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[5], Glenn Fu[7], David A. Hinds[7], Huaming Chen[6], Kelly A. Frazer[7], Daniel H. Huson[4], Bernhard Schölkopf[3], Magnus Nordborg[5], Gunnar Rätsch[2], Joseph R. Ecker[6,8], and Detlef Weigel[1,8†]

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

*These authors contributed equally to this work.

[†]To whom correspondence should be addressed. E-mail: weigel@weigelworld.org

One-sentence summary: Resequencing of 20 *Arabidopsis thaliana* strains with high-density oligonucleotide arrays identified hundreds of thousands of polymorphisms, allowing a systematic description of the distribution of polymorphisms in the genome, and enabling future whole-genome population genomic studies.

The genomes of individuals from the same species vary in sequence as a result of diverse evolutionary processes. To examine the patterns of and 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 non-random among gene families, and genes mediating interaction with the biotic environment harbor 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 dramatic example in which almost all variation was removed in a 500 kb window.

Comprehensive polymorphism data are essential for the systematic identification of sequence variants affecting phenotypes (*1*). Despite progress with new technologies, however, 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 successfully to identify a large fraction of the SNP variation in human and mouse (*3, 4*). Here, we have applied this technology to 20 wild strains (accessions) of *A. thaliana*, for which a single genome sequence from the Col-0 accession was generated in 2000 (*5*). The resulting polymorphism dataset captures much of the common sequence variation that distinguishes *A. thaliana* accessions. We have used this information to determine in a systematic fashion the types of sequences and genes that differ between accessions, and to provide a high resolution description of the genome-wide distribution of polymorphism 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 accessions, such as Ler-1 (Table S1). Col-0, the reference accession, was included to aid in polymorphism detection. For all but one of the 20 accessions, 1,213 fragments of ~500 bp in length and spaced throughout the genome had been sampled by dideoxy sequencing, identifying between 2,266 to 3,949 nucleotide substitutions per accession relative to Col-0 (*6*). This dataset, hereafter called “2010”, allowed us to assess precision and recall for polymorphic predictions.

Whole genome amplified DNA from each accession was hybridized to resequencing microarrays designed to interrogate > 99.99% of bases in the 119 Mb reference genome sequence (*7*). On these arrays, each position was queried with forward and reverse strand probe quartets consisting of 25-mer oligonucleotides (Fig. S1). Within a probe quartet, all four nucleotides were represented at the central oligonucleotide base, and differences in relative intensities across probe quartets indicated potential SNPs. For tightly linked SNPs, however, all probes harbor at least one mismatch. This suppresses hybridization, and thereby confounds SNP detection (Fig. S1).

Two computational methods were employed 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 particularly favorable for hybridization-based SNP detection (*7*). Accordingly, recovery of SNPs using a model-based algorithm (MB) (*3, 4*) was better for coding regions (36% versus 15%), at a three-fold lower FDR (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 using Support Vector Machines (SVMs) (*8, 9*) for SNP identification (*7*) (Figs. S3-S8). The training step exploited 2010 data, and as input we used information for all probes corresponding to positions within a 9 bp window centered on candidate polymorphisms (*7*). Apart from hybridization data, we also 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. 1A, B and Table S3).

At a 2% FDR, precision and recall for the MB and ML methods were similar; however, the two methods were complementary, with 60% of predictions made using 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 indel was higher for the MB method, while recall for SNPs separated by 7 to 30 bp from a nearby polymorphism was two-fold higher for the ML method. For very closely linked SNPs (<7 bp), recovery was low with both methods (~3%). FDRs for either method peaked in regions of low hybridization quality, an effect of polymorphism but also of other factors (*7*) (Fig. 2E).

For subsequent analyses, we combined all MB predictions with ML predictions supported at a 2% FDR. The resulting dataset, "MBML2", consisted of 648,570 nonredundant SNPs (*7, 10*) (Fig. 2), an average of 1 polymorphic position per 166 nonrepetitive positions in the genome. Within MBML2, the FDR for SNPs supported by both methods is very low (0.2%), while FDRs for those supported by only one method are markedly higher (Tables 1 and S3). A caveat for our error estimates is that 2010, which we used for specificity and sensitivity assessment of the two prediction methods, is underrepresented for noncoding sequences and sequences highly dissimilar to the reference (*6, 7*).

