Genome-wide patterns of single-feature polymorphism in *Arabidopsis thaliana*

Justin O. Borevitz*†‡§, Samuel P. Hazen¶, Todd P. Michael*, Geoffrey P. Morris‡, Ivan R. Baxter‖, Tina T. Hu**, Huaming Chen†, Jonathan D. Werner*, Magnus Nordborg**, David E. Salt‖, Steve A. Kay¶, Joanne Chory*††, Detlef Weigel*‡‡, Jonathan D. G. Jones§§, and Joseph R. Ecker*†§

*Plant Biology Laboratory, †Genomic Analysis Laboratory, and ††Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037; ‡Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637; ¶Department of Biochemistry, The Scripps Research Institute, La Jolla, CA 92037; ‖Department of Horticulture, Purdue University, West Lafayette, IN 47907; **Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089; ‡‡Department of Molecular Biology, Max Planck Institute for Developmental Biology, D-72076 Tübingen, Germany; and §§Sainsbury Laboratory, John Innes Centre, Colney Lane, Norwich NR4 7UH, United Kingdom

Contributed by Joseph R. Ecker, June 7, 2007 (sent for review April 23, 2007)

We used hybridization to the ATH1 gene expression array to interrogate genomic DNA diversity in 23 wild strains (accessions) of *Arabidopsis thaliana* (arabidopsis), in comparison with the reference strain Columbia (Col). At <1% false discovery rate, we detected 77,420 single-feature polymorphisms (SFPs) with distinct patterns of variation across the genome. Total and pair-wise diversity was higher near the centromeres and the heterochromatic knob region, but overall diversity was positively correlated with recombination rate ($R^2 = 3.1\%$). The difference between total and pair-wise SFP diversity is a relative measure contrasting diversifying or frequency-dependent selection, similar to Tajima's D, and can be calibrated by the empirical genome-wide distribution. Each unique locus, centered on a gene, has a diversity and selection score that suggest a relative role in past evolutionary processes. Homologs of disease resistance (*R*) genes include members with especially high levels of diversity often showing frequency-dependent selection and occasionally evidence of a past selective sweep. Receptor-like and S-locus proteins also contained members with elevated levels of diversity and signatures of selection, whereas other gene families, bHLH, F-box, and RING finger proteins, showed more typical levels of diversity. SFPs identified with the gene expression array also provide an empirical hybridization polymorphism background for studies of gene expression polymorphism and are available through the genome browser http://signal.salk.edu/cgi-bin/AtSFP.

evolutionary genomics | gene array | nucleotide diversity

DNA sequence polymorphism studies at the genic level have been conducted for several years, but extensive studies at the whole-genome level are just beginning. These kinds of studies, which will ultimately involve the complete sequencing of multiple individuals of the same species and thus an absolute understanding of genome variation, make it possible to contrast the pattern of variation at particular loci with the genomic background and test for evidence of selection in a manner that is robust to confounding demographic factors such as population structure or bottlenecks (1). Early molecular population genetic studies in *Arabidopsis* revealed a pattern of polymorphism that was highly variable among loci, as well as an excess of rare alleles compared with neutral expectations (2). Recombination among accessions was evident from the different gene genealogies at different loci. Genome-wide nuclear markers revealed some population structure and isolation by distance underlying an overall star-like phylogeny (3–5). As marker densities increased, it became apparent that linkage disequilibrium was highly variable across the genome ranging from 25 to 200 kb, with patterns of variation often suggestive of past or ongoing selection (6, 7). With greater density of genotyping, naive scans at the whole-genome level may allow for detection of selective events as well as association mapping (8). Accordingly, high-density oligonucleotide arrays with 25-mer features, initially designed to query expression levels of transcripts, provide a straightforward tool to query sequence polymorphisms among strains (9).

These so-called single-feature polymorphisms (SFPs) detect sequence polymorphisms in or near the 25-mer array feature rather than precise nucleotide polymorphisms and have been used to identify a large number of polymorphisms in several organisms, including *Saccharomyces cerevisiae* (10), *Arabidopsis thaliana* (11), *Oryza sativa* (domesticated rice) (12), *Anopheles gambiae* (African malaria mosquito) (13), *Plasmodium falciparum* (protozoan parasite) (14), and *Hordeum vulgare* (domesticated barley) (15, 16). SFPs segregate with expected frequencies and are generally biallelic when diversity is low (<1% sequence variation), allowing relatively rapid mapping of qualitative (11, 17–19) and quantitative traits (13, 20–22). In addition to their application as molecular markers for genetic linkage studies, SFPs are complementary to single-locus approaches previously used to estimate population genetic parameters (23) in *Arabidopsis* (6) and *Drosophila melanogaster* (24). An important experimental design consideration is the tradeoff between quantity and quality. Population studies based on SNPs generally benefit from high-quality data but typically sparse genome coverage, whereas SFPs tend to be of lower quality yet exceptionally abundant. A power analysis evaluating decay of linkage disequilibrium across various simulated study design parameters indicated that the abundance of lower-quality SFPs compensates for a lack of SNP density (25). In addition, the probability of detecting potential causative polymorphism is considerably greater (17).

We investigated genome-wide polymorphism in 23 *Arabidopsis* accessions through SFP genotyping. Centromeres exhibited high levels of variation; however, overall diversity was positively correlated with recombination rate. The genome-wide distribution of diversity highlights that regions have unusually high or low levels and are likely to be under selection. Plant disease resistance-like (*R*) genes, receptor-like protein (RLP) genes, and S-locus protein genes families showed increased levels of diversity relative to families with similar numbers of genes. Furthermore, individual gene family members were identified as clear outliers, suggesting their evolutionary importance.

