Common Sequence Polymorphisms Shaping Genetic Diversity in Arabidopsis thaliana

Richard M. Clark, *et al.*

Science **317**, 338 (2007);
DOI: 10.1126/science.1138632

Common Sequence Polymorphisms Shaping Genetic Diversity in *Arabidopsis thaliana*

Richard M. Clark,[1] Gabriele Schweikert,[1,2,3]* Christopher Toomajian,[4]* Stephan Ossowski,[1]* Georg Zeller,[1,2,5]* Paul Shinn,[6] Norman Warthmann,[1] Tina T. Hu,[4] Glenn Fu,[7] David A. Hinds,[7] Huaming Chen,[6] Kelly A. Frazer,[7] Daniel H. Huson,[5] Bernhard Schölkopf,[3] Magnus Nordborg,[4] Gunnar Rätsch,[2] Joseph R. Ecker,[6,8] Detlef Weigel[1,8]†

The genomes of individuals from the same species vary in sequence as a result of different evolutionary processes. To examine the patterns of, and the forces shaping, sequence variation in *Arabidopsis thaliana*, we performed high-density array resequencing of 20 diverse strains (accessions). More than 1 million nonredundant single-nucleotide polymorphisms (SNPs) were identified at moderate false discovery rates (FDRs), and ~4% of the genome was identified as being highly dissimilar or deleted relative to the reference genome sequence. Patterns of polymorphism are highly nonrandom among gene families, with genes mediating interaction with the biotic environment having exceptional polymorphism levels. At the chromosomal scale, regional variation in polymorphism was readily apparent. A scan for recent selective sweeps revealed several candidate regions, including a notable example in which almost all variation was removed in a 500-kilobase window. Analyzing the polymorphisms we describe in larger sets of accessions will enable a detailed understanding of forces shaping population-wide sequence variation in *A. thaliana*.

Comprehensive polymorphism data are essential for the systematic identification of sequence variants affecting phenotypes (*1*). Despite progress with new technologies, direct resequencing of individual genomes is not yet cost effective for most organisms (*2*). High-density oligonucleotide arrays provide an alternative approach for polymorphism detection and have been used to identify a large fraction of the SNP variation in the human and the mouse (*3*, *4*). We applied this technology to 20 wild accessions of *A. thaliana*, for which a genome sequence from a single accession was generated in the year 2000 (*5*). The resulting polymorphism data set captures much of the common sequence variation in the worldwide *A. thaliana* population. We used this information to systematically determine the types of sequences and genes that differ between accessions and to provide a high-resolution description of the genome-wide distribution of polymorphisms in this multicellular reference organism.

Sample selection, array design, and polymorphism detection. For polymorphism discovery, we selected accessions with maximal genetic diversity (*6*, *7*). In addition, we chose several commonly used strains, such as L*er*-1 (table S1). Col-0, the reference accession, was included as a control. For 19 of the 20 accessions (*7*), 1213 fragments of ~500 base pairs (bp) in length, which were spaced throughout the genome, had previously been sampled by dideoxy sequencing; between 2266 and 3949 nucleotide substitutions per accession relative to Col-0 had been identified (*6*). This data set, called "2010," allowed us to assess the quality of our polymorphism predictions.

Whole-genome amplified DNA from each accession was hybridized to resequencing microarrays interrogating >99.99% of bases in the 119-Mb reference genome sequence (*7*). Each position was queried with forward- and reverse-strand probe quartets consisting of oligonucleotides of length 25 (fig. S1). Within a probe quartet, all four nucleotides were represented at the central position, and differences in relative intensities across probe quartets indicated potential SNPs. For tightly linked SNPs, however, all probes harbor at least one mismatch, hybridization is suppressed, and SNP detection is confounded (fig. S1).

We used two computational methods to detect SNPs at 105,920,272 positions that were not highly repetitive (*7*) (table S2 and fig. S2). In *A. thaliana*, the sequence composition (i.e., GC content) and low polymorphism levels typical for coding sequences are favorable for hybridization-based SNP detection (*7*). Accordingly, recovery of SNPs with a model-based (MB) algorithm (*3*, *4*) was higher for coding than for noncoding regions (36 versus 15%) at a corresponding FDR that was only one-third as high (Fig. 1A). An average of 96,814 SNPs were identified per accession by the MB method, for a total of 456,956 nonredundant SNPs (Fig. 1B and table S3).

We also developed a machine learning (ML) method with support vector machines (*8*, *9*) for SNP identification (*7*) (figs. S3 to S8). The training step exploited the 2010 data, and as input we used information for all oligonucleotide probes corresponding to positions within a 9-bp window centered on candidate polymorphisms (*7*). In addition to hybridization data, we included as inputs sequence characteristics and genome-wide repetitiveness of probes (tables S5 and S6). The ML method assigns a probability to each prediction, and we generated 440,657 to 1,074,055 nonredundant SNP predictions over a corresponding range of FDRs from 2 to 10% (*7*, *10*). Performance of the ML method was inferior to the MB method for coding sequences but superior for noncoding sequences (Fig. 1, A and B, and table S3).