Apart from SNPs, deletions are detectable on high-density oligonucleotide arrays as regions of reduced hybridization (*11*). A similar signal is expected for sequences highly dissimilar to the reference (Fig. S1). A heuristic algorithm was developed 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 dataset, which was ascertained by PCR, missing data corresponds 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 from 2010. We also attempted direct validation of 382 PRPs by PCR and sequencing, obtaining complete or partial sequence data for 171 products. We identified 124 deletions of between 50 bp to over 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, a potential factor 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 also developed a base calling algorithm to identify positions identical to the reference at low false discovery rates (*7*). Between 80.3 to 92.3% of coding, and 39.7 to 61.2% of intergenic positions, were predicted as 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 non-redundant MBML2 SNPs on the 26,541 annotated 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, 1,227 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. As 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 that large-effect SNPs resulted predominately 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 therefore only moderately higher than for an average SNP in MBML2 (*7*). In total, 1,614 genes harbor at least one large-effect SNP. In addition, the coding regions of 1,191 genes are at least partially included in PRPs, i.e., are highly polymorphic or deleted. The overlap between the two classes is greater than expected by chance (χ^2-test, $P<10^{-16}$). Together, large-effect SNPs and PRPs, hereafter referred to simply as "major-effect changes", affected 2,495, or 9.4%, of *A. thaliana* genes.

The number of genes harboring major-effect changes varies significantly according to annotation support, duplication status, and gene family (Fig 3A; χ^2 tests, $P<10^{-53}$ for differences within each category). Correction for gene size and repetitive content does not appreciably change the observed patterns (Fig. S11). By annotation, genes known to be expressed but otherwise lacking functional support or high homology ("expressed unknown"), as well as genes without expression support ("not expressed"), were overrepresented. In addition, of 836 *A. thaliana* genes that either lack or have only moderate homology to *Populus trichocarpa* (*7*), the closest sequenced genome to *Arabidopsis* (*13*), 26.0% harbored major-effect changes (compared to 8.9% for all other genes). Poor gene annotation almost certainly contributes to this finding, although rapid gene evolution may also play a part. Consistent with relaxed purifying selection following recent gene duplication (*14*), tandem duplicates are 3.2 and 2.2-fold overrepresented for major-effect changes relative to segmentally duplicated and non-duplicated genes, respectively (Fig. 3A). Segmentally duplicated genes in *A. thaliana* resulted from ancient genome-wide duplications (*5, 15*), and that these genes harbor relatively few major-effect changes is consistent with earlier work suggesting strong purifying selection for duplicates that have persisted over long evolutionary time frames (*13, 14*).

Analysis of individual gene families provided additional insights. Families involved in basic biological processes (e.g., ribosomal function), as well as families involved in transcriptional regulation (e.g., MYB and bHLH transcription factors), harbor relatively few major-effect changes (Fig. 3A). In contrast, 60% of NB-LRR (Nucleotide Binding-Leucine Rich Repeat) genes (*7*) and 15% of RLK (Receptor-Like Kinase) 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 polymorphic, with presence/absence polymorphisms being 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. RLK genes, while having diverse functions (*20*), 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 resistance and growth (*22*).

We also found major-effect changes in 143 members of the F-box superfamily, which comprises over 660 genes in *A. thaliana* (*23*) (Fig. 3A). This finding, in combination with other data (*13, 24*), shows that 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 pathogens (*24*), experimental support for this hypothesis is lacking. The polymorphisms here described provide a resource for ascribing biological roles to members of this large, yet poorly characterized, gene family.

Signatures of selection by SNP type and gene family

To assess the extent to which the variation we describe 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, we observed a skew towards low frequency variants at nonsynonymous relative to synonynmous sites (*6*) (Fig. 3B). This frequency skew is much more dramatic for SNPs that introduce premature stop codons, albeit less extreme for other large-effect SNPs. That SNPs introducing premature stops codons tend to be rare suggests that, at least under natural settings, these changes are associated with fitness costs.

While allele frequency distributions across gene families were broadly similar, NB-LRR genes were an exception (Figs. 3B and S12). Here, both nonsynonymous and synonymous variants were strongly skewed towards high frequency compared to the genome average (Fig. 3B). This shift is a hallmark of some form of balancing selection (perhaps in the form of regional adaptation), and agrees with earlier work, based on fewer family members, that suggested this mode of selection to be common 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, over 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 (Figs. 3A and S12). Although F-box genes harbored the second-highest occurrence of major-effect changes, allele frequency distributions were similar to the average (Fig. 3B).

