR (1.5±0.1%) and TPR (98±0.1%) remained independent of the coverage depth (5-45x coverage) (Table S5).

Pseudogenomes and variant annotation

Pseudogenomes were generated by combining reference and variant calls including indels, with uncalled sites represented as Ns, plus an index for finding Col-0 annotated regions.

Variants were annotated with SnpEff (release 4.1L) and the SnpEff *A. thaliana* database (release 2015-01-08) (Cingolani et al., 2012). Derived alleles were extracted from a three way alignment of *A. thaliana* (TAIR10), *A. lyrata* and *C. orientalis* (unpublished PacBio assemblies) calculated with progressive Cactus (Paten et al., 2011a, 2011b). Sub-alignments with more than one sequence from one of the species, or one of the species missing were discarded. The remaining alignments were screened for identical sites. 35 Mb of the *A. thaliana* genome was marked as derived in this way. Allele density spectrums were smoothed and plotted with the sm package in GNU R (Lenth, 2009).

Genome-wide association studies

Seeds for all 1135 accessions were surface-sterilized with chlorine gas. Seeds were distributed in pots with four replicates in a randomized block design, each replicate corresponding to one block. Plants were grown in growth chambers with the following settings: after 6 days of stratification in the dark at 4ºC, constant temperature of 10°C or 16°C with 16 hours light / 8 hours darkness, 65% humidity. All trays within a block were moved to a new shelf and rotated 180º every other day to minimize position effects. Flowering time was scored as days until first open flower. Genome-wide association mapping was done on the means of the four replicates for both Phenotypes (10°C and 16°C) independently, using an approximation of the mixed model that has been described previously (Kang et al., 2010, pipeline available at: https://github.com/arthurkorte/GWAS).

We compared the 1001G SNPS to the 250k SNP-array data from Horton et al. (2012). Out of the 214,051 SNPs called with the 250k array, 207,096 are called as SNPs in the 1001G data, with 192,498 (93.0%) being biallelic, and 14,598 (7.0%) multi-allelic. For 530 of the biallelic SNPs, the inferred state between the two datasets differs. These are distributed across all chromosomes according to local SNP density (Supplemental Figure S7) and are unlikely to negatively affect GWAS (Supplemental Fig S8). In general, given the much higher estimated error rates in the 250k SNP-array data (Atwell et al. 2010; Horton et al., 2012), disagreement with the 1001G data are likely to result from errors in the 250k SNP-array data (Supplemental Figure S9).

Population genetic analyses

We used Beagle v3 (Browning and Browning, 2009) to impute missing SNPs based on linkage disequilibrium with default parameters, followed by GERMLINE v1.5.1 (Gusev et al., 2009) for error-tolerant and computationally efficient identification of Identity-by-Descent (IBD) regions on these imputed SNPs. Pairwise IBD segments were detected as long continuous stretches with a minimum length of 10 kb, merged from slices containing 100 identical SNPs and allowing for maximally two mismatches.

MSMC input was created parsing a VCF file including all sites (variant and non-variant). Filtered sites or sites were any of the focal individuals had missing genotypes were excluded from the count of "called sites". MSMC was run in the two haplotype mode with the option --fixedRecombination, were haplotypes of different inbred individuals were used together. Scaled times were converted to years assuming a generation time of one year and a mutation rate of 7*10^(-9) (Ossowski et al., 2010). Coalescent rate was calculated as 1/(relative effective population size).

To infer the relationship between relicts and non-relicts for all individual genes, we estimated the genealogy between two non-relict and two relict accessions. This analysis was done separately for each pairs of Iberian relicts while fixing Col-0 and L*er*-0 as the two non-relicts in the 4-taxon test. From the start of each gene, we searched for non-recombining region by performing four-gamete tests for all consecutive SNP pairs until a recombination event among the four accessions was detected. Gene genealogy was estimated from the phylogenetically informative sites (doubletons) within the non-recombining region, and genes with less than two informative sites were excluded.