Author contributions: J.O.B. and J.R.E. designed research; J.O.B., S.P.H., T.P.M., I.R.B., J.D.W., and D.W. performed research; J.O.B., S.P.H., T.P.M., G.P.M., I.R.B., T.T.H., H.C., M.N., D.E.S., S.A.K., J.C., J.D.G.J., and J.R.E. contributed new reagents/analytic tools; J.O.B. and G.P.M. analyzed data; and J.O.B., S.P.H., T.P.M., D.W., and J.D.G.J. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Abbreviations: SFP, single-feature polymorphism; RLP, receptor-like protein; FDR, false discovery rate; *R* genes, disease-resistance genes.

§To whom correspondence may be addressed. E-mail: borevitz@uchicago.edu or ecker@salk.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0705323104/DC1.

EVOLUTION

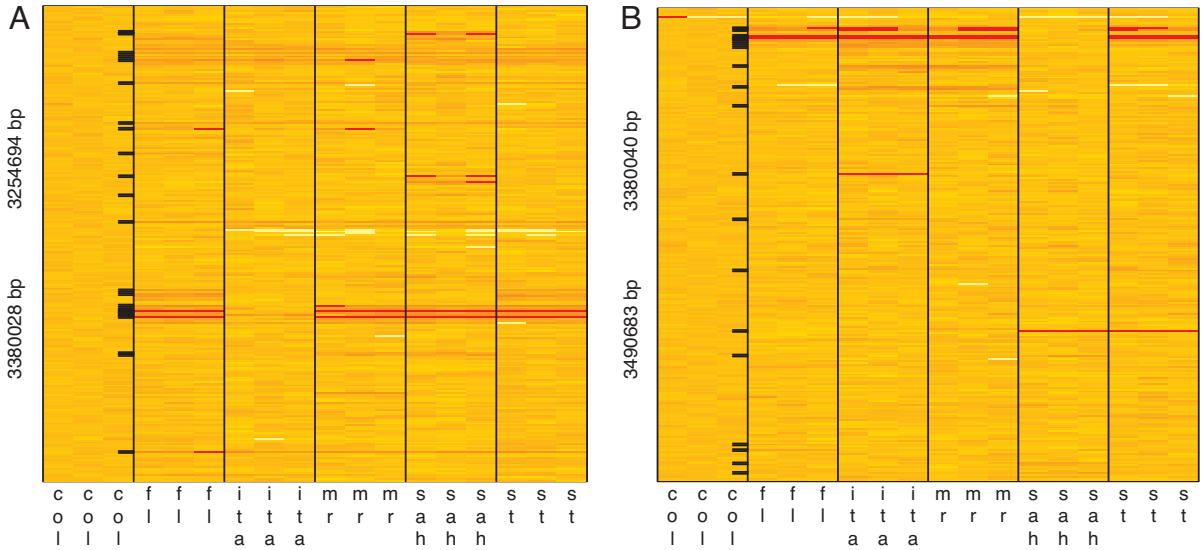

Fig. 1. SFP haplotype structure across *Arabidopsis* accessions. Feature intensities are shown as a heat map across three replicate columns of six accessions from experiment 3. Lower and higher relative SFP hybridization intensities are in red and white, respectively, as compared with the reference Col. Those in orange do not show significant variation but are included to show genotyping density. Rows correspond to 250 consecutive 25-mer features across adjacent genes and thus are not equally spaced. Black tick marks in Col show significant SFPs from the D stat threshold. (*A*) The haplotype patterns are clearly seen where Ita is similar to Col, whereas Fl, Mr, and St are quite similar to each other, and Sah has a third pattern. (*B*) The patterns change along the chromosome due to ancestral recombination events moving a Col haplotype onto Sah and a new haplotype onto Ita. Heat maps of the entire genome are available for each of five experiments (http://naturalvariation.org/accessionSFP/supplement/HapMapImages).

Results and Discussion

Genome-Wide Patterns of Polymorphism. We measured SFPs among 23 *Arabidopsis* accessions using a standard gene expression array, ATH1, with 202,878 unique 25-bp oligonucleotide features (26). Three replicates of each accession were compared against replicates of the reference strain, Col (Col-0 or Col-*gl1*). After spatial correction (11) and quantile normalization (27), modified Student's *t* tests were applied to identify SFPs with significantly lower hybridization values than Col. Instances of hybridization intensity greater than Col were assumed to be duplicated loci with an unknown physical location within the genome and were therefore removed from further consideration. At an experiment-wise <0.1% false discovery rate (FDR), 77,420 SFPs were identified including 46,911 nonsingletons. The matrix of 23 accessions by 202,878 feature positions contains 7% called SFPs, 70% called non-SFPs, and 23% missing data (see *Materials and Methods*). The abundance of rare variants is a product of the *Arabidopsis* mating system, population history (6), and our methodology, because each accession is compared against Col [supporting information (SI) Fig. 5]. If Col harbors a rare allele, the SFP frequency can be underestimated when the null hypothesis is rejected independently for each accession. To estimate error rates and sensitivity, we exploited known SNPs derived from dideoxy sequencing data (6). We selected significance thresholds balancing false positives and false negatives across the large number of features tested on the array. There are up to 1,563 array features with corresponding sequence data at a subset of the 23 accessions (SI Table 1). For example, in the Lz accession, there are 1,179 array features with sequence data. Forty-six (4%) features contain SNPs, whereas other features have no sequence variants. Among the 1,179 total features with sequence information, 49 SFPs were identified, 10 of which harbor no known SNPs within the exact 25-bp feature (10/1,179 = <1% false-positive rate, 10/49 = ≈20%; FDR). On the other hand, seven SNPs were not detected by array hybridization at our selected SFP threshold (7/46 = ≈15% false-negative rate). In a previous study, we reported a 66% and 43% SNP false-negative rate at a conservative and modest threshold, respectively, where the aim was a low false-discovery rate (11). The range in error rates among the accessions (SI Table 1) reflects subtle differences in sequence variation among accessions, experimental effects, and stochastic error due to the relatively small number of SNPs under consideration. A SNP may remain undetected if the feature on the array exhibits poor hybridization properties in general, or the polymorphism resides at the edge of the 25 mer (11, 15, 28). The false-negative rate can be improved with greater replication or alternatively by lowering the threshold, which comes at a cost of increased false positives. Detected SFPs that do not contain a sequence polymorphism may be statistical false positives or are likely the result of polymorphisms adjacent to the 25 mer, which may alter feature intensity via differential labeling (29).