Genome-wide patterns of polymorphism

Finally, turning to broader patterns of variation, we found significant non-random distribution of polymorphism levels across the genome (Table S12 and Fig. 4), in agreement with an earlier result based on sparer sampling (*6*) (Fig. S13). Regions of elevated average polymorphism extend from the centromeres to beyond the pericentromeric regions. Clusters of NB-LRR genes (*7, 25*) were also associated with elevated levels of polymorphism (e.g., between 21 and 25 Mb on chromosome 1).

As these patterns are found in synonymous sites, which appear to evolve near-neutrally (*26*), it is unlikely they can be explained by a simple relaxation of selective constraint or the direct action of diversifying selection on many sites. Possible fundamental differences in the mutation rate on a chromosomal scale, which can soon be assessed using outgroup data (*27*), could explain the patterns. However, this does not seem to be the case for base composition, as patterns are also evident in specific subclasses of SNPs, e.g., A/T SNPs (Fig. S14). Nor do repetitive sequences appear to explain this; our data largely exclude repeats, and polymorphism levels are not strongly correlated with repeat density (Table S12).

We therefore hypothesize that regions of increased polymorphism are partly due to above-average coalescence times, possibly due to gene conversion by non-homologous recombination or balancing selection. Random genetic drift and population structure can contribute to a high variance in coalescence times, yet they are not expected to produce the systematic patterns we see. Gene conversion, which is expected in duplicated regions (*27*), and which has previously been identified in NB-LRR gene families (*28*), would cause coalescence events between non-homologous regions. While gene conversion may contribute to the observed pattern, it seems unlikely to be the full explanation because: elevated synonymous site polymorphism remains around NB-LRR gene clusters even after these genes are removed (Fig. S15); and the patterns are observed for intergenic sites (Fig. 4), where the potential for non-homologous gene conversion should be low.

Alternatively, balancing selection, which can increase coalescence times for regions linked to selectively maintained polymorphisms (*29*) and may be more easily detected in selfing organisms (*30*), has been detected in *A. thaliana* (*22, 31, 32*), and may be a general force influencing patterns of variation in NB-LRR genes. Clusters of tightly linked genes subject to balancing selection may underlie the high polymorphism associated with NB-LRR gene clusters (*25*) similar to what has been observed in the vertebrate MHC (*33*). Our findings, combined with previous studies (*17*), indicate that balancing selection, perhaps in the form of adaptation to local environments, is a major force in ensuring diversity at disease resistance loci in *A. thaliana*. Our data also suggest that selection at disease resistance loci, which in the case of NB-LRR genes have a clustered distribution in the genome, may underlie chromosomal-scale patterns of sequence diversity and non-neutral evolution similar to that observed in other species (*34*).

In contrast, regions of low polymorphism might reflect recent positive selection, or “selective sweeps” (*35, 36*) characterized by extensive haplotype sharing. A recent study using 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* (*37*). The current data set confirmed extensive haplotype sharing of up to 600 kb around *FRI* (Fig. S16), as well as haplotype sharing around other low-frequency candidate alleles (*37*) (Fig. S17).

We looked for evidence of additional sweeps in the form of extensive haplotype sharing across at least 50 kb (Figs. 5 and S18-S19). Because of its composition and size, our sample is only suited for discovering species-wide selective 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 (Figs. 5). The two unaffected accessions, Cvi-0 and Lov-5, come from the periphery of the range of *A. thaliana* and may have escaped the sweep because of different selective environments or geographic isolation. The region of most extreme haplotype sharing extends from 20.34-20.49 Mb and contains 50 annotated genes (Table S13) including *RPP27*, which functions in resistance to the pathogen downy mildew (*38*). Haplotype sharing patterns revealed several additional candidates for sweeps (Figs. S18-S21). With the SNPs identified in this project, and the ability to determine their frequencies in hundreds to thousands of accessions (*39*), the goal of understanding the forces shaping diversity at global, regional, and local scales will soon be within reach.