Two additional four-taxon gene tree analyses were also conducted. One with *A. lyrata*, *A. thaliana* Col-0 and two relicts, and the other with *A. lyrata*, *A. thaliana* Col-0, one relict, and one non-relict. We used CONSEL (Shimodaira and Hasegawa, 2001) and RAxML (Stamatakis, 2014) to generate Maximum Likelihood gene trees, and determined their significance with the AU test (significance threshold 0.05) (Shimodaira, 2002). The deviation from null distribution was assessed for each chromosome via one-tailed Student's t-tests.

SPA geographic projections and SNP gradient analyses were performed using SPA v1.13 (Yang et al. 2012), with default parameters. One-tailed *p*-values were calculated by first transforming SPA scores to *z*-scores.

Climate data (both historical and recent) were obtained from WorldClim (www.worldclim.org). Recent data were extracted from the Current Conditions Bioclim rasters (Hijmans et al. 2005, http://www.worldclim.org/current). Historical data were extracted from the CMIP5 Multi-Model Ensemble dataset (MPI-ESM-P, http://www.worldclim.org/paleo-climate). Data was sourced from the 30 arc-second rasters, with 2.5 arc-minute rasters as a fallback if collection locations fell between raster cells.

We used a mixed model approach to identify variants associated with latitude and six representative Bioclim variables, while controlling for the potentially confounding effects of population structure. The Bioclim variables included in the analysis were annual mean temperature, annual precipitation, mean temperature during warmest quarter, mean temperature during the coldest quarter, precipitation in the wettest quarter and precipitation during the driest quarter. We used GEMMA (Zhou and Stephens, 2012) to infer the correlation between each variant with frequency >5% in the total sample and each of the seven variables. First, we estimated a relatedness matrix for each chromosome using the '-gk 1' option in GEMMA. Then, we assessed evidence for correlation in a linear mixed model framework using Rao's score test. We identified the set of variants with a false discovery rate <5% for each variable (R package: p.adjust, method="fdr") and used a comparison to the closest outgroup relative, *A. lyrata*, to determine the ancestral state.

To identify candidate selective sweeps, we used OmegaPlus (Kim and Nielsen, 2004; Alachiotis et al., 2012). The grid size was chosen to ensure the ⍰ statistic was evaluated, on average, every 500 bp while the minimum and maximum regions considered were 10-kb and 100-kb, respectively. For the F_{ST} scan, Weir and Cockerham's ⍰ (Weir and Cockerham, 1984) was calculated at each SNP, to identify genomic regions that have diverged among these groups. To identify GO terms (TAIR) overrepresented in the top 1% results from the F_{ST} scan, which were summarized in 10-kb windows, we omitted gene-models with low confidence (evidence-code: 'inferred from electronic annotation') and any biological category represented by only one gene-model. We used Storey's approach (Storey and Tibshirani, 2003) to correct for multiple testing.

Data release

Although some of the original data have been released in conjunction with prior publications (Cao et al., 2011; Gan et al., 2011; Schneeberger et al., 2011; Schmitz et al., 2013; Long et al., 2013; Hagmann et al., 2015), we uploaded raw reads in fastq format for all 1135 final accessions to NCBI SRA with id SRP056687. We are releasing the full VCF variant files from the intersection of the GMI (GATK) and MPI (SHORE) pipelines for each accession on the http://1001genomes.org project website under

http://1001genomes.org/data/GMI-MPI/releases/v3.1/intersection_snp_short_indel_vcf/.

We also produced VCF files that include information on quality reference calls:

http://1001genomes.org/data/GMI-MPI/releases/v3.1/intersection_snp_short_indel_vcf_with_quality_reference/.

The combined Full Genome VCF file that includes information for all genomes (132 GB):

http://1001genomes.org/data/GMI-MPI/releases/v3.1/1001genomes_snp-short-indel_with_tair10_only_ACGTN.vcf.gz.

