Large-Scale Identification and Analysis of Genome-Wide Single-Nucleotide Polymorphisms for Mapping in *Arabidopsis thaliana*

Karl J. Schmid,[1] Thomas Rosleff Sörensen,[2] Ralf Stracke,[2] Ottó Törjék,[3] Thomas Altmann,[3] Tom Mitchell-Olds,[1] and Bernd Weisshaar[2,4]

[1]*Max-Planck-Institute of Chemical Ecology, Jena, Germany;* [2]*Max-Planck-Institute for Plant Breeding Research, Cologne, Germany;* [3]*Max-Planck-Institute of Molecular Plant Physiology, Golm, Germany*

Genetic markers such as single nucleotide polymorphisms (SNPs) are essential tools for positional cloning, association, or quantitative trait locus mapping and the determination of genetic relationships between individuals. We identified and characterized a genome-wide set of SNP markers by generating 10,706 expressed sequence tags (ESTs) from cDNA libraries derived from 6 different accessions, and by analysis of 606 sequence tagged sites (STS) from up to 12 accessions of the model flowering plant *Arabidopsis thaliana*. The cDNA libraries for EST sequencing were made from individuals that were stressed by various means to enrich for transcripts from genes expressed under such conditions. SNPs discovered in these sequences may be useful markers for mapping genes involved in interactions with the biotic and abiotic environment. The STS loci are distributed randomly over the genome. By comparison with the Col-0 genome sequence, we identified a total of 8051 SNPs and 637 insertion/deletion polymorphisms (InDel). Analysis of STS-derived SNPs shows that most SNPs are rare, but that it is possible to identify intermediate frequency framework markers that can be used for genetic mapping in many different combinations of accessions. A substantial proportion of SNPs located in ORFs caused a change of the encoded amino acid. A comparison of the density of our SNP markers among accessions in both the EST and STS datasets, revealed that Cvi-0 is the most divergent accession from Col-0 among the 12 accessions studied. All of these markers are freely available via the internet.

[EST sequences longer than 50 bp have been submitted to EMBL/GenBank under accession numbers CB255604–CB265223. STS sequences longer than 50 bp have been submitted to EMBL/GenBank under accession numbers BV007447–BV012320.]

Arabidopsis thaliana is currently the most important model organism for plant research (Meinke et al. 1998) and the first plant whose genome has been nearly completely sequenced (Arabidopsis Genome Initiative [AGI] 2000). With the genome sequence at hand, research goals have shifted toward elucidating the function of all genes of this flowering plant and uncovering how genetic networks act to control complex phenotypic traits using forward and reverse genetic approaches (Somerville and Dangl 2000).

Traditionally, most genetic studies in *A. thaliana* involved mutants, crosses, and mapping populations derived from the genetically distinct Columbia (Col-0) and Landsberg *erecta* (L*er*) accessions (e.g., Lister and Dean 1993). More recent work has shown that different accessions of *A. thaliana* harbor large amounts of heritable genetic variation that can lead to marked phenotypic differences among individuals (for review, see Alonso-Blanco and Koornneef 2000). This naturally occurring genetic variation is an important source for identifying functionally important genes by quantitative trait locus (QTL), linkage disequilibrium (LD), or association mapping approaches (Alonso-Blanco and Koornneef 2000, Steinmetz et al. 2000).

Such genetic mapping approaches and, in addition, the map-based positional cloning of chemically induced mutants (Lukowitz et al. 2000; Jander et al. 2002) require large sets of genetic markers that are polymorphic between individuals or populations. The class of markers that has recently attracted much interest is that of single nucleotide polymorphisms (SNPs). SNPs are abundant in the genome and suitable for high-throughput genotyping (Cho et al. 1999). In *A. thaliana*, more than 37,000 SNPs have already been identified by comparing a partial shotgun sequence from the L*er* accession with the nearly complete genome sequence of the Col-0 accession (www.arabidopsis.org/Cereon/index.html; Jander et al. 2002).