Among the 373 features known to have SNPs by sequence identification across the 16 accessions where sequence was available (6), 304 are biallelic (82%), 57 are triallelic (15%), and 12 (3%) reside in a simple sequence repeat with greater than three alleles (SI Fig. 6). Subsequent analysis of genomic diversity assesses levels and patterns of variation across 50-kb regions. In this way, the rare and unaccounted-for observations of multiallelism are averaged away across the ≈100 features per region. Furthermore, regional patterns are compared with the rest of the genome that also contains the unobserved multiallelism bias. The entire SFP collection has been made available in a Web-searchable format, allowing users to search genomic regions or genes for natural variation (http://signal.salk.edu/cgi-bin/AtSFP).

An important question that can be addressed with high-density genotyping is, "How does the variation among accessions, or genealogy, change along the genome?" Fig. 1 shows an example of raw data. The haplotype patterns can be easily visualized by plotting relative hybridization differences in false color. For example, the accessions Col and Ita share a common haplotype for 3.25–3.38 Mb; Fl, Mr, and St share a different haplotype, and Sah a third (Fig. 1*A*). At 3.38–3.49 Mb, the pattern has been shuffled. Here Ita and Mr share correlated SFPs, St and Fl share another pattern, and Col and Sah a third. Ancestral recombination has mixed the haplotypes (Fig. 1*B*). The levels of variation vary widely, and the patterns change frequently across the genome. Images are available for all regions and all experiments (http://naturalvariation.org/accessionsSFP/

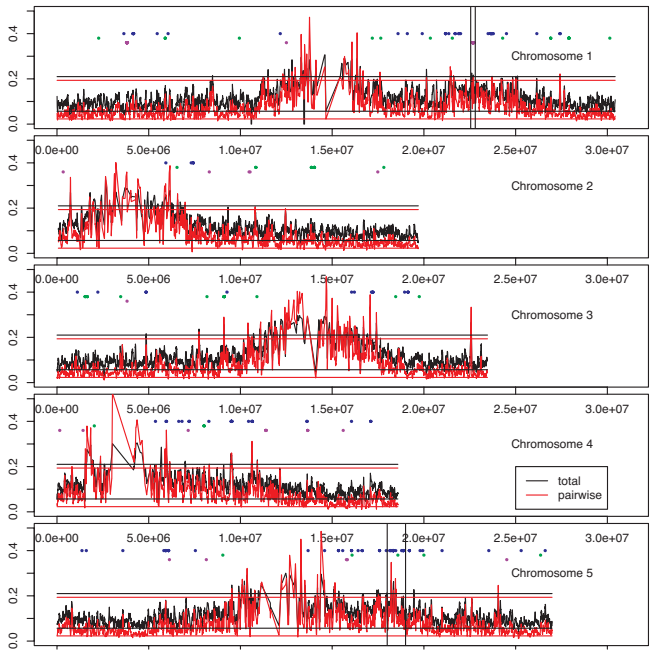

Fig. 2. Diversity along the five *Arabidopsis* chromosomes. The *y* axis shows SFP diversity at each 50-kb interval centered on a gene. The centromeres on all chromosomes, as well as the heterochromatic knob near the top of chromosome 4, are highly elevated compared with the rest of the genome, although estimates are less dense there as well. Variation spikes are also rampant. The blue dashes represent *R* genes, the green dashes represent RLPs, and the purple dashes represent S-locus proteins. Both total diversity (black) and pairwise diversity (red) reveal unusually high or low levels of variation exceeding the 2.5% genome-wide thresholds (horizontal black line, θ; red line, π). Vertical lines demark regions shown in detail in Fig. 3.

supplement/HapMapImages). A correlation dendrogram of the 23 accessions reveals genome-wide relatedness due to population structure as well as increased diversity among West Asian strains (SI Fig. 7). The west Asian strains Shah (Shahdara, Tajikistan), Kas (Kasmir, India), and Sorbo (Tajikistan) are closely related, as reported (5), and greatly diverge from other strains when compared on the array against the reference genotype Col. The increased diversity is consistent with Southwest Asia being part of an ancient glacial refugia (4).

Regional Diversity on a Genome Scale. We next explored variation within the genome by calculating SFP diversity in 50-kb sliding windows along the genome and found highly variable regions dispersed throughout with some gross patterns apparent (Fig. 2). We calculated both total diversity (black lines) and average pairwise diversity (red lines), which are expected to be equal under neutrality but show differences under nonneutral, selective, or demographic scenarios (30). Similar to previous reports, the distribution of both measures shows a strong skew toward low diversity and rare variant regions (Fig. 4 and SI Fig. 8, red bars) (6, 7). To test how the patterns of variation extend across genic polymorphism to the chromosome level, the gene order with their SFP genotypes was randomly shuffled with respect to position along the chromosome. Diversity was then recalculated in 50-kb sliding windows along the genome. This shuffled data set established false discovery thresholds under the null hypothesis that the observed patterns of variation are limited to within a gene (see yellow bars in Fig. 4 and SI Fig. 8). In addition, the empirical distribution across the genome was used to determine the frequency of variation observed at a particular region.

Highly variable regions include the centromeres as well as a region on chromosome four corresponding to the heterochromatic knob (Fig. 4) (31). The high levels of diversity observed at these heterochromatic regions described as having a low-recombination rate may be due to stochastic error from low gene and feature density. The prevalence of insertion/deletion polymorphisms near centromeres and other heterochromatin is also a likely contributor to this pattern in our data (32).