Conclusion

We have used array-based methods to generate a comprehensive polymorphism resource for *A. thaliana.* The SNP set we describe is highly applicable for linkage disequilibrium mapping studies. In addition, we identified hundreds of thousands of polymorphisms in both coding and noncoding regions, thereby 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 (*40*). In contrast to these plants, *A. thaliana* has a compact genome consisting largely of unique sequences. Nevertheless, our data emphasize that even for species with streamlined genomes, substantial sequence variation can and does exist between individuals.

In the laboratory, mutations identified in phenotypic screens typically have drastic 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 constraint under natural conditions. A possible explanation could be variation in copy number for genic sequences; 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 (Jul, 2005).
7. Supporting tables, figures and materials and methods are available as supporting material on Science Online.
8. B. Schölkopf, A. Smola, *Learning with Kernels* (MIT-Press Cambridge, 2002).
9. V. N. Vapnik, *The Nature of Statistical Learning Theory* (Springer, 1995).
10. SNP and PRP datasets along with effects on genes and pseudochromosome sequences are at The Arabidopsis Information Resource (TAIR) (http://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.
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. Y. Belkhadir, R. Subramaniam, J. L. Dangl, *Curr Opin Plant Biol* **7**, 391 (2004).
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. P. Andolfatto, *Nature* **437**, 1149 (2005).
27. Eight-fold genome coverage for the close *A. thaliana* relatives, *A. lyrata* and *Capsella rubella*, is being generated by the Joint Genome Institute (http://www.jgi.doe.gov/).
28. M. Mondragon-Palomino, B. S. Gaut, *Mol Biol Evol* **22**, 2444 (2005).
29. R. R. Hudson, N. L. Kaplan, *Genetics* **120**, 831 (1988).
30. M. Nordborg, *Genetics* **146**, 1501 (1997).
31. J. Kroymann, T. Mitchell-Olds, *Nature* **435**, 95 (2005).
32. D. Tian, H. Araki, E. Stahl, J. Bergelson, M. Kreitman, *Proc Natl Acad Sci U S A* **99**, 11525 (2002).

33. M. Nordborg, H. Innan, *Genetics* **163**, 1201 (2003).
34. A. L. Hughes, M. Nei, *Nature* **335**, 167 (1988).
35. N. L. Kaplan, R. R. Hudson, C. H. Langley, *Genetics* **123**, 887 (1989).
36. J. M. Smith, J. Haigh, *Genet Res* **23**, 23 (1974).
37. C. Toomajian *et al.*, *PLoS Biol* **4**, e137 (2006).
38. M. Tor *et al.*, *Plant Physiol* **135**, 1100 (2004).
39. A. C. Syvanen, *Nat Genet* **37 Suppl**, S5 (2005).
40. M. Morgante, *Curr Opin Biotechnol* **17**, 168 (2006).

41. We thank G. Nielson and H. Huang for bioinformatics support, R. Gupta and M. Morenzoni for information management, C. Shindo, C. Dean, J. Borevitz, and T. Altman for contributing seed stocks, and K. Schneeberger for helpful discussions. Supported by Innovation Funds of the Max Planck Society, NIH (HG002790 to M. Waterman, and GM62932 to DW), NSF (DEB-0115062 to MN, and DBI-0520253 to JRE), an NIH NRSA fellowship to CT, and core funding from the Max Planck Society. DW is a director of the Max Planck Institute.

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

	Mean # SNPs per accession by method [FDR (%)]			
SNP type	Total	MB only	MB ∩ ML	ML only
Coding	53,700 [2.0]	11,379 [3.2]	27,833 [0.1]	14,488 [4.8]
Intron+UTR	29,395 [3.1]	5,762 [9.6]	11,652 [0.4]	11,981 [2.6]
Intergenic	60,478 [3.5]	22,395 [7.3]	17,976 [0.3]	20,107 [3.6]
All	143,572 [2.8]	39,536 [6.5]	57,461 [0.2]	46,575 [3.7]