One limitation of SNP markers identified in pairwise comparisons is their unknown frequency in a population. Information on SNP frequency is very useful for two reasons. First, most current SNP genotyping methods require the synthesis of a primer, or a pair of primers, to be used in primer extension or PCR amplification reactions. In contrast to rare SNPs, intermediate-frequency SNPs are polymorphic in many different combinations of accessions. Thus, in large-scale QTL mapping or positional-cloning projects that involve hundreds or thousands of markers or other additional accessions, it will be cost-effective to synthesize primers only for those SNPs

[4]**Corresponding author.**
E-MAIL weisshaa@mpiz-koeln.mpg.de; FAX 49-221-5062 851.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.728603.

13:1250–1257 ©2003 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/03 $5.00; www.genome.org

that are likely to be polymorphic between different combinations of parental lines. Second, if a SNP is to be used as a marker in LD or association mapping projects, its frequency carries important information, because the frequency of a given SNP in the population is correlated with the expected size of a region that is in LD with this marker (for review, see Nordborg and Tavaré 2002). The population structure as well as past selection and demographic events that shape genome-wide patterns of LD and levels of polymorphism can be inferred by analysis of the frequency, LD, and distribution of SNPs among individuals and are important parameters in such mapping projects (Cardon and Bell 2001).

There have been some studies in *A. thaliana* to obtain information on population structure (Bergelson et al. 1998), population history (Miyashita et al. 1998; Sharbel et al. 2000) and large-scale patterns of linkage disequilibrium (Hagenblad and Nordborg 2002; Haubold et al. 2002; Nordborg et al. 2002). However, to our knowledge, no genome-wide marker set is currently available for positional cloning and QTL mapping in nonstandard accessions (i.e., accessions other than Col-0 and L*er*), or for LD and association mapping using *A. thaliana* populations.

Another limitation of currently available SNP sets is that they have not been enriched for markers in genes related to interactions with the environment. Such genes may contribute to quantitative traits and thus may be the main targets of mapping studies directed at uncovering the genetic architecture of naturally occurring phenotypic variation.

Here, we describe a genome-wide set of SNP markers that attempts to overcome these limitations. The markers were generated from more than 10,000 expressed sequence tags (ESTs) from 6 accessions of *A. thaliana* and ~600 sequence-tagged sites (STSs) from 12 accessions. The SNP discovery process was aimed at (1) obtaining a genome-wide set of evenly spaced SNPs, whose frequency in the population is known and, therefore, are useful for both QTL mapping and association studies, and (2) discovering SNPs located in coding regions of genes that are expressed under conditions of environmental stress.

RESULTS

SNPs Derived From ESTs

A total of 10,706 ESTs were generated from cDNA libraries that were derived from 6 different accessions of *A. thaliana* (Table 1). The sequences were subjected to stringent quality filtering, including vector clipping and quality trimming (see Methods), and 7465 high-quality reads were clustered separately for every accession. This resulted in 5289 distinct clusters. Of these, 4240 (80%) consist of only one sequence read (singlets) and 1049 (20%) of clusters with at least 2 reads. The high proportion of singlet reads shows that the libraries are of low redundancy. In the following, the 5289 clusters or singlets will be commonly referred to as clustered ESTs.

We compared all clustered ESTs against the Col-0 genome sequence with BLASTN and subsequently pairwise aligned them with their best hits. A total of 1108 (21%) clustered ESTs were excluded from SNP detection. The reasons were that the BLAST matches (1) were too short (<80 bp, $n = 11$ clustered ESTs), (2) showed a high-sequence divergence (>3%) to the best-matching Col-0 sequence ($n = 76$), or (3) did hit unannotated or incompletely annotated (e.g., no ATG start codon defined) genome regions ($n = 309$). In addition, in 712 cases, the sim4 and/or Wise2 alignments could not be interpreted because of questionable exon-intron structure, incorrect EST assembly, chimeric clones, or incorrect gene prediction. The remaining 4176 clustered ESTs could be mapped to 2907 different annotated genes, which represent about 12% of all currently annotated genes of *A. thaliana* (as of November 2002).