When the FDR for the ML method was at 2%, the FDR and recovery for the ML and MB methods were similar; however, the two methods were complementary, with 60% of predictions made with only one of the methods (Fig. 1C). This resulted, in part, from differing performance in polymorphic regions (Fig. 1D). Recall for SNPs more than ~30 bp from another SNP or insertion or deletion (indel) was higher for the MB method, whereas recall for SNPs separated by 7 to 30 bp from a nearby polymorphism was about two times as high for the ML method. For very closely linked SNPs (<7 bp), recovery was low with both methods (~3%). FDRs for both methods peaked in regions of low hybridization quality (Fig. 1E), an effect of sequence divergence but also of other factors (*7*).

For subsequent analyses, we combined all MB predictions with ML predictions supported at a 2% FDR. The resulting data set, "MBML2," consisted of 648,570 nonredundant SNPs (*7*, *10*) (Fig. 2), an average of one polymorphic site per 166 nonrepetitive positions in the genome. Within MBML2, SNPs supported by both methods have a very low FDR of ~0.2%, whereas SNPs supported by only one method have correspondingly higher FDRs (Table 1 and table S3). A caveat for our error estimates is that 2010 data, which we used for specificity and sensitivity assessment of the two prediction methods, are underrepresented for noncoding sequences, repeats, and sequences not similar to the reference (*6*, *7*).

Apart from SNPs, deletions or sequences highly dissimilar to the reference are detectable on high-density arrays as regions of reduced hybridization (*11*) (fig. S1). We developed a heuristic algorithm to identify tracts of reduced hybridization extending over more than ~200 bp (*7*) (figs. S9 and S10). The median length of 13,470 polymorphic region predictions (PRPs) generated across all accessions with this algorithm was 589 bp; the longest was 41.2 kb (*10*). In the 2010 data set, which was ascertained by polymerase chain reaction (PCR), missing data correspond in part to highly polymorphic or deleted regions. Consistent with high specificity for PRPs, a 162-fold overrepresentation was observed between PRPs and absent data in 2010. We also attempted validation of 382 PRPs by PCR and sequencing, ob-

[1]Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany. [2]Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tübingen, Germany. [3]Department of Empirical Inference, Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany. [4]Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA. [5]Center for Bioinformatics Tübingen, Tübingen University, 72076 Tübingen, Germany. [6]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA. [7]Perlegen Sciences, 2021 Stierlin Court, Mountain View, CA 94043, USA. [8]Plant Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: weigel@weigelworld.org

tained complete or partial sequence data for 171 products, and identified 124 deletions ranging from 50 bp to more than 10 kb. In all other cases, PRPs corresponded to clusters of SNPs or small indels (table S11). Many deletions or clusters of polymorphisms extended beyond PRP boundaries, potentially contributing to the high failure rate for validation attempts (~55%). Where sequence data were available, 98.6% of bases in PRPs were either deleted or within 6 bp of a SNP or indel polymorphism (*7*). Nearly 4.1% of the reference genome sequence was included in PRPs, with transposon and pseudogene sequences overrepresented 3.5-fold (Fig. 2).

To complement polymorphism predictions, we developed a base-calling algorithm to identify positions identical to the reference at low FDRs (*7*). Between 80.3 and 92.3% of coding positions and between 39.7 and 61.2% of intergenic positions were predicted to be the same as the reference in the different accessions (table S8). We combined these reference base calls with MBML2 to generate pseudochromosome sequences for each of the 20 accessions (*10*).

Effects of polymorphisms on genes. To characterize genome evolution in *A. thaliana*, we assessed effects of the nonredundant MBML2 SNPs on the 26,541 annotated protein-coding genes (*12*). In addition to SNPs resulting in 109,979 amino acid changes, we identified many SNPs with large effects on gene integrity. In this class, 1227 introduce premature stop codons, 156 alter initiation methionine residues, and 435 lead to nonfunctional splice donor or acceptor sites (*10*) (table S9). Also, 198 SNPs remove annotated stop codons, resulting in longer open reading frames. Given that large-effect SNPs are expected to be uncommon in the genome relative to all SNPs, FDRs for this SNP subclass might differ from that for MBML2 as a whole. To rule out the possibility that large-effect SNPs resulted predominantly from false SNP calls, we assayed 701 of these predictions directly (table S9). Dideoxy-sequencing validated 650 SNPs, including 413 resulting in premature stop codons (table S10). At 7.3%, the FDR for large-effect SNPs is moderately higher than for an average SNP in MBML2 (*7*). In total, 1614 genes harbor at least one large-effect SNP. In addition, the coding regions of 1191 genes are at least partially included in PRPs; that is, they are highly polymorphic or deleted. The overlap between the two classes is greater than expected by chance ($\chi^2 = 186.1$, df = 2, $P < 10^{-20}$). Together, large-effect SNPs and PRPs, hereafter referred to simply as "major-effect changes," affected 2495, or 9.4%, of *A. thaliana* protein-coding genes.