We analyzed closely the relationship between SFP diversity and recombination rate variation using a recently generated very high-density genetic linkage map that precisely defined 676 recombination events in 98 Col/L*er* recombinant inbred lines (33). The number of recombination events in a 1-Mb region centered on a gene was calculated here for every gene and compared with the total SFP diversity within a 50-kb window centered on that gene. Overall, there was a positive correlation with recombination rate and diversity ($R^2 = 3.1\%$), as seen previously (33). In this study, however, the diversity estimates span 23 accessions rather than just the mapping population. This observed positive correlation is consistent with previous observations in several organisms (34–36). However, many outliers exist especially in the heterochromatic region (SI Fig. 9). Heterochromatin makes up a large physical portion of the genome, as seen in Fig. 2, but few unique genic data points are available here to estimate diversity or recombination leading to the observed positive correlation across genic regions in the genome. With more comprehensive genome-wide data, this relationship should be revisited.

A region on chromosome five exhibited a particularly high level of diversity (Fig. 2), as measured by both total (θ) and pair-wise diversity (π). Within this region, three peaks exceeded the upper 2.5% tail of the chromosome-wide diversity distribution (Fig. 3 *A* and *B*), suggesting these are among the fastest-evolving regions in the genome. The two measures of diversity are sensitive to different patterns. Total diversity is sensitive to rare changes, whereas pairwise diversity is greater when there are only a small number of common haplotypes. The difference between the two measures is analogous to Tajima's D statistic, a test for selection (30). When pairwise diversity is higher (positive Tajima's D), this suggests balancing or frequency-dependent selection, whereas negative Tajima's D indicates a past selective sweep. The empirical distribution of the Tajima's D statistic calculated from SFP data in 50-kb intervals is skewed toward negative statistics (red bars, Fig. 4 and SI Fig. 8), similar to a previous reports based on *Arabidopsis* sequence data (6). This is also true from the permuted null distribution but to a lesser extent (yellow bars in Fig. 4 and SI Fig. 10). Nevertheless, the two patterns of selection can be identified that are rare in the genome and/or unlikely to occur in the null distribution.

Three high-diversity subregions with singleton or clusters of *R* genes were identified (blue dashes, 18.2, 18.3, and 18.44 Mb) (Fig. 3 *A* and *B*). The central high-diversity locus spans the *RPS4 R* gene (37). The Tajima's D scores at the *RPS4* region, however, are found quite often in the genome-wide distribution, i.e., no particular pattern was apparent. Tajima's D scores are strikingly high in the flanking regions, implicating underlying candidate genes as evolutionarily important. The left-most locus (≈18.22 Mb) contains four *R*-like genes, whereas the right-most (≈18.44-Mb) locus contains one putative *R* gene and a cluster of five unknown genes (not shown). Further downstream, another two relatively high-diversity regions flank a single (18.78 Mb), or a cluster (18.87 Mb) of, *R* gene homologs. The pattern of variation at the cluster (18.87 Mb) shows the opposite pattern, i.e., an excess of rare variants due to a potential past selective sweep (Fig. 4*B*). More extensive population genetic sequencing analysis is needed to confirm and resolve the predicted evolutionary signatures, and functional assays are warranted to understand their roles. Selection scores for each gene in the genome are available (SI Dataset 1).

Another region that exhibited unusual levels and patterns of diversity contains a cluster of S-locus protein kinase genes involved in sporophytic self incompatibility in the Brassicaceae (38) (Fig. 4

EVOLUTION

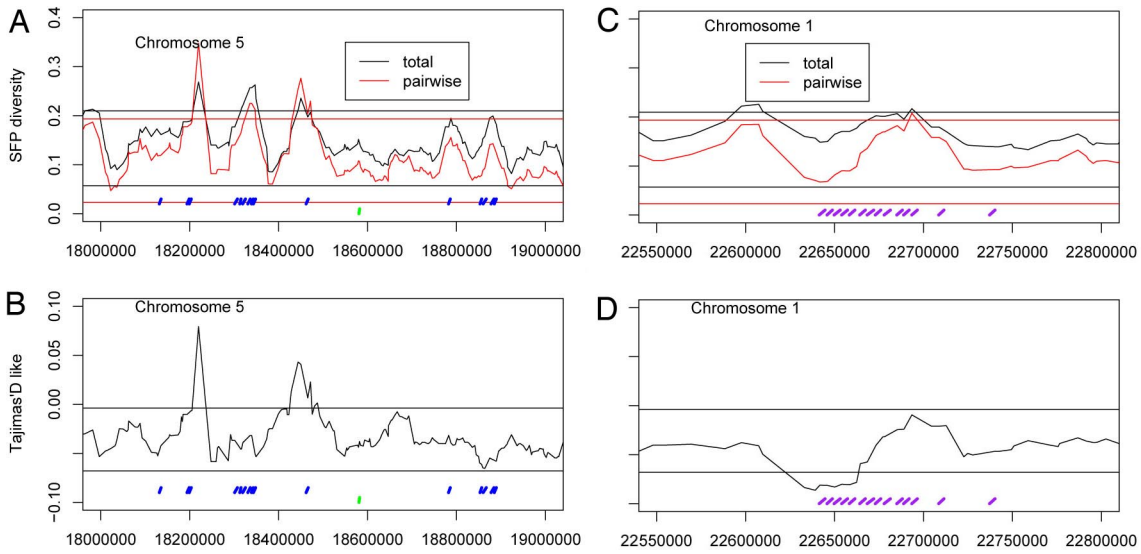

Fig. 3. SFPs reveal local regions of diversity and patterns of selection in the *R* and S-locus like genes. (*A*) A closer look at 18–19 Mb on chromosome 5 reveals interesting patterns of variation. (*B*) Contrasting the different measures of diversity allows one to infer the type of selective forces acting at different adjacent loci. For the gene clusters at 18.2 and 18.44 Mb, pairwise diversity is higher than total diversity, suggesting a pattern of frequency-dependent selection rare in the genome. This pattern is observed only at some of the *R* genes (blue bars) in the cluster. The pattern of variation at the central locus, 18.33 *RPS4*, however, is common in the genome. (*C*) Another region with high diversity spans a cluster of S-locus proteins, the gene family involved in self incompatibility. (*D*) Regions spanning this cluster show both rare negative and positive Tajima's D scores, suggesting a selective sweep near 22.65 Mb and frequency-dependent selection near 22.7 Mb. Plots of the entire genome are shown in SI Fig. 6. Colors and thresholds are as in Fig. 2.