The standard merged group VCF file without information on invariant positions (~18 GB):

http://1001genomes.org/data/GMI-MPI/releases/v3.1/1001genomes_snp-short-indel_only_ACGTN.vcf.gz.

The variant annotated SnpEff VCF file (~17 GB):

http://1001genomes.org/data/GMI-MPI/releases/v3.1/1001genomes_snpeff_v3.1/1001genomes_snp-short-indel_only_ACGTN_v3.1.vcf.snpeff.gz.

Individual pseudogenome files in gzipped FASTA format:

http://1001genomes.org/data/GMI-MPI/releases/v3.1/pseudogenomes/.

The imputed SNP matrix (317 MB):

http://1001genomes.org/data/GMI-MPI/releases/v3.1/SNP_matrix_imputed_hdf5/1001_SNP_MATRIX.tar.gz.

Accession metadata, including group membership:

http://1001genomes.org/tables/1001genomes-accessions.html

The phenotypes for flowering time scored at 10°C and 16°C:

http://1001genomes.org/tables/1001genomes-FT10-FT16_and_1001genomes-accessions.html

We created several tools to facilitate using this dataset under http://tools.1001genomes.org.
Our strain ID web application:

http://tools.1001genomes.org/strain_id.

A tool to download specific regions of psuedogenomes:

http://tools.1001genomes.org/pseudogenomes.

Online visualization tool to view ADMIXTURE group membership and genetic composition:

http://1001genomes.github.io/admixture-map.

SUPPLEMENTAL TABLES

Table S1. Gene tree analyses, Related to Figure 4.

Mean and standard deviation of genes with resolved four-sample topology supporting the expected or unexpected relationship among Col-0, Ler-0, and two Iberian relicts.

	Gene count	**Proportion**
Concordant	3867.23 (+- 968.83)	0.71 (+- 0.11)
Discordant	1327.98 (+- 383.91)	0.26 (+- 0.10)

Table S2. SNPs from the climate correlation analysis with FDR<0.05. Bold denotes nonsynonymous variants, Related to Figure 5.

Variable	Chr	Position	Frequency	p-value	GeneID	Gene Name
Annual precipitation	3	8201102	0.057	3.14E-07	AT3G23060	CYP705A33
	4	**13187833**	0.108	3.48E-08	AT4G25970	PSD3
	4	13191070	0.108	3.72E-08	AT4G25980	
	5	17883475	0.069	4.26E-07	AT5G44390	
	5	17883492	0.069	3.50E-07	AT5G44390	
	5	17883501	0.067	4.22E-07	AT5G44390	
	5	17883508	0.068	4.79E-07	AT5G44390	
	5	21818955	0.066	5.09E-07	Intergenic	
	5	21818969	0.068	4.11E-07	Intergenic	
Precipitation wettest quarter	1	1697341	0.081	9.48E-07	Intergenic	
	1	16373400	0.1	8.67E-07	AT1G43387	TE gene
	3	334271	0.09	5.46E-07	Intergenic	
	3	7341449	0.128	1.23E-07	AT3G20940	CYP705A30
	3	**7341468**	0.123	6.43E-07	AT3G20940	CYP705A30
	3	**7345743**	0.101	6.04E-08	AT3G20960	CYP705A33
	3	8296396	0.209	3.44E-07	AT3G23240	ERF1
	4	**9814059**	0.083	1.58E-06	AT4G17610	tRNA/rRNA methyltransferase
	4	**13187833**	0.108	8.75E-07	AT4G25970	PSD3
	4	13191070	0.108	5.22E-07	AT4G25980	
	5	9063610	0.048	5.55E-10	AT5G44390	FAD-binding
	5	9072657	0.07	2.95E-07	Intergenic	
	5	17883475	0.069	2.14E-07	AT5G44390	
	5	17883476	0.069	1.10E-06	AT5G44390	
	5	17883492	0.069	3.38E-07	AT5G44390	
	5	17883501	0.067	3.61E-07	AT5G44390	

	5	17883508	0.068	1.28E-07	AT5G44390	
	5	21818969	0.068	1.33E-06	Intergenic	

Table S3. IDs of Iberian relicts and their closest matching (unique) non-relict, Related to Figure 5.