The number of genes harboring major-effect changes varies significantly according to annotation support ($\chi^2 = 239.2$, df = 2, $P < 10^{-20}$), duplication status ($\chi^2 = 256.4$, df = 2, $P < 10^{-20}$), and gene family ($\chi^2 = 311.6$, df = 12, $P < 10^{-20}$) (Fig. 3A). Correction for gene size and repetitive content does not appreciably change the observed patterns (fig. S11). By annotation, genes known to be expressed but lacking functional support or high homology ("Expressed unknown"), as well as genes without expression support ("Not expressed"), are overrepresented. In addition, of 836 *A. thaliana* genes that either lack or have only moderate similarity

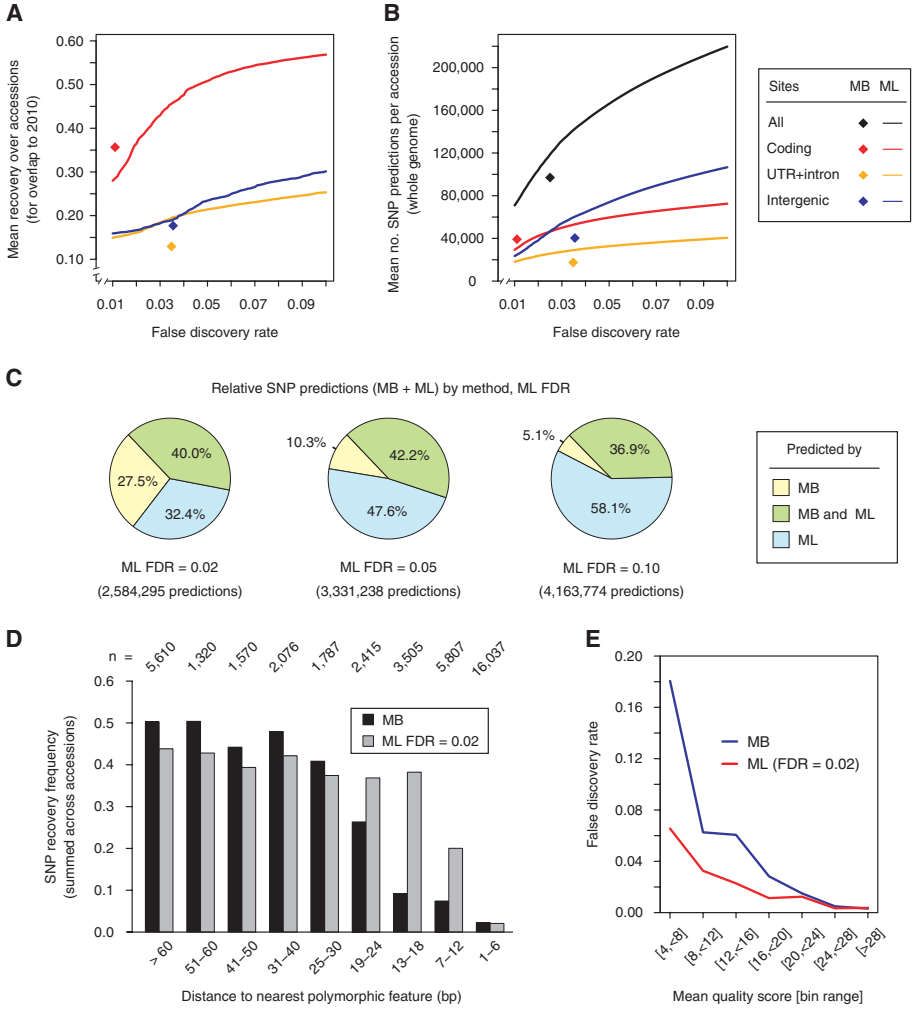
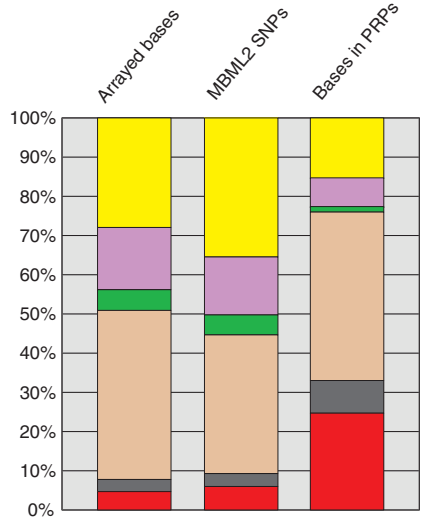


Fig. 1. Comparison of SNP detection methods. (**A**) FDR-dependent recovery of 48,700 known SNPs in 2010 fragments by the MB and ML methods. Because of small sample size for untranslated region (UTR) SNPs in 2010, this group was combined with intron sequences. (**B**) FDR-dependent recovery across the entire genome by both methods with precision estimates from 2010. The FDR for all SNPs is also given with a correction for both methods to account for the different sequence composition of 2010 and the whole genome. (**C**) Overlap between genome-wide MB and ML calls across all accessions. (**D**) Recovery frequency for SNPs as a function of distance to the nearest polymorphic feature. Analysis was based on 2010 sites with sufficient flanking information to assign bin membership. Sample sizes per bin are shown at the top. (**E**) FDRs for MB and ML predictions as a function of the mean quality score for the forward- and reverse-strand probe quartets.