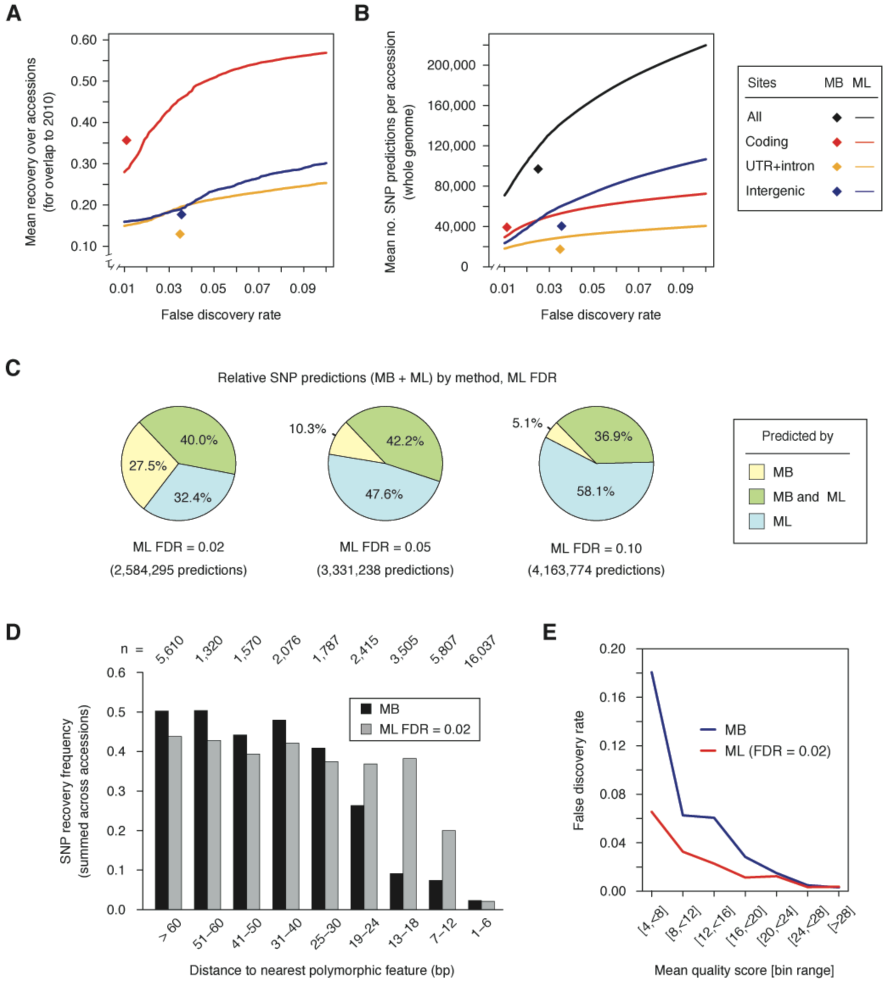
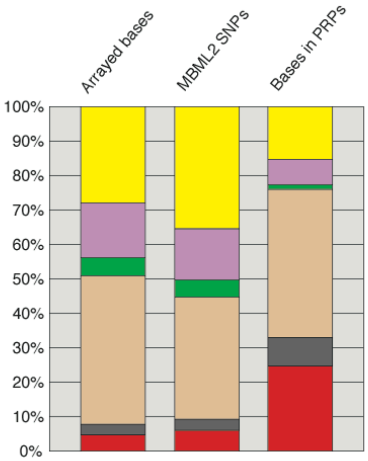


Fig. 1. Comparison of SNP detection methods. (A) FDR-dependent recovery of 51,100 known SNPs in 2010 fragments by the MB and ML methods. Because of small sample size for UTR SNPs in 2010, this group was combined with intron sequences. (B) FDR-dependent recovery across the entire genome by both methods using precision estimates from 2010. The FDR for all SNPs is also given, and for the MB method the FDR was weighted by sequence type to account for the different sequence composition of 2010 and the whole genome. (C) Overlap between genome-wide MB and ML calls. (D) Recovery frequency for SNPs as a function of distance to the nearest polymorphic feature. Analysis based on 2010 sites with sufficient flanking information to assign bin membership. Sample sizes per bin are shown at 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 to sequences tiled on arrays.