C and *D*, purple dashes). *Arabidopsis* is self-compatible, and the chromosome one S-locus homologous loci are annotated as pseudogenes in Col-0. Nonetheless, high levels of diversity clearly exist for these genes (22.7 Mb), and patterns of variation (Tajima's D) are noteworthy (Fig. 3 *C* and *D*). S-locus protein genes at 22.65 Mb show a rare pattern, suggesting a selective sweep (<1% genome-wide lower tail), whereas at 22.7-Mb, positive values indicate frequency-dependent selection (<3% genome-wide upper tail). Active self-incompatibility genes in self-incompatible species are well known to have been under strong frequency-dependent selection (39). When an active receptor and ligand from the self-incompatible sister species *Arabidopsis lyrata* is transformed into *Arabidopsis*, some strains were rendered self-incompatible, confirming that in some cases the downstream mechanism retains functionality (38). Cryptic polymorphism in *Arabidopsis* was revealed in these transgenic lines for the degree of self incompatibility (40). A functional S-locus maps to an orthologous region on chromosome four; however, we see normal levels of diversity at this locus in our study. If the chromosome one locus had residual self-incompatibility function in the past, perhaps there was active selection for the degeneration at 22.65 Mb. Many independent mutation hits, allowing selfing, would provide a strong advantage when populations were isolated or pollinators absent. Conjecture aside, there is a striking pattern in this region that is rare in the genome and warrants functional assays to test such hypotheses.

Other regions in the genome show rare levels of diversity and patterns of selection. Underlying a handful of these loci are a certain class of candidate genes, the RLPs (SI Fig. 11 and below). Genome-wide plots of diversity and Tajima's D statistic are provided (SI Fig. 10) and can be searched at our Web site (http://naturalvariation.org/accessionSFP) or in tabular form (SI Dataset 1).

Gene Family Patterns. There are 150 genes annotated in the *Arabidopsis* genome that encode nucleotide-binding/leucine-rich repeat-(NB-LRR) type proteins, the most common type of *R* genes (41). One hundred eighteen are uniquely represented on the array. We compared the levels and patterns of NB-LRR diversity to the 46 unique RLP genes [of 56 total (42)] and 39 unique S-locus protein genes [of 39 + 9, S-locus receptor kinase genes (43) and S-locus glycoprotein genes (44)]. Fig. 4*B* shows the relative frequency distribution of total diversity and Tajima's D scores across *R* gene homologs, RLP genes, and S-locus genes in comparison to the genomic and null distribution (Fig. 4*A*). Diversity was much higher for all three gene families relative to the genome-wide or null distributions (notice the enrichment in the right tail of the distribution). The pattern of variation (Tajima's D) is also noteworthy for these gene families (Fig. 4*D*). Strikingly, there are 7 and 35 *R* gene homologs with scores exceeding the lower and upper 2.5% tails of the genome-wide null distribution, respectively, where approximately four are expected (Fig. 4*D*, blue bars vs. Fig. 4*A*, yellow bars). *R* genes evolve by positive selection, accumulating many amino acid changes (41, 45). Balancing selection has been inferred from polymorphism at several *R* genes and *R* gene homologs before (46–48) and may be due to a fitness tradeoff associated with resistance (49). However, the pattern suggestive of a selective sweep has also been observed and may be due to an evolutionary arms race of competing pathogen virulence and plant resistance (50). More recently, 27 *R* gene homologs in *Arabidopsis* were surveyed for polymorphisms, identifying members with very high or low levels of nonsynonymous changes, suggesting that different forms of selection have shaped evolution at these loci (51). We found rare genomic patterns suggestive of both types of selective forces with a simple scan across the entire family. In this way, we highlight which *R* genes are most likely to have been or are perhaps still under active selection, and which show patterns indistinguishable from the rest of the genome that could now be pseudogenes. Precisely how many genes are actually under selection, however, is difficult to know without an appropriate null model.

Like R proteins, RLPs have leucine-rich repeats but do not contain a nucleotide-binding site and belong to a protein family including Cf-*R* genes in tomato (42). Their functional importance can be inferred by investigation of their evolutionary signatures. There is an enrichment of high-diversity members (Fig. 4*A*, green bars). In addition, four and eight RLPs were found in the lower and upper tails of the Tajima's D distribution, respectively (Fig. 4*D*, green bars), where only two to three total are expected (Fig. 4*C*,