Sequence type	SNP no.	bp in PRPs
Coding	230,186	747,541
Intron	96,069	368,916
UTR	32,771	66,254
Intergenic	229,335	2,076,046
Pseudogene	21,552	401,282
Transposon	38,657	1,196,637

Fig. 2. Distribution of SNPs in MBML2 and positions included in PRPs compared with sequences tiled on arrays.

to genes in *Populus trichocarpa* (*7*), the closest sequenced genome to *A. thaliana* (*13*), 26.0% harbor major-effect changes, compared with 8.9% of all other genes. Poor gene annotation likely contributes to this effect, but rapid gene evolution may also play a part.

Table 1. SNPs identified per accession in MBML2 with FDR and recovery assessed against 2010.

	Mean no. SNPs per accession by method [Mean FDR (%): Mean recovery (%)]			
SNP type	Total	MB only	MB ∩ ML	ML only
Coding	53,700 [2.0: 48.0]	11,379 [3.2: 11.3]	27,833 [0.1: 24.6]	14,488 [4.8: 12.1]
Intron+UTR	29,395 [3.1: 20.5]	5,762 [9.6: 4.3]	11,652 [0.4: 8.8]	11,981 [2.6: 7.5]
Intergenic	60,478 [3.5: 24.4]	22,395 [7.3: 7.7]	17,976 [0.3: 10.2]	20,107 [3.6: 6.5]
All	143,572 [2.8: 27.7]	39,536 [6.5: 7.0]	57,461 [0.2: 13.7]	46,575 [3.7: 7.8]

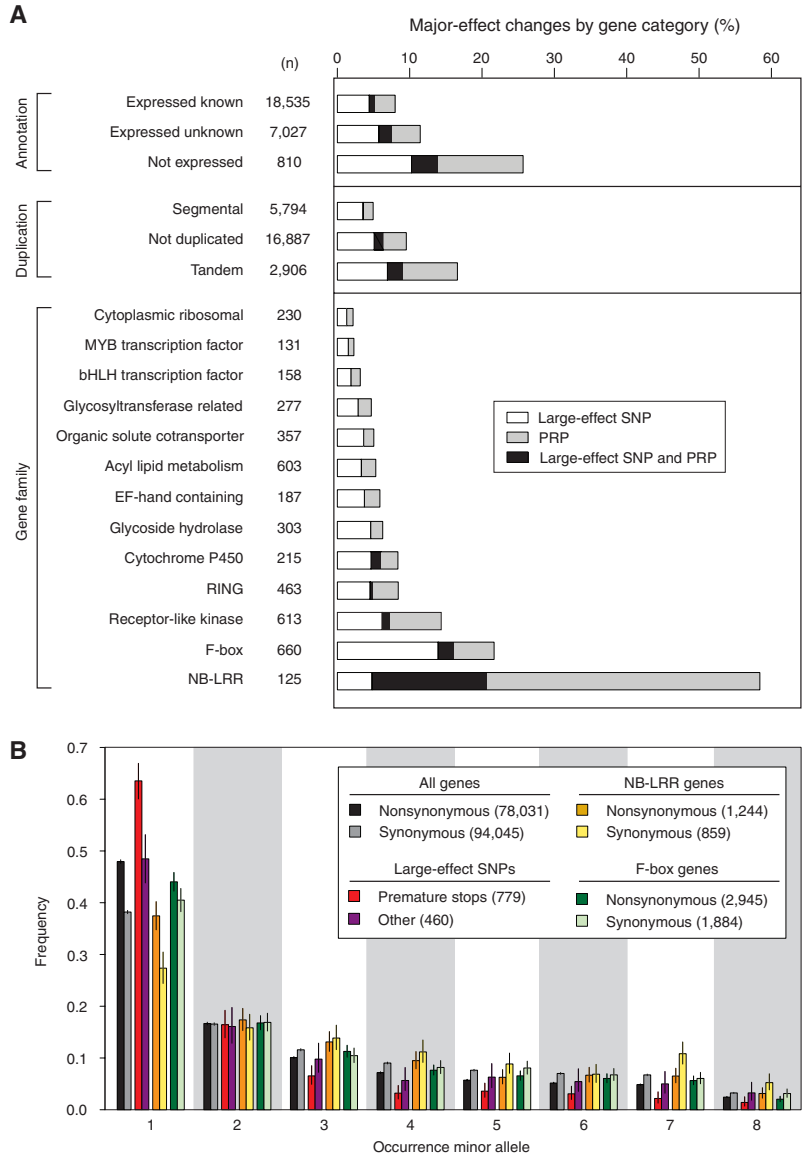


Fig. 3. Distribution of major-effect changes and allele frequencies. (**A**) Fraction of genes affected by large-effect SNPs or PRPs by gene category. "Large-effect SNP and PRP" (black) denotes a gene harboring both types of polymorphism, either within the same accession or in different accessions. Genes that were entirely masked as repetitive, and for which no SNPs could be predicted, were excluded from analysis. RING, Really Interesting New Gene. (**B**) Minor allele frequency by SNP type and gene family. Only positions with complete data for at least 16 of the 20 accessions were assessed. The number of polymorphic positions included in the analysis is shown in the inset. For large-effect SNPs, "Other" includes nonfunctional splice site changes, substitutions in initiation methionine codons, and substitutions that remove termination codons. Error bars denote 95% confidence intervals for binomial expectations.