To obtain the number of annotated genes that were for the first time tagged by sequences in our EST set, we analyzed the stringently filtered EST sequences ($n = 7465$). Of these, 574 sequences (479 clustered ESTs) did not match any of the currently annotated genes, and 201 sequences did match a total of 177 genes that had no EST hit before. We concluded that the majority of the genes tagged by the new EST sequences has been tagged before ($n = 3229$), but that there is a significant number of newly tagged genes. Possibly, gene expression profiles are different among the accessions, an option that makes this EST data set useful for gene annotation in *A. thaliana*.

We evaluated the pairwise alignments (clustered EST against genome sequence) for mismatches and were able to identify 4327 SNPs and 18 insertion/deletion polymorphisms (InDels). InDels in noncoding regions were not included because the sim4 alignment program does not have gap penalties and tends to produce inaccurate alignments around InDels. The distribution of SNPs on the five chromosomes shows that the whole genome is well covered by EST-derived SNPs with the exception of the centromeric regions that contain few transcribed genes (AGI 2000).

Due to the low redundancy of the cDNA libraries, 2621 (60%) of SNPs are derived from only one EST sequence. The majority of SNPs are located in coding regions (3432; 79%), and among coding SNPs, a significant proportion (1101 of 3432; 32%) did cause an amino acid replacement. Two SNPs lead to a nonsense codon. Among InDel polymorphisms, 10 are in-frame and 8 out-of-frame.

Because most EST-derived SNPs are derived from only a single sequencing reaction, they need to be considered as hypothetical. To estimate the proportion of false positives among these SNPs, we de-

Table 1. Summary of EST-Derived SNPs

Accession	ESTs	Clustered ESTs	Mapped ESTs[a]	High-quality bp	SNPs	SNP/bp sequenced	InDels[b]
Ak-2	1,248	746	620	311,040	878	1:354	8
C24	2,719	1,193	916	398,384	1,137	1:350	5
Cvi-0	1,774	877	685	322,322	1,345	1:240	4
Ei-2	2,088	944	717	322,698	861	1:386	2
L*er*	1,248	712	590	285,822	840	1:340	5
Nd-1	1,629	817	648	291,330	839	1:347	1
Total	10,706	5,289	4,176	1,941,596	4,327[c]	1:336	18[c]

[a] Mapped ESTs refer to ESTs that could be aligned with annotated genes (see text).
[b] Only InDels in ORFs were counted (see text).
[c] Due to redundancy (e.g., intermediate frequency polymorphisms), the overall number of SNPs and InDels is lower than the sum of all accessions.

signed primers for 96 amplicons covering genomic regions with EST-derived SNPs and used them to generate and sequence PCR products similar to STS generation (see below). From 96 polymorphic sites that were analyzed, 92 were confirmed to match the expected Col-0 sequence from MAtDB, 2 did display a difference, and 2 analyses failed due to PCR or sequencing problems. The two cases with a difference may reflect rare differences in the Col-0 stocks used for genome sequencing (both are from IGF BACs – therefore, our Col-0 stock might be more similar to the TAMU Col-0). The PCR failure rate for the other accessions was higher than for Col-0 (on which the primers were designed), but in 81 cases, data for both the targeted SNP and the Col-0 sequence were available. Of these, only three turned out to be incorrect (96% confirmation rate). In addition, among 1858 SNPs that are located in EST clusters of at least 2 sequence reads, only 8 differ between the individual sequence reads and appear to be sequencing errors or reverse transcriptase-induced mutations. We therefore conclude that due to our stringent quality criteria, a very high proportion of EST-derived SNPs are true polymorphisms.