Iberian relict	Iberian non-relict	Haversine distance (km)
9832	9862	10.15
9837	9873	13.15
9947	9855	18.92
9533	9531	19.58
9871	9841	22.33
9905	9843	26.68
9542	9822	30.13
9869	9522	26.70
9600	9943	28.90
9543	9900	87.33
9598	9578	31.93
9555	9556	36.23
9545	9544	38.67
9550	9590	36.89
9887	9534	40.23
9549	9903	49.35
9554	6961	66.95
9944	9537	140.40
9574	9514	170.62
9583	9541	221.71
9879	9518	261.95

Table S4. Top five significant Gene Ontology (GO) terms enriched in relicts compared to geographically close non-relicts. Ties among *q* - values are ranked according to results from Fisher’s Exact test, Related to Figure 5.

Rank	Biological Process	Enrichment	FDR q
1	Flower development	6.3	0.00059
2	Positive regulation of abscisic acid mediated signaling pathway	15.4	0.00059
3	Embryo sac development	7.7	0.00059
4	Embryo development	6.2	0.00059
5	Positive regulation of flower development	9.5	0.00059

Table S5. Error rate dependencies, Related to Experimental Procedures.

Error rates as a function of sequencing depth and genomic context based on read data from a single accession (L*er*).

Coverage/ Annotation	TP	FP	FN	TPR	FNR	FDR
5	185,402	4,424	3,042	98.39%	1.61%	2.33%
7	314,276	7,504	4,246	98.67%	1.33%	2.33%
9	373,409	9,233	4,982	98.68%	1.32%	2.41%
12	404,520	10,668	5,549	98.65%	1.35%	2.57%
14	423,364	11,711	5,995	98.60%	1.40%	2.69%
17	434,153	12,463	6,286	98.57%	1.43%	2.79%
20	443,682	13,168	6,504	98.56%	1.44%	2.88%
22	450,936	13,805	6,732	98.53%	1.47%	2.97%
24	456,516	14,303	6,907	98.51%	1.49%	3.04%
27	461,342	14,815	7,036	98.50%	1.50%	3.11%
29	465,272	15,269	7,199	98.48%	1.52%	3.18%
32	468,692	15,661	7,285	98.47%	1.53%	3.23%
34	471,446	16,085	7,397	98.46%	1.54%	3.30%
37	474,048	16,441	7,399	98.46%	1.54%	3.35%
39	476,257	16,710	7,494	98.45%	1.55%	3.39%
41	478,062	16,904	7,618	98.43%	1.57%	3.42%
45	479,863	17,140	7,702	98.42%	1.58%	3.45%
5’ UTR	22,823	208	245	98.94%	1.06%	0.90%
3’ UTR	28,400	381	281	99.02%	0.98%	1.32%
Exon	143,361	2,201	1,202	98.52%	0.83%	1.51%
Intron	81,572	1,176	1,126	98.64%	1.36%	1.42%
Intergenic	264,123	9,410	3,485	96.62%	2.05%	4.99%
Repetitive	99,705	9,410	3,485	96.62%	3.38%	8.62%
Non-repetitive	380,158	7,730	4,217	98.90%	1.10%	1.99%

SUPPLEMENTAL REFERENCES

Alachiotis, N., Stamatakis, A., and Pavlidis, P. (2012). OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. Bioinformatics *28*, 2274-2275.

Browning, B.L., and Browning, S.R. (2009). 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.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C.*, et al.* (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet *43*, 956-963.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain *w^{1118}; iso-2; iso-3*. Fly *6*, 80-92.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156-2158.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M.*, et al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet *43*, 491-498.