Consistent with relaxed purifying selection following recent gene duplication (*14*), tandem duplicates are 3.4- and 1.7-fold overrepresented for major-effect changes relative to segmentally duplicated and nonduplicated genes, respectively (Fig. 3A). Segmentally duplicated genes in *A. thaliana* resulted from ancient genome-wide duplications (*5*, *15*); these genes harbor relatively few major-effect changes, which is consistent with earlier work suggesting that duplicates persisting over long evolutionary time frames are under strong purifying selection (*13*, *14*).

Analysis of individual gene families provided additional insights. Families involved in basic biological processes (such as ribosomal function), as well as families involved in transcriptional regulation [such as MYB and basic helix-loop-helix (bHLH) transcription factors], harbor relatively few major-effect changes (Fig. 3A). In contrast, nearly 60% of nucleotide-binding leucine-rich repeat (NB-LRR) genes (*7*) and 15% of receptor-like kinase (RLK) genes harbor at least one major-effect change. The only function assigned to members of the NB-LRR gene family is in strain-specific resistance to pathogens (*16*), and receptors of this class can be exceedingly variable, as presence and absence polymorphisms are common in *A. thaliana* and other plants (*17*–*19*). Our data indicate that this extends to the majority of NB-LRR genes in the *A. thaliana* genome. Although they have diverse functions (*20*), RLK genes have also been implicated in race-specific pathogen defense (*21*). Thus, the finding that RLK genes are overrepresented for major-effect changes raises the possibility that this is, similar to NB-LRR genes, a consequence of fitness trade-offs between pathogen defense and growth (*22*).

We found major-effect changes in 143 members of the F-box superfamily, which comprises more than 660 genes in *A. thaliana* (*23*) (Fig. 3A). This finding, in combination with other data (*13*, *24*), shows that F-box genes have undergone rapid birth and death in the *A. thaliana* genome. Although F-box genes have been proposed to evolve quickly in response to pathogen pressure (*24*), experimental support for this hypothesis is lacking. The polymorphisms we describe provide a resource for ascribing biological roles to members of this large, yet not well-characterized, gene family.

Signatures of selection by SNP type and gene family. To assess the extent to which the variation we observed has been shaped by selection, we examined allele frequency distributions for different classes of polymorphisms. Consistent with general expectations for selective constraints in coding sequences, there is a skew toward low-frequency variants at nonsynonymous relative to synonymous sites (*6*) (Fig. 3B). This skew is most notable for SNPs that introduce premature stop codons and less extreme for other large-effect SNPs. The tendency of SNPs causing premature stops to be rare suggests that, at least under natural settings, these changes are often associated with fitness costs.

Although allele frequency distributions across gene families are broadly similar, NB-LRR genes are an exception (Fig. 3B and fig. S12). Here,

both nonsynonymous and synonymous variants are strongly skewed toward high frequency compared with the genome average (Fig. 3B). This shift is a hallmark of some type of balancing selection (perhaps in the form of regional adaptation), and agrees with earlier work, on the basis of fewer family members, that suggested this mode of selection to be not uncommon for NB-LRR genes in *A. thaliana* (*17*, *19*, *22*). An additional prediction of balancing selection is a higher-than-average level of polymorphism because of the maintenance of relatively ancient, highly diverged alleles. Consistent with this expectation, more than 50% of NB-LRR genes are at least partially included in PRPs (Fig. 3A), many of which correspond to highly dissimilar sequences. Although less extreme, a similar allele frequency skew and high number of PRPs were observed for RLK genes (Fig. 3A and fig. S12). Although F-box genes harbor the second-highest occurrence of major-effect changes, allele frequency distributions are similar to the average (Fig. 3B).

Genome-wide patterns of polymorphism. Turning to broader patterns of variation, we found a markedly nonrandom distribution of polymorphism levels across the genome (Fig. 4). Regions of high polymorphism extend from the centromeres to beyond the pericentromeric regions. Similarly, clusters of NB-LRR genes (*7*, *25*) are associated with high levels of polymorphism (e.g., between 21 and 25 Mb on chromosome 1).

It is difficult to determine the reasons for these patterns. Given that they are evident for both synonymous and intergenic polymorphism, direct selection on the polymorphic sites seems unlikely, although selection on linked sites may either increase or decrease variation (*6*). Mutation rates may also differ between chromosomal regions because of differences in base composition. Finally, there are almost certainly biases in the array-based resequencing—e.g., due to regional differences in repeat content (*7*).