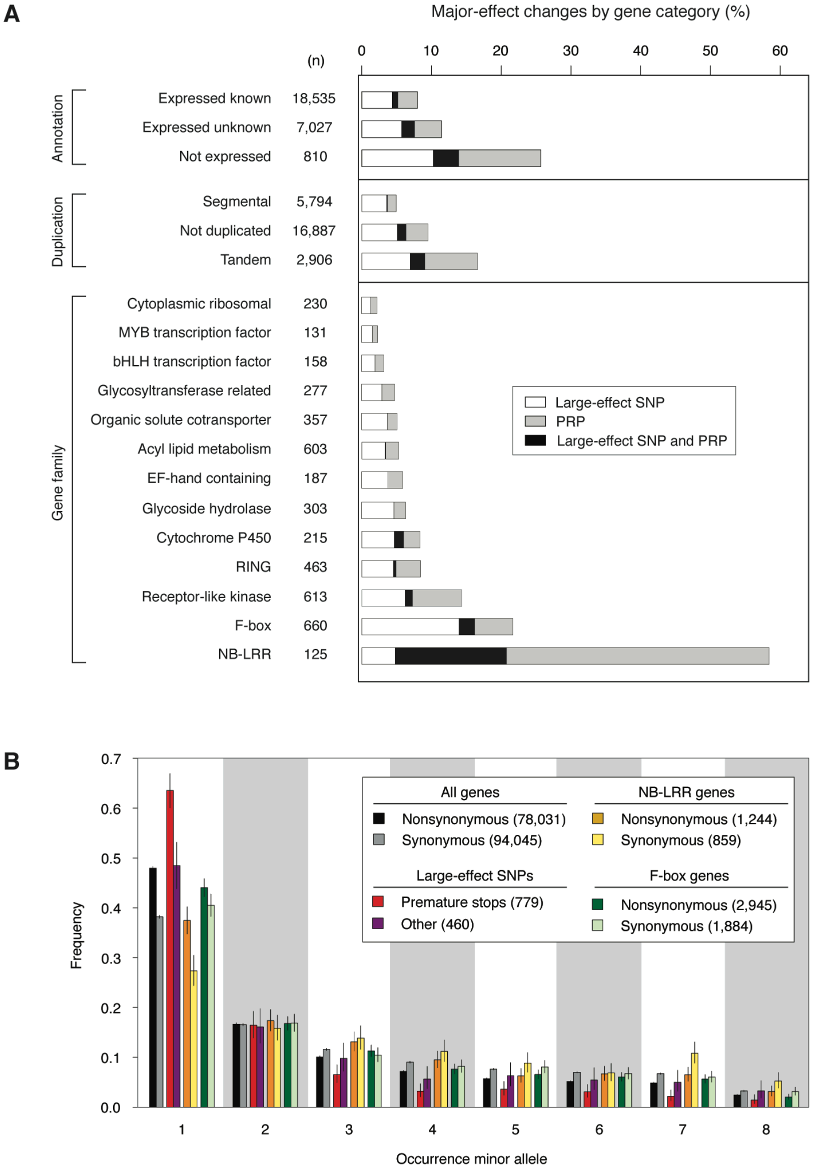


Fig. 3. Distribution of major-effect changes and allele frequencies. (A) Fraction of genes affected by major-effect SNPs or PRPs by gene category. "Large-effect SNP and PRP" denotes a gene harboring both types of polymorphism, either within the same accession or in different accessions. Genes that were entirely masked as repetitive, and for which no SNPs could be predicted, were excluded from analysis. (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.