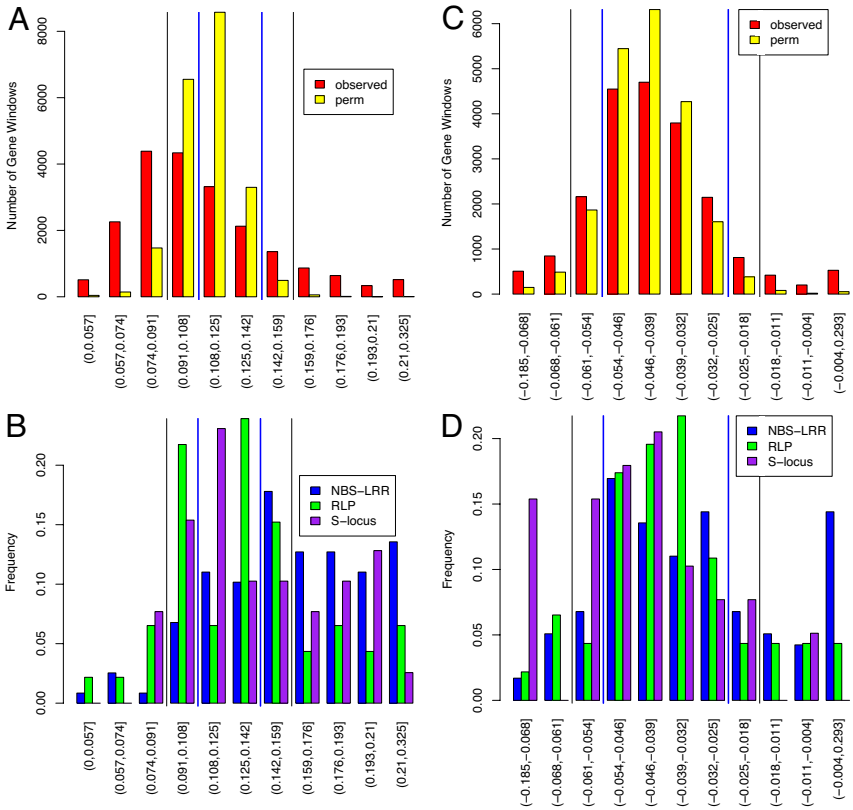

Fig. 4. Distribution of diversity and selection statistics at the genome and gene family level. The vertical black lines delimit 95% of the gene position-shuffled null distribution ($P < 0.05$ outside). The vertical blue lines represent a <2% FDR under the same null distribution. (*A*) Empirical distribution of diversity in 50-kb windows in red is shown relative to the gene position shuffled null distribution in yellow. (*B*) In comparison to the genome-wide distribution directly above (red), the diversity seen in select gene families is elevated or shifted to the right. (*C*) The empirical (red) and null (yellow) Tajima's D distribution is shown. (*D*) Tajima's D distribution in select gene families is enriched in both the lower and upper tail, suggesting that selection has acted on regions where these genes reside. SI Fig. 8 shows that the distribution across control gene families is similar to the genome-wide distribution.

yellow bars). These would be the first members to test by functional characterization to investigate allelic phenotypic effects. The S-locus protein gene family also shows an excess of members from high-diversity gene regions, and the overall pattern of diversity suggests an excess of rare variants (Fig. 4*D*, purple bars). This could be due to active degradation of such genes if residual activity resulted in partial sterility.

As a control, we looked at several other gene families that have members which also often occur in tandem clusters (SI Fig. 8). One hundred twenty-seven members of the *bHLH* transcription factor gene family have been annotated (http://arabidopsis.med.ohio-state.edu) (52), 88 of which are uniquely represented on the array. RING-type E3 Ubiquitin ligase proteins (53) and F-box-containing subunits of the SCF proteasome (54) are some of the largest gene families in *Arabidopsis*, with 444 and 672 members, respectively; 356 and 346 are uniquely represented on the array. Investigation of the evolutionary signatures of these three families did not show particular striking levels or patterns of variation. However, several members were identified that have unusually low or high amounts of diversity with rare patterns of variation providing an evolutionary annotation to aid functional studies.

Conclusions

We profiled patterns of genome-wide diversity using a relatively rapid and cost-effective platform, genomic DNA hybridization to commercially available gene expression arrays. A relative statistic measuring variation that suggests a certain type of selective pressure was calculated for every gene (SI Dataset 1). The procedures used are relatively easy to perform and thus can be applied in any organism with a high-density oligonucleotide array and a physical map. Prior knowledge of global SNPs is not needed. The standard ascertainment bias due to comparison to a reference genome must be acknowledged, but this caveat pertains to other technologies as well where a reference genome is used for sequence assembly. As array designs and densities improve, along with sequencing technology, this basic method of evolutionary annotation can be applied to nonmodel organisms where ecological knowledge provides a context to understand the driving forces behind the observed evolutionary patterns. Finally, the union of diversity scanning, genetic markers for mapping functional phenotypic variation, and polymorphism-controlled gene expression variation provides a comprehensive approach for determining the genetic basis of adaptation (55).

Materials and Methods

Plant Growth Conditions. Plants were grown in soil under greenhouse conditions, and a single leaf was extracted from each plant for each biological replicate at 3 weeks of age for experiments 1, 2, 4, and 5. For experiment 3, seedlings were grown on agar plates for 1 week.

Array Protocol and SFP Analysis. Hybridization and analysis were performed as described (26) in five independent experiments: experiment 1, Col-0 (CS6673), Cvi-0 (CS6675), Kas-1 (CS6751), L*er* (CS20), Shah (C6180), and Bay-0 (CS6608); experiment 2, Col(*gl1*) (CS3879), Est (CS6173), KendL (Lehle seed WT-16-03), Mt-0 (CS6799), Nd-1 (CS1636), Sorbo (CS931), Van-0 (CS6884), Ws-2