Gan, X., Stegle, O., Behr, J., Steffen, J.G., Drewe, P., Hildebrand, K.L., Lyngsoe, R., Schultheiss, S.J., Osborne, E.J., Sreedharan, V.T.*, et al.* (2011). Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature *477*, 419-423.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. USA *108*, 1513-1518.

Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., and Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. Genome Res *19*, 318-326.

Hagmann, J., Becker, C., Müller, J., Stegle, O., Meyer, R.C., Wang, G., Schneeberger, K., Fitz, J., Altmann, T., Bergelson, J.*, et al.* (2015). Century-scale methylome stability in a recently diverged *Arabidopsis thaliana* lineage. PLoS Genet. *11*, e1004920.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. Int. J. Climat. *25*, 1965-1978.

Horton, M., Hancock, A.M., Huang, Y.S., Toomajian, C., Atwell, S., Auton, A., Muliyati, W., Platt, A., Sperone, F.G., Vilhkálmsson, B.J.*, et al.* (2012). Genome-wide pattern of genetic variation in worldwide *Arabidopsis thaliana* accessions from the *RegMap* panel. Nat Genet *44*, 212-216.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. *42*, 348-354.

Kim, Y., and Nielsen, R. (2004). Linkage disequilibrium as a signature of selective sweeps. Genetics *167*, 1513-1524.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. Genome Biol. *5*, R12.

Lenth, R.V. (2009). Response-surface methods in R, using rsm. J. Stat. Software *32*.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Long, Q., Rabanal, F.A., Meng, D., Huber, C.D., Farlow, A., Platzer, A., Zhang, Q., Vilhjálmsson, B.J., Korte, A., Nizhynska, V., *et al.* (2013). Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. Nat. Genet. *45*, 884-890.

Ossowski, S., Schneeberger, K., Clark, R.M., Lanz, C., Warthmann, N., and Weigel, D. (2008). Sequencing of natural strains of *Arabidopsis thaliana* with short reads. Genome Res *18*, 2024-2033.

Ossowski, S., Schneeberger, K., Lucas-Lledo, J.I., Warthmann, N., Clark, R.M., Shaw, R.G., Weigel, D., and Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science *327*, 92-94.

Paten, B., Diekhans, M., Earl, D., John, J.S., Ma, J., Suh, B., and Haussler, D. (2011a). Cactus graphs for genome comparisons. J. Comput. Biol. *18*, 469-481.

Paten, B., Earl, D., Nguyen, N., Diekhans, M., Zerbino, D., and Haussler, D. (2011b). Cactus: Algorithms for genome multiple sequence alignment. Genome Res. *21*, 1512-1528.

Platzer, A., Nizhynska, V., and Long, Q. (2012). TE-Locate: A tool to locate and group transposable element occurrences using paired-end next-generation sequencing data. Biology *1*, 395-410.

Qi, J., Liu, X., Shen, D., Miao, H., Xie, B., Li, X., Zeng, P., Wang, S., Shang, Y., Gu, X., et al. (2013). A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. Nat. Genet. *45*, 1510-1515.

Schmitz, R.J., He, Y., Valdes-Lopez, O., Khan, S.M., Joshi, T., Urich, M.A., Nery, J.R., Diers, B., Xu, D., Stacey, G., *et al.* (2013). Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. Genome Res. *23*, 1663-1674.

Schneeberger, K., Ossowski, S., Ott, F., Klein, J.D., Wang, X., Lanz, C., Smith, L.M., Cao, J., Fitz, J., Warthmann, N., *et al.* (2011). Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. Proc. Natl. Acad. Sci. USA *108*, 10249-10254.

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. Syst. Biol. *51*, 492-508.

Shimodaira, H., and Hasegawa, M. (2001). CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics *17*, 1246-1247.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312-1313.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc Natl Acad Sci USA *100*, 9440-9445.

Weir, B.S., and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. Evolution *38*, 1358-1370.

Yang, W.Y., Novembre, J., Eskin, E., and Halperin, E. (2012). A model-based approach for analysis of spatial structure in genetic data. Nat Genet *44*, 725-731.