We believe that all three explanations contribute to the observed patterns. In a multiple regression, the estimated polymorphism levels are significantly correlated with several variables, including repeat density, GC content, fraction of missing data, distance from the centromere, and density of NB-LRR genes (table S12). The patterns are, however, not simply an artifact of the resequencing technology; they are also evident in resequencing data obtained with other techniques (*6*, *26*) (fig. S13). For the regions around NB-LRR gene clusters, polymorphism was elevated even when the NB-LRR genes themselves were excluded (fig. S14), and polymorphism was also elevated in intergenic DNA. Pervasive balancing selection acting on these genes, as suggested by the results in Fig. 3B and other studies (*17*), is a likely explanation. Balancing selection can increase coalescence times for regions linked to selectively maintained polymorphisms (*27*), a phenomenon that should be more easily detected in selfing organisms (*28*) and that has been reported for *A. thaliana* (*22*, *29*, *30*). Clusters of tightly linked genes subject to balancing selection, such as NB-LRR genes (*25*) (Fig. 4), may give rise to regions of high polymorphism similar to what has been observed for the vertebrate major histocompatibility complex genes (*31*, *32*). The forthcoming *A. lyrata* genome sequence (*33*) will be instrumental in analyzing these data further, because it will allow divergence to be estimated between these two closely related species. This will be essential for determining the relative importance of selection versus mutation-rate variation.

In contrast, regions of low polymorphism might reflect recent positive selection, or "selective sweeps" (*34*, *35*) characterized by extensive haplotype sharing. A study with 2010 data found strong evidence for two separate partial sweeps involving inactivation of *FRI*, a major determinant of flowering time in natural populations of *A. thaliana* (*36*). The

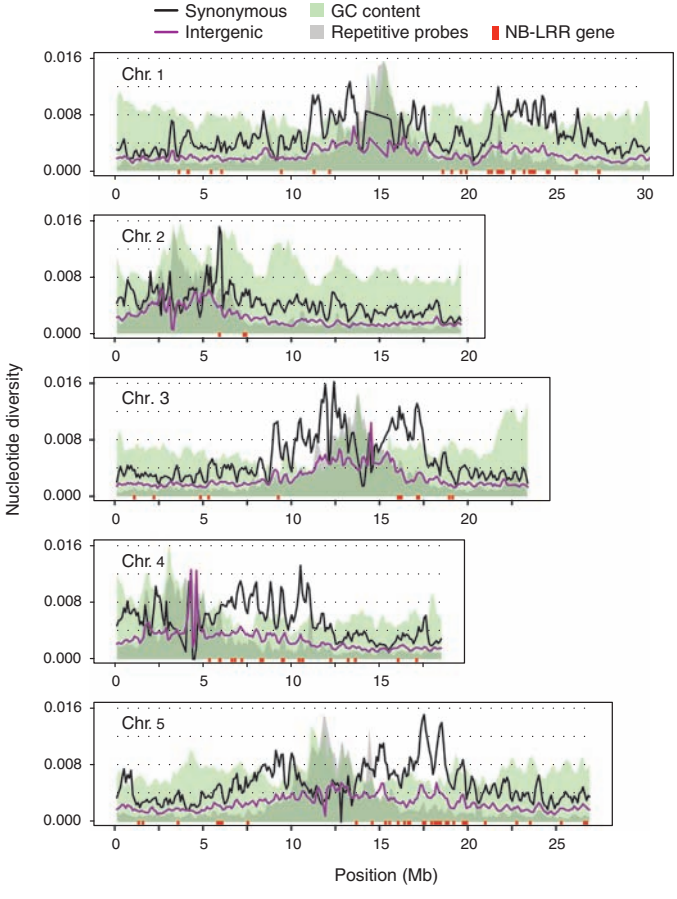


Fig. 4. Genome-wide pattern of nucleotide diversity. Average pairwise nucleotide diversity is plotted for both fourfold degenerate synonymous and intergenic sites along each chromosome with sliding windows of 250 kb (counted from all sites) with an offset of 100 kb (*7*). GC content in each window was calculated from sites called in the Col-0 sample and has been rescaled so 35% is at the bottom of each plot and 47.5% is at the top. The broad peaks of repetitive probe density on each chromosome correspond to the centromeric and pericentromeric regions. Repeat content has been rescaled so 100% is at the top of each plot and 0% is at the bottom. Levels of polymorphism for both fourfold degenerate and intergenic sites are significantly negatively correlated with the distance to the centromere and positively correlated with the number of NB-LRR genes in nonoverlapping 50-kb windows (table S12). Polymorphism is reduced at intergenic relative to synonymous sites, which is partly due to lower recovery of SNPs in intergenic regions (e.g., Fig. 1A).

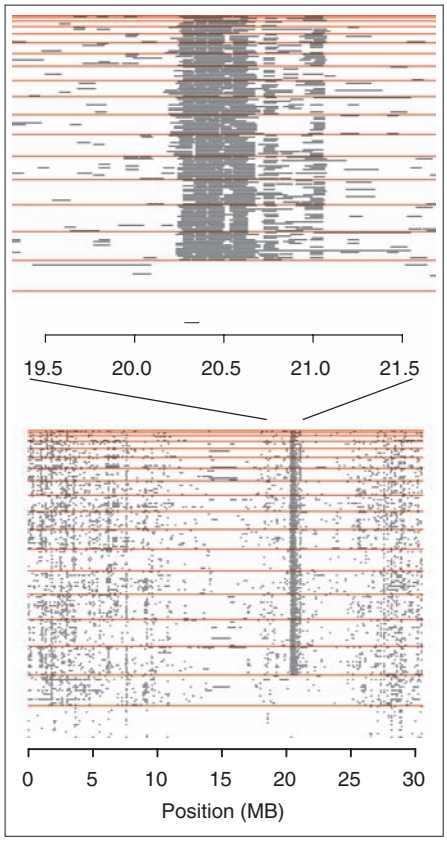


Fig. 5. Regions of extensive pairwise haplotype sharing along chromosome 1. Accession pairs are sorted along the *y* axis. Horizontal red lines demarcate comparisons using one accession. Each possible pairwise comparison is shown only once. Black lines indicate regions of very high similarity between a pair of accessions. The region between 20 and 21 Mb exhibits extensive haplotype sharing over nearly 500 kb in all but two accessions, Cvi-0 and Lov-5, which are shown at the bottom.