SNPs Derived From STS Sequences

We designed eight different sets of STS primers to amplify genomic segments of about 600-bp length (see Methods; Tables 2 and 3). In addition to the six accessions selected for SNP generation by cDNA sequencing, five other accessions (Table 2) and Col-0 as control were selected for the STS approach. We tested a total of 606 primer pairs for PCR amplification, from which 595 (98%) yielded a PCR product from at least 1 accession. A total of 88% of all PCR reactions yielded sufficient amounts of DNA for sequencing, suggesting that the failure of PCR reactions was most probably due to sequence variation in the primer-binding sites. We assembled the forward and reverse read separately for each STS and accession, and used the resulting 5098 consensus sequences for comparisons with the Col-0 genome sequence.

Most STS sequences were highly similar to the Col-0 sequence, but 103 consensus sequences showed an overall divergence of >3% and were excluded from SNP detection. The comparison of the remaining 4955 consensus sequences led to the identification of 3773 SNPs and 619 InDel polymorphisms (Table 3). Due to PCR failure or low quality sequence, some SNP positions could not be genotyped in all 12 accessions (Fig. 1A). Among SNPs, 2922 (77%) are located in regions of the genome annotated as noncoding and 869 (23%) in coding regions. Among InDels, 617 were noncoding and only 2 were coding. We were able to determine the coding status of 857 SNPs in coding regions with the Wise2 program and found 410 (48%) replacement SNPs and 447 (52%) silent SNPs. Seven polymorphisms were nonsense mutations that lead to a premature stop codon.

Polymorphisms based on two sequence reads from both strands of the PCR products can be considered to be confirmed and the remaining ones as hypothetical. Using this criterion, 2331 (62%) of the SNPs and 343 (55%) of all InDels are confirmed polymorphisms for at least 1 accession. To test the reliability of our automated SNP detection, we generated 355 STS from the Col-0 accession. A total of 160,708 nonredundant high-quality base pairs was obtained from this accession and 8 (SNPs and InDels) sequence differences to the genome sequence were observed, which leads to an estimated proportion of 0.043% false-positive polymorphisms that likely result either from rare differences in the Col-0 stocks used for genome sequencing, or from errors in our sequence data. Only 1 of these differences had the status of a confirmed polymorphism among a total of 79,362 confirmed base calls from the Col-0 accession (estimated proportion of false-positive SNPs, 0.0012%). This demonstrates that the confirmed SNPs are of high reliability.

To estimate the proportion of SNPs that are rare polymorphisms, we calculated the relative frequencies of all SNPs in the STS sample, whose allelic states have been determined from at least 8 of 12 accessions ($n = 2640$), including the reference genome sequence. Figure 1B shows that most SNPs in our sample are rare and segregate at low frequencies. A total of $n = 1344$ (51%) SNPs with a sampling depth of at least 8 accessions occur as singletons.

Table 2. Summary of STS-Derived SNPs

Accession	Sequence count[a]	High-quality bp	SNPs	InDels	SNPs/bp[b]	% divergence to Col-0[c]
C24	472	212,757	1,006	158	1:212	0.734
CS22491	452	213,646	988	171	1:216	0.749
Cvi-0	492	213,234	1,186	217	1:180	0.895
Ei-2	440	186,619	722	125	1:259	0.588
Gü-0	470	218,942	727	138	1:301	0.542
L*er*	486	216,760	1,057	184	1:206	0.751
Lz-0	462	211,138	925	157	1:228	0.660
Nd-1	473	214,844	829	157	1:259	0.641
Wei-0	465	214,110	860	147	1:250	0.602
Ws-0	299	133,947	568	115	1:236	0.684
Yo-0	444	200,321	751	138	1:267	0.635
Total	4,995	2,236,318	3,773[d]	619[d]	n.d.[e]	0.680

[a] Consensus sequences assembled from individual sequence reads (both directions from PCR fragments were sequenced).
[b] SNPs from each accession in comparison to Col-0 are evaluated, which counts alleles identical in non-Col-0 accessions independently for each accession.
[c] Average percent divergence calculated from pairwise alignment with the Col-0 genome sequence.
[d] Due to redundancy of alleles, the overall number of SNPs and InDels is lower than the sum of SNPs from all accessions.
[e] (n.d.) Not determined.