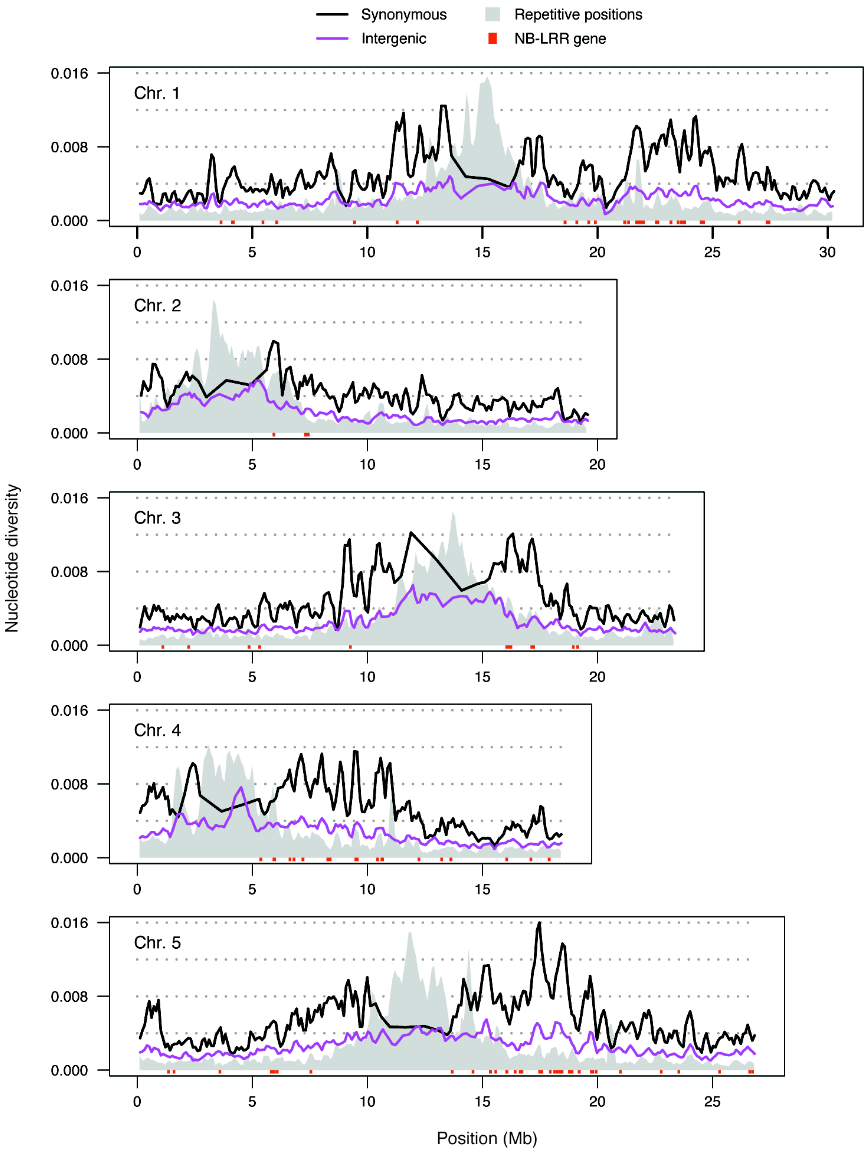


Fig. 4. Genome-wide pattern of nucleotide diversity. Average pairwise nucleotide diversity is plotted for both 4-fold degenerate synonymous sites (black line) and intergenic sites (magenta line) along each chromosome using sliding windows of 5 bins of 50 kb (counted from all sites) with an offset of 2 bins. Grey shading indicates the content of repetitive probes in each window, with the broad peaks on each chromosome corresponding to the centromeric and pericentromeric regions. Repeat content has been rescaled so the maximum possible value is at the top of each plot. The location of NB-LRR genes is denoted above the x-axis. Levels of polymorphism for both 4-fold 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 each bin (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 2A).

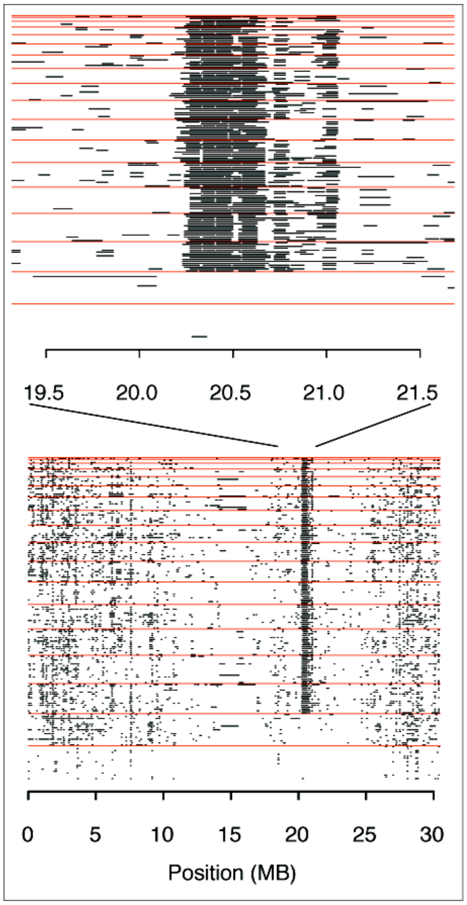


Fig. 5. Regions of extensive pairwise haplotype sharing along chromosome 1. Black lines indicate regions of very high similarity between a pair of accessions (row). The region between 20 and 21 Mb exhibits extensive haplotype sharing over nearly 500 kb in all but two accessions. Accession pairs are sorted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