current data set confirmed extensive haplotype sharing of up to 600 kb around *FRI* (fig. S15), as well as haplotype sharing around other low-frequency candidate alleles (*36*) (fig. S16).

We looked for evidence of additional sweeps in the form of extensive haplotype sharing across at least 50 kb (Fig. 5 and figs. S17 to S19). Because of its composition and size, our sample is only suited for discovering species-wide sweeps. We did not find evidence of a recent sweep affecting all accessions. However, on chromosome 1 all but two accessions were nearly identical for approximately 500 kb (Fig. 5). The two unaffected accessions, Cvi-0 and Lov-5, are from the periphery of the *A. thaliana* range and may have escaped the sweep because of different selective environments or geographic isolation. The region of most extreme haplotype sharing extends from 20.34 to 20.49 Mb and contains 50 annotated genes (table S13). There are several additional candidates for sweeps affecting a smaller number of accessions (figs. S17 to S20). With the SNPs identified in this project and the ability to determine their frequencies in hundreds to thousands of accessions (*37*), the goal of understanding the forces shaping diversity at global, regional, and local scales will soon be within reach.

Conclusions. We used array-based methods to generate a comprehensive polymorphism resource for *A. thaliana*. Our SNP data set is highly applicable for linkage disequilibrium mapping studies. In addition, we identified hundreds of thousands of polymorphisms in both coding and noncoding regions, providing an important resource for both evolutionary genetic and functional studies. Recently, studies in plants with large, repetitive genomes, like maize (genome size ~2.5 Gb), have shown that as much as 50% of sequences can differ between strains (*38*). In contrast to these plants, *A. thaliana* has a compact genome consisting largely of unique sequences. Nevertheless, our data highlight that even for species with streamlined genomes, individuals can differ substantially in genic content.

Mutations identified in laboratory phenotypic screens typically have marked phenotypic effects that are likely detrimental in the wild. The genes segregating for major-effect changes in our population have few known mutant phenotypes (tables S10 and S11), but nonetheless, allele frequency patterns suggest functional constraints under natural conditions. Variation in copy number for genic sequences may explain this observation; in a given accession, higher constraint may be observed if a paralog is absent. Nevertheless, as highlighted by the current study, many genes harboring major-effect changes in wild populations are likely to mediate interactions with the environment. Ultimately, experiments under more natural conditions will be required to fully appreciate the functional relevance of such sequence variation.