Divergence Relationships Among Accessions

A comparison of the average sequence divergence to Col-0 among accessions indicates that the Cvi-0 accession is substantially more divergent from Col-0 than the other accessions in both the EST and STS data sets (Tables 1 and 3). Concerning the EST data, the number of base pairs that have to be analyzed to detect a SNP varies >1.5-fold between the most (Cvi-0) and the least divergent (Ei-2) accession. In the STS data, the corresponding difference in the divergence of Cvi-0 and Gü-0 to Col-0 is 1.6-fold. The overall level of divergence is ~1.5-fold higher in the STS than in the EST data set, because the former contains a higher proportion of noncoding genomic regions.

Table 3. Summary of STS-Derived SNPs by STS Set

STS set	Count	% successful amplifications	Genes tagged	High-quality bp	SNPs	SNPs/bp sequenced	InDels
CAPS	122	85	62	427,841	653	1:655	131
TIGR	89	96	53	320,807	609	1:527	69
CYS	56	85	26	185,588	254	1:731	67
CHR4	48	90	24	177,132	449	1:395	43
Je-1MB	100	88	67	414,536	637	1:651	125
Go-1MB	111	n/a	55	477,476	829	1:576	134
ATHA_EST	48	97	43	208,490	309	1:675	47
BDRUM_EST	32	47	7	18,080	33	1:548	3
Total	606	88	337	2,236,318	3,773	1:595	619

A distance tree of all 12 accessions based on STS-derived SNPs that have been genotyped in all accessions confirms that Cvi-0 is the most divergent and Gü-0 the least divergent accession to Col-0 (Fig. 2A). A consensus tree resulting from a bootstrap analysis of the same data shows that the topology is essentially identical to the distance-based tree, but individual nodes (except the node connecting Col-0 with Gü-0) are only weakly supported, and Cvi-0 no longer appears to be the most divergent accession (Fig. 2B), suggesting that the overall large genetic distance between Cvi-0 and Col-0 is due to a limited number of more (still <3%) divergent loci.

By use of the STS-derived SNPs, it is also possible to calculate the proportion of the observed genetic variation that segregates among the widely used Col-0, L*er*, and Cvi-0 accessions (e.g., Lister and Dean 1993; Alonso-Blanco et al. 1998a). Of 620 SNPs, 311 (50%) are polymorphic among these 3 accessions. If SNPs are partitioned into singlets (lower frequency polymorphisms) and multiplets (higher frequency polymorphisms), then 121 (37%) of 328 singlets are polymorphic among Col-0, L*er*, and Cvi-0, and 190 (65%) of 292 multiplets are polymorphic among these 3 accessions.