EVOLUTION

(CS2360), and C24 (CS906); experiment 3, Col-2 (CS907), Fl-1 (CS6706), Ita-0 (CS6097), Mr-0 (CS6795), Sah-0 (CS6917), and St-0 (CS6863); experiment 4, Col (CS6673), Bur-0 (CS6643), and Lz-0 (CS6788) (20); and experiment 5, Col arrays from experiments 1 and 2 vs. Ts-1 (CS1552), Tsu-1 (CS1640), and Ws-0 (CS1602). Two biological replicates were used for Kas-1, KendL, C24, Shah, and Tsu-1 due to failures, with all three biological replicates for all others. Feature intensities of accessions were compared with Col by calculating a D statistic (13) within each experiment, and a permutation FDR threshold was applied. When a particular feature exhibited reduced hybridization in a test accession relative to Col, ensuring the precise genomic position would be known from the reference physical map, a value of one was assigned to the SFP. Duplication, where the test accession had a greater hybridization signal than Col, and ambiguous SFPs, with marginal D statistics, were treated as missing data. D statistics falling within 90% of the permutation distribution were called 0, i.e., no hybridization polymorphism (SI Fig. 5). The zero-value genotype (non-SFP call) is accepting the null hypothesis without evidence against it, and this results in a lower sensitivity (see SI Table 1). A final caveat is that a single accession was used as a reference; therefore, rare alleles in the reference strain may be estimated as more common due to type 2 error resulting in overestimates of diversity at a locus. That being said, the results reported here compare relative patterns across all genomic loci observed under this calling method. Analysis scripts, raw data, and supplementary material are available at http://naturalvariation.org/accessionSFP.

Diversity in Sliding Windows. The polymorphism detection and genotyping power are stronger when averaged across 50-kb intervals containing ≈100 (86–126 inner-quartile range) 25-bp features. The added resolution more than makes up for the lower per-site accuracy (23, 25) compared with dideoxy sequencing of a single ≈500-bp fragment every 100 kb (6). Features within ±25 kb of the central probe in a gene were used to estimate diversity. We calculated total and pairwise diversity in sliding windows of 50 kb along the genome in gene-size steps, ≈3.8 kb (2.2–6.6 kb inner-quartile range), because features are not evenly spaced on the ATH1 expression array. Total diversity is the number of SFPs in a window, scaled by the average number of accessions observed to account for missing data. This is an estimate of Watterson's θ developed for general biallelic markers, which is sensitive to rare variants (56). Pairwise diversity is the average number of differences in haplotype calls across all pairs of strains (also scaled for missing data) and is an estimate of Tajima's π (57), which is sensitive to common variants. A skewed distribution toward low values was observed for both θ and π (Fig. 4 and SI Fig. 8). Of less importance here is the magnitude of the estimate, but rather the relative differences in diversity between regions, because the empirical genome-wide distribution is known. Contrasting these estimates of diversity reveals different patents of selection [Tajima's D (30) = $(\pi - \theta)$/normalization)], with negative values indicating a selective sweep, and positive values indicating balancing or frequency-dependent selection. SFP estimates of Tajima's D discussed here were unnormalized.

We thank Tsegaye Dabi for help with growing plants and Thomas Gal for help with array annotation. J.O.B. was a Helen Hay Whitney Fellow, and concluding work was supported by National Institutes of Health (NIH) Grant R01GM073822. We acknowledge funding from NIH Grant GM62932 (to J.C., D.W., J.D.W., and T.P.M.). J.D.G.J. was supported by a Gatsby Foundation grant to the Sainsbury Laboratory. T.P.M. is also supported by NIH Fellowship 5 F32 GM068381, and J.C. is also supported by the Howard Hughes Medical Institute. In addition, this work was supported by NIH Grants GM56006 and GM67837 (to S.A.K.), by a Ruth L. Kirschstein NIH Postdoctoral Fellowship (to S.P.H.), and by National Science Foundation Grants DBI 0313578 and DBI 0420126 (to J.R.E.). D.E.S. and I.R.B. were supported by National Science Foundation Arabidopsis 2020 Grant IOB 0419695.