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. *44*, 821-824.

AUTHOR AFFILIATIONS

Carlos Alonso-Blanco[1], Jorge Andrade[2], Claude Becker[3], Felix Bemm[3], Joy Bergelson[2], Karsten M. Borgwardt[4], Eunyoung Chae[3], Todd Dezwaan[5], Wei Ding[3], Joseph R. Ecker[6], Moises Exposito-Alonso[3], Ashley Farlow[7,8], Joffrey Fitz[3,9], Xiangchao Gan[10], Dominik G. Grimm[3,4], Angela Hancock[2,11,12], Stefan R. Henz[3], Svante Holm[13], Matthew Horton[2,7], Mike Jarsulic[2], Randall A. Kerstetter[14], Arthur Korte[7], Pamela Korte[7], Christa Lanz[3], Cheng-Ruei Lee[7], Dazhe Meng[7], Todd P. Michael[5], Richard Mott[10], Ni Wayan Muliyati[2], Thomas Nägele[12,15], Matthias Nagler[12], Viktoria Nizhynska[7], Magnus Nordborg[7], Polina Yu. Novikova[7], F. Xavier Picó[16], Alexander Platzer[7], Fernando A. Rabanal[7], Alex Rodriguez[2], Beth A. Rowan[3], Patrice A. Salomé[3], Karl Schmid[17], Robert J. Schmitz[6], Ümit Seren[7], Felice Gianluca Sperone[2], Mitchell Sudkamp[14], Hannes Svardal[7], Matt M. Tanzer[5], Donald Todd[5], Samuel L. Volchenboum[2], Congmao Wang[3,18], George Wang[3], Xi Wang[3], Wolfram Weckwerth[12,15], Detlef Weigel[3], Xuefeng Zhou[5]

[1]Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas, Madrid-28049, Spain; [2]Dept. of Ecology and Evolution, University of Chicago, Chicago, IL 60637, USA; [3]Dept. of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany; [4]Dept. of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland; [5]Monsanto, Research Triangle Park, NC 27709, USA; [6]Salk Institute for Biological Studies, La Jolla, CA 92037, USA; [7]Gregor Mendel Institute (GMI), Austrian Academy of Sciences, Vienna Biocenter (VBC), 1030 Vienna, Austria; [8]Centre for Systems Genomics, The University of Melbourne, Australia; [9]Tropic IT Ltd., Central, Hong Kong; [10]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK; [11]Max F. Perutz Laboratories, 1030 Vienna, Austria; [12]Dept. of Ecogenomics and Systems Biology, University of Vienna, 1090 Vienna, Austria; [13]Department of Natural Sciences, Mid-Sweden University, 851 70 Sundsvall, Sweden; [14]Monsanto, Chesterfield, MO 63017, USA; [15]Vienna Metabolomics Center, University of Vienna, 1090 Vienna, Austria; [16]Estación Biológica de Doñana, Consejo Superior de Investigaciones Científicas, Sevilla-41092, Spain; [17]Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany; [18]Institute of Digital Agriculture, Zhejiang Academy of Agricultural Sciences, Hangzhou, Zhejiang, 310021, PR China

Current addresses:

Center for Research Informatics, The University of Chicago, Chicago, Illinois, USA (JA); Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany (XG); intomics, 2800 Lyngby, Denmark (SRH); Center for Computational and Theoretical Biology, University Würzburg, 97074 Würzburg, Germany (AK, PK); Ibis Biosciences, Carlsbad, CA 92008, USA (TM); Genetics Institute, University College London, London WC1E 6BT, UK (RM); Dept. of Chemistry and Biochemistry, UCLA, Los Angeles, CA 90095, USA (PAS); Dept. of Genetics, University of Georgia, Athens, GA 30602, USA (RJS); Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, UK (HS); Monsanto, Chesterfield, MO 63017, USA (DT); Bayer Cropscience, 9052 Gent, Belgium (XW)