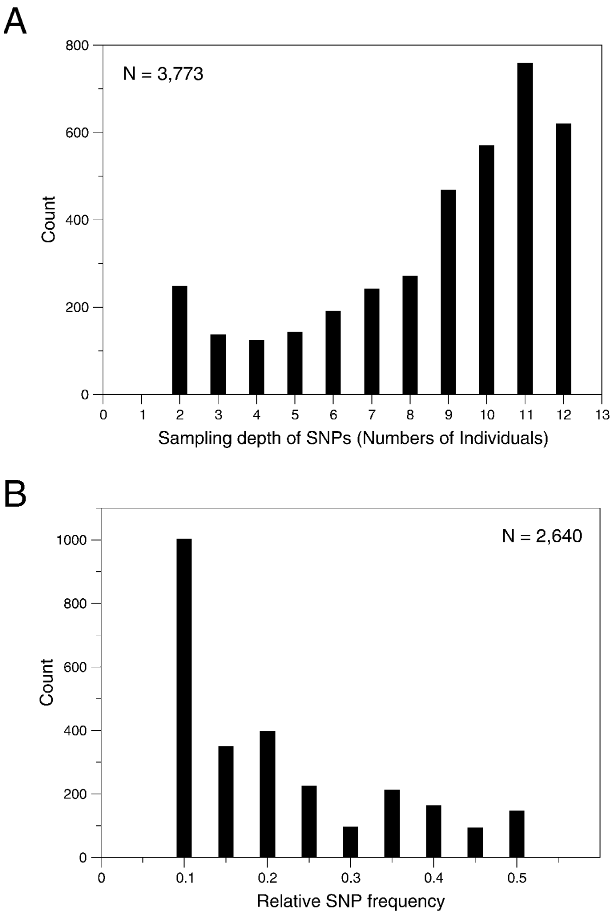

Figure 1 (*A*) Sampling depth of SNPs derived from STS sequences. Sampling depth is defined as the number of different accessions from which the SNP was sequenced. (*B*) Frequency distribution of SNPs with a sampling depth of eight or more accessions.

GABI-MASC SNP Database

To provide public access to markers generated in this study, a Web interface was created that can be accessed under www.mpiz-koeln.mpg.de/masc/. It is possible to retrieve SNP markers on the basis of information on genomic location, hits to protein coding genes, or differences between two or more accessions that were surveyed in this study. The Web interface provides additional information such as restriction sites, upstream and downstream sequences, and primer sequences that will aid the application of these markers in large-scale mapping experiments by use of various genotyping methods (e.g., CAPS, primer extension, Pyrosequencing). The information can be used by researchers without restrictions.

DISCUSSION

The availability of the genome sequence greatly facilitates the discovery of genetic polymorphisms in model species, because the sequencing of additional individuals and comparison with genome sequence can be automated and performed on a large scale. We used EST and STS sequence data to identify a total of 8051 unique SNP and 637 InDel polymorphisms, and thus significantly increased the number of available genetic markers for *A. thaliana*. The inclusion of new accessions into the SNP discovery process and the genotyping of SNPs up to 12 accessions provides genome-wide sets of SNPs for many pairwise combinations. These sets can be used for QTL mapping and as SNPs of different frequencies and types in association and LD mapping projects. The use of cDNA libraries from stressed individuals should enrich our marker collection for genes that interact with the environment and control quantitative traits.

Overall, the data confirm the observation made in earlier genome-wide and multilocus surveys of genetic variation (Bergelson et al. 1998; Miyashita et al. 1999; Arabidopsis Genome Initiative 2000; Sharbel et al. 2000) that, despite its selfing nature, a substantial amount of genetic variation segregates in *A. thaliana*, which is useful in genetic mapping experiments and in studies of the evolutionary history and population structure of this model organism. The mean sequence divergence to Col-0 calculated from our data of 0.68% is well within the range observed in sequence surveys of other loci in *A. thaliana* (Kuittinen and Aguadé 2000).