1. Toomajian C, Hu TT, Aranzana MJ, Lister C, Tang C, Zheng H, Zhao K, Calabrese P, Dean C, Nordborg M (2006) *PLoS Biol* 4:e137.
2. Yoshida K, Kamiya T, Kawabe A, Miyashita NT (2003) *Genes Genet Syst* 78:11–21.
3. Miyashita NT, Kawabe A, Innan H (1999) *Genetics* 152:1723–1731.
4. Sharbel TF, Haubold B, Mitchell-Olds T (2000) *Mol Ecol* 9:2109–2118.
5. Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, Hagenblad J, Kreitman M, Maloof JN, Noyes T, Oefner PJ, *et al.* (2002) *Nat Genet* 30:190–193.
6. Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R, *et al.* (2005) *PLoS Biol* 3:e196.
7. Schmid KJ, Ramos-Onsins S, Ringys-Beckstein H, Weisshaar B, Mitchell-Olds T (2005) *Genetics* 169:1601–1615.
8. Borevitz JO, Chory J (2004) *Curr Opin Plant Biol* 7:132–136.
9. Hazen SP, Kay SA (2003) *Trends Plants Sci* 8:413–416.
10. Winzeler EA, Richards DR, Conway AR, Goldstein AL, Kalman S, McCullough MJ, McCusker JH, Stevens DA, Wodicka L, Lockhart DJ, *et al.* (1998) *Science* 281:1194–1197.
11. Borevitz JO, Liang D, Plouffe D, Chang HS, Zhu T, Weigel D, Berry CC, Winzeler E, Chory J (2003) *Genome Res* 13:513–523.
12. Kumar R, Qiu J, Joshi T, Valliyodan B, Xu D, Nguyen HT (2007) *PLoS ONE* 2:e284.
13. Turner TL, Hahn MW, Nuzhdin SV (2005) *PLoS Biol* 3:e285.
14. Kidgell C, Volkman SK, Daily J, Borevitz JO, Plouffe D, Zhou Y, Johnson JR, Le Roch K, Sarr O, Ndir O, Mboup S, *et al.* (2006) *PLoS Pathogenet* 2:e57.
15. Rostoks N, Borevitz JO, Hedley PE, Russell J, Mudie S, Morris J, Cardle L, Marshall DF, Waugh R (2005) *Genome Biol* 6:R54.
16. Cui X, Xu J, Asghar R, Condamine P, Svensson JT, Wanamaker S, Stein N, Roose M, Close TJ (2005) *Bioinformatics* 21:3852–3858.
17. Hazen SP, Borevitz JO, Harmon FG, Pruneda-Paz JL, Schultz TF, Yanovsky MJ, Liljegren SJ, Ecker JR, Kay SA (2005) *Plant Physiol* 138:990–997.
18. Hazen SP, Schultz TF, Pruneda-Paz JL, Borevitz JO, Ecker JR, Kay SA (2005) *Proc Natl Acad Sci USA* 102:10387–10392.
19. Rus A, Baxter I, Muthukumar B, Gustin J, Lahner B, Yakubova E, Salt DE (2006) *PLoS Genet* 2:e210.
20. Werner JD, Borevitz JO, Uhlenhaut H, Ecker JR, Chory J, Weigel D (2005) *Genetics* 170:1197–1207.
21. Werner JD, Borevitz JO, Warthmann N, Trainer GT, Ecker JR, Chory J, Weigel D (2005) *Proc Natl Acad Sci USA* 102:2460–2465.
22. Wolyn DJ, Borevitz JO, Loudet O, Schwartz C, Maloof J, Ecker JR, Berry CC, Chory J (2004) *Genetics* 167:907–917.
23. Jiang R, Marjoram P, Borevitz JO, Tavare S (2006) *Genetics* 173:2257–2267.
24. Shapiro JA, Huang W, Zhang C, Hubisz MJ, Lu J, Turissini DA, Fang S, Wang HY, Hudson RR, Nielsen R, *et al.* (2007) *Proc Natl Acad Sci USA* 104:2271–2276.
25. Kim S, Zhao K, Jiang R, Molitor J, Borevitz JO, Nordborg M, Marjoram P (2006) *Genetics* 173:1125–1133.
26. Borevitz J (2006) *Methods Mol Biol* 323:137–145.
27. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) *Bioinformatics* 19:185–193.
28. Ronald J, Akey JM, Whittle J, Smith EN, Yvert G, Kruglyak L (2005) *Genome Res* 15:284–291.
29. Naef F, Lim DA, Patil N, Magnasco M (2002) *Phys Rev E* 65:040902.
30. Tajima F (1989) *Genetics* 123:585–595.
31. Fransz PF, Armstrong S, de Jong JH, Parnell LD, van Drunen C, Dean C, Zabel P, Bisseling T, Jones GH (2000) *Cell* 100:367–376.
32. *Arabidopsis* Genome Initiative (2000) *Nature* 408:796–815.
33. Singer T, Fan Y, Chang HS, Zhu T, Hazen SP, Briggs SP (2006) *PLoS Genet* 2:e144.
34. Hellmann I, Ebersberger I, Ptak SE, Paabo S, Przeworski M (2003) *Am J Hum Genet* 72:1527–1535.
35. Wright SI, Foxe JP, DeRose-Wilson L, Kawabe A, Looseley M, Gaut BS, Charlesworth D (2006) *Genetics* 174:1421–1430.
36. Takahashi A, Liu YH, Saitou N (2004) *Mol Biol Evol* 21:404–409.
37. Gassmann W, Hinsch ME, Staskawicz BJ (1999) *Plant J* 20:265–277.
38. Nasrallah ME, Liu P, Nasrallah JB (2002) *Science* 297:247–249.
39. Igic B, Kohn JR (2001) *Proc Natl Acad Sci USA* 98:13167–13171.
40. Liu P, Sherman-Broyles S, Nasrallah ME, Nasrallah JB (2007) *Curr Biol* 17:734–740.
41. Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW (2003) *Plant Cell* 15:809–834.
42. Fritz-Laylin LK, Krishnamurthy N, Tor M, Sjolander KV, Jones JD (2005) *Plant Physiol* 138:611–623.
43. Shiu SH, Karlowski WM, Pan R, Tzeng YH, Mayer KF, Li WH (2004) *Plant Cell* 16:1220–1234.
44. Shiu SH, Bleecker AB (2003) *Plant Physiol* 132:530–543.
45. Mondragon-Palomino M, Meyers BC, Michelmore RW, Gaut BS (2002) *Genome Res* 12:1305–1315.
46. Mauricio R, Stahl EA, Korves T, Tian D, Kreitman M, Bergelson J (2003) *Genetics* 163:735–746.
47. Stahl EA, Dwyer G, Mauricio R, Kreitman M, Bergelson J (1999) *Nature* 400:667–671.
48. Tian DC, Araki H, Stahl E, Bergelson J, Kreitman M (2002) *Proc Natl Acad Sci USA* 99:11525–11530.
49. Tian D, Traw MB, Chen JQ, Kreitman M, Bergelson J (2003) *Nature* 423:74–77.
50. Bergelson J, Kreitman M, Stahl EA, Tian D (2001) *Science* 292:2281–2285.
51. Bakker EG, Toomajian C, Kreitman M, Bergelson J (2006) *Plant Cell* 18:1803–1818.
52. Bailey PC, Martin C, Toledo-Ortiz G, Quail PH, Huq E, Heim MA, Jakoby M, Werber M, Weisshaar B (2003) *Plant Cell* 15:2497–2502.
53. Stone SL, Hauksdottir H, Troy A, Herschleb J, Kraft E, Callis J (2005) *Plant Physiol* 137:13–30.
54. Gagne JM, Downes BP, Shiu SH, Durski AM, Vierstra RD (2002) *Proc Natl Acad Sci USA* 99:11519–11524.
55. Zhang X, Richards EJ, Borevitz JO (2007) *Curr Opin Plant Biol* 10:142–148.
56. Watterson GA (1975) *Theor Popul Biol* 7:256–276.
57. Tajima F (1983) *Genetics* 105:437–460.