References and Notes

1. The International HapMap Consortium, *Nature* **437**, 1299 (2005).
2. J. Kling, *Nat. Biotechnol.* **23**, 1333 (2005).
3. D. A. Hinds *et al.*, *Science* **307**, 1072 (2005).
4. N. Patil *et al.*, *Science* **294**, 1719 (2001).
5. The Arabidopsis Genome Initiative, *Nature* **408**, 796 (2000).
6. M. Nordborg *et al.*, *PLoS Biol.* **3**, e196 (2005).
7. Materials and methods are available as supporting material on *Science* Online.
8. B. Schölkopf, A. Smola, *Learning with Kernels* (MIT Press, Cambridge, MA, 2002).
9. V. N. Vapnik, *Estimation of Dependences Based on Emperical Data* (Springer, New York, 1982; reprinted by Springer, New York, 2006).
10. SNP and PRP data sets along with effects on genes and pseudochromosome sequences are hosted at The *Arabidopsis* Information Resource (TAIR) (www.arabidopsis.org/).
11. D. A. Hinds, A. P. Kloek, M. Jen, X. Chen, K. A. Frazer, *Nat. Genet.* **38**, 82 (2006).
12. TAIR annotation Version 6 (www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=3702).
13. G. A. Tuskan *et al.*, *Science* **313**, 1596 (2006).
14. M. Lynch, J. S. Conery, *Science* **290**, 1151 (2000).
15. G. Blanc, K. Hokamp, K. H. Wolfe, *Genome Res.* **13**, 137 (2003).
16. J. D. Jones, J. L. Dangl, *Nature* **444**, 323 (2006).
17. E. G. Bakker, C. Toomajian, M. Kreitman, J. Bergelson, *Plant Cell* **18**, 1803 (2006).
18. M. R. Grant *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 15843 (1998).
19. J. Shen, H. Araki, L. Chen, J. Q. Chen, D. Tian, *Genetics* **172**, 1243 (2006).
20. S. H. Shiu, A. B. Bleecker, *Sci. STKE* **2001**, RE22 (2001).
21. W. Y. Song *et al.*, *Science* **270**, 1804 (1995).
22. E. A. Stahl, G. Dwyer, R. Mauricio, M. Kreitman, J. Bergelson, *Nature* **400**, 667 (1999).
23. E. Lechner, P. Achard, A. Vansiri, T. Potuschak, P. Genschik, *Curr. Opin. Plant Biol.* **9**, 631 (2006).
24. J. H. Thomas, *Genome Res.* **16**, 1017 (2006).
25. B. C. Meyers, A. Kozik, A. Griego, H. Kuang, R. W. Michelmore, *Plant Cell* **15**, 809 (2003).
26. K. J. Schmid, S. Ramos-Onsins, H. Ringys-Beckstein, B. Weisshaar, T. Mitchell-Olds, *Genetics* **169**, 1601 (2005).
27. R. R. Hudson, N. L. Kaplan, *Genetics* **120**, 831 (1988).
28. M. Nordborg, *Genetics* **146**, 1501 (1997).
29. J. Kroymann, T. Mitchell-Olds, *Nature* **435**, 95 (2005).
30. D. Tian, H. Araki, E. Stahl, J. Bergelson, M. Kreitman, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 11525 (2002).
31. A. L. Hughes, M. Nei, *Nature* **335**, 167 (1988).
32. M. Nordborg, H. Innan, *Genetics* **163**, 1201 (2003).
33. Eightfold coverage for *A. lyrata* and *Capsella rubella* is being generated by the Joint Genome Institute (www.jgi.doe.gov/).
34. N. L. Kaplan, R. R. Hudson, C. H. Langley, *Genetics* **123**, 887 (1989).
35. J. M. Smith, J. Haigh, *Genet. Res.* **23**, 23 (1974).
36. C. Toomajian *et al.*, *PLoS Biol.* **4**, e137 (2006).
37. A. C. Syvanen, *Nat. Genet.* **37** (Suppl.), S5 (2005).
38. M. Morgante, *Curr. Opin. Biotechnol.* **17**, 168 (2006).

39. We thank G. Nielson and H. Huang for bioinformatics support; R. Gupta and M. Morenzoni for information management; T. Altman, J. Borevitz, C. Dean, and C. Shindo for seed stocks; J. Gagne, D. Gingerich, R. Vierstra, L. Sterck, and Y. van de Peer for providing gene family or homology information; and K. Schneeberger for helpful discussions. Supported by Innovation Funds of the Max Planck Society, NIH (HG002790 to M. Waterman, and GM62932 to J. Chory and D.W.), NSF (DEB-0115062 to M.N., and DBI-0520253 to J.R.E.), an NIH National Research Service Award fellowship to C.T., and core funding from the Max Planck Society. D.W. is a director of the Max Planck Institute. Sequence data have been deposited in GenBank (accession codes EI100660 to EI102044).

Supporting Online Material

www.sciencemag.org/cgi/content/full/317/5836/338/DC1
Materials and Methods
Figs. S1 to S20
Tables S1 to S15
References and Notes

11 December 2006; accepted 7 June 2007
10.1126/science.1138632

REPORTS

Imaging the Surface of Altair

John D. Monnier,[1]* M. Zhao,[1] E. Pedretti,[2] N. Thureau,[3] M. Ireland,[4] P. Muirhead,[5] J.-P. Berger,[6] R. Millan-Gabet,[7] G. Van Belle,[7] T. ten Brummelaar,[8] H. McAlister,[8] S. Ridgway,[9] N. Turner,[8] L. Sturmann,[8] J. Sturmann,[8] D. Berger[1]

Spatially resolving the surfaces of nearby stars promises to advance our knowledge of stellar physics. Using optical long-baseline interferometry, we constructed a near-infrared image of the rapidly rotating hot star Altair with a resolution of <1 milliarcsecond. The image clearly reveals the strong effect of gravity darkening on the highly distorted stellar photosphere. Standard models for a uniformly rotating star cannot explain our findings, which appear to result from differential rotation, alternative gravity-darkening laws, or both.

Whereas solar astronomers can take advantage of high-resolution, multi-wavelength, real-time imaging of the Sun's surface, stellar astronomers know most stars—whether located parsecs or kiloparsecs away—as simple points of light. To discover and understand the processes around stars unlike the Sun, we must rely on stellar spectra averaged over the entire photosphere. Despite their enormous value, spectra alone have been inadequate to resolve central questions in stellar astronomy, such as the role of angular momentum in stellar evolution (*1*), the production and maintenance of magnetic fields (*2*), the launching of massive stellar winds (*3*), and the interactions between very close binary companions (*4*).

Fortunately, solar astronomers no longer hold a monopoly on stellar imaging. Long-baseline visible and infrared interferometers have enabled the cataloging of photospheric diameters of hundreds of stars and high-precision dynamical masses for dozens of binaries, offering exacting constraints for theories of stellar evolution and stellar atmospheres (*5*). This work requires an angular resolution of ~1 milliarcsecond (mas) (1 part in 2×10^8, or 5 nanoradians) for resolving even nearby stars, which is more than an order of