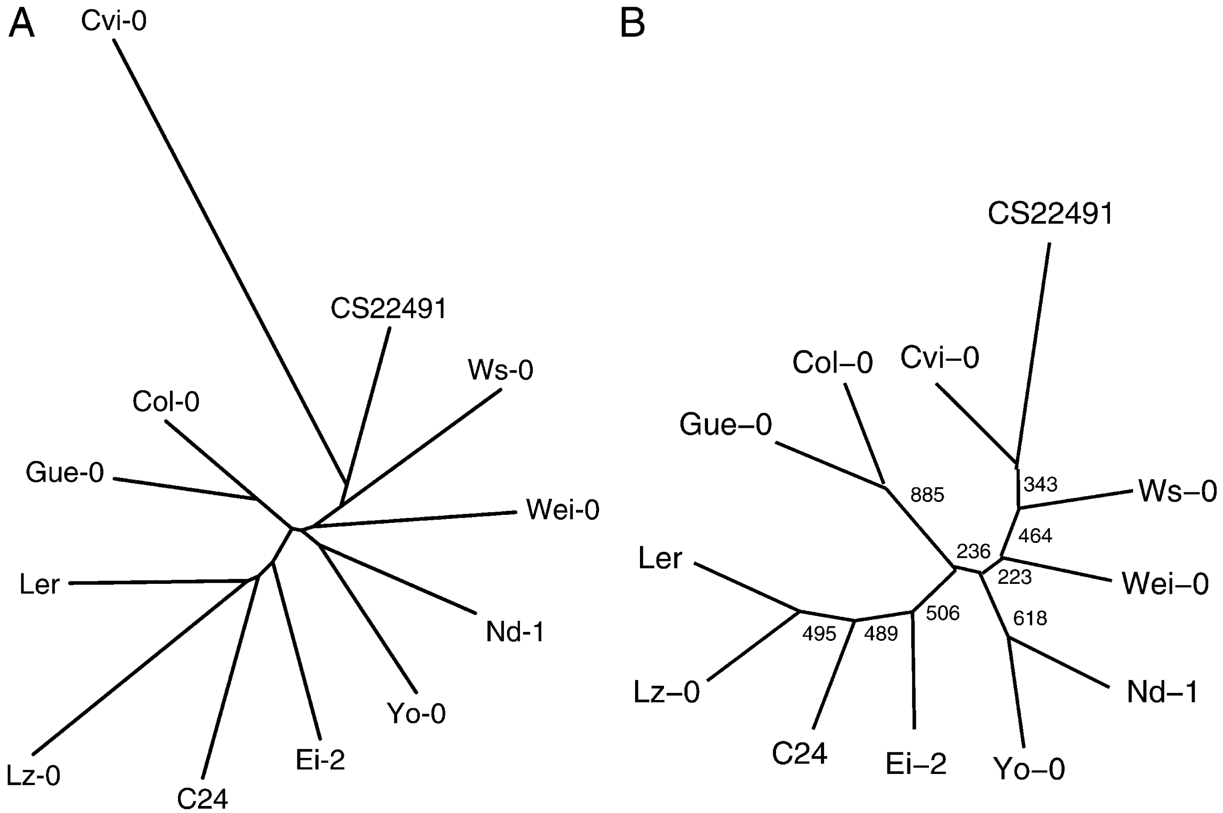

Figure 2 Genealogy of 12 accessions of *A. thaliana* accessions based on SNPs. (*A*) Neighbor-joining (NJ) tree based on sequence between accessions. (*B*) Consensus NJ tree of 1000 bootstrap simulations. Numbers indicate how often a given node was supported in the bootstrap sampling.

The EST and STS datasets can be considered to be independent samples of genome sequence diversity. The main differences between both sets of SNPs are (1) a larger number of base pairs that are necessary to find a SNP in EST than in STS sequence data, and (2) a higher ratio of silent to replacement polymorphisms in EST (~2:1) than in STS data (~1:1), suggesting that the genes tagged by the ESTs are more constrained in their sequence evolution than the essentially randomly chosen loci of the STS set. This notion is supported by a lower average frequency of replacement than silent or noncoding polymorphisms in STS (data not shown). Among the different STS sets, the ratio of silent to replacement polymorphisms varies considerably (Table 3) and reflects differences in evolutionary constraints among protein-coding genes.

Our data support the existence of a certain degree of population structure in *A. thaliana*, for which evidence was also found in a genome-wide survey of AFLP markers (Sharbel et al. 2000) but not in other studies (e.g., Bergelson et al. 1998). Among all accessions studied, Gü-0 is the most similar and Cvi-0 the most divergent accession in a comparison with Col-0. Although it has been noted before that the Cvi-0 accession may be genetically divergent from other accessions (Alonso-Blanco et al. 1998b), more stringent analyses of multilocus data did not support this notion (Breyne et al. 1999; Miyashita et al. 1999; Sharbel et al. 2000). This is consistent with the result of our bootstrap analysis (Fig. 2B), which suggests that the overall distance of Cvi-0 is similar to other accessions, but that the observed total distance is probably due to a few, highly divergent loci. It will be interesting to further investigate whether the high divergence of these loci in the Cvi-0 accession is due to adaptive evolution and contributes to the phenotypic divergence of Cvi-0 that has been noted before.

Our SNP markers are of immediate use for QTL mapping and positional cloning experiments with the accessions used for this study. We have extracted several subsets of ~100 evenly spaced framework markers of intermediate frequency and with an average spacing of ~1.3 Mb for more than 6 combinations of the Col-0, C24, Nd-1, Cvi-0, and L*er* accessions. These SNPs are currently used for genotyping of novel RILs and NILs made from these accessions and for determining the relationship of ~300 accessions that are currently kept at stock centers.

A long-term goal of this project is to apply SNP markers as tools for association and LD mapping approaches to identify genes that underlie naturally occurring genetic variation. Population genetics theory predicts that in a predominantly selfing organism like *A. thaliana*, large genomic regions can be expected to be in LD, suggesting that a limited set of markers may be sufficient for fine-scale mapping of interesting genes (Nordborg 2000). However, before being able to carry out such projects, both practical and statistical issues need to be resolved (for review, see Cardon and Bell 2001; Tabor et al. 2002). These include, for example, the investigation of the underlying population structure in the species, the classification of markers by their functional effects, and the prioritization of markers most suitable for population-based mapping. We are further developing bioinformatics tools for a high-throughput and automated characterization of genetic polymorphisms to be able to carry out such analyses. By using the SNPs identified in this study, we have now started a detailed investigation of evolutionary forces that may have affected the observed patterns of genotypic variation in the markers described here.

METHODS

Selection of *A. thaliana* Accessions

A total of 12 accessions was chosen for the survey. These include six accessions used previously for genetic mapping (Col-0, Cvi-0, L*er*, Nd-1, C24, Ws-0) and six additional accessions (Ei-2, CS22491, Gü-0, Lz-0, Wei-0, Yo-0) with a high-average genetic distance to other accessions as determined from AFLP data (Sharbel et al. 2000). All accessions are available from stock centers.

Construction of cDNA Libraries

Accessions Cvi-0, L*er*, Nd-1, and Ei-2 were grown under long-day conditions in soil. After 6 weeks (10 weeks in the case of Ei-2), whole adult plants of each ecotype were treated for 24 h with different stresses (8 plants per stress condition) as follows: (1) at 4°C in the dark, (2) at 37°C in the dark, (3) lying in the laboratory after removing from the soil, (4) in the greenhouse after wounding leaves with a foreceps, (5) in the laboratory by watering with a 150-mM NaCl solution, (6) at 26°C in UV-containing white light. Equal quantities of stressed plant material were pooled and ground in a mortar. A cDNA library was made from inflorescences collected at different stages from the Ak-2 accession. Total RNA from 1 g of plant material was isolated using the RNAeasy Maxi Kit (QIAGEN), and mRNA was purified using the mRNA purification kit (Amersham Biosciences). mRNA was reverse transcribed into cDNA using a *Not*I primer-adapter (5′-pGACTAG TAGTTCTAGATCGCGAGCGGCCGCCC(T)18VN-3′) and the Superscript Plasmid System for cDNA Synthesis and Plasmid Cloning (Invitrogen). After *Sal*I adapter (primers: 5′-TC GACCCACGCGTCCG-3′ and 5′-pCGGACGCGTGGG-3′) ligation, *Not*I digestion, and cDNA size fraction, fragments >500 bp were ligated directionally into the vector pSPORT1. The cDNA-libraries were electrotransformed into *Escherichia coli* TOP10 cells (Invitrogen). A pre-existing λ-ZAPII library derived from seedlings of accession C24 (provided by Dieter Berger, MPT-MP) was mass excised, and *E. coli* XL1-Blue MRF′ cells (Stratagene) were infected with the phagemids. After plating, single clones of each cDNA library were picked into 384-well microtiter plates filled with dYT/freezing-buffer/Ampicillin medium using a Q-PIX robot (Genetix). Libraries were replicated using a BioGrid robot (BioRobotics).

A total of 9696 ordered library clones were used for EST generation (number of clones subjected to EST sequencing per accession were as follows: C24, 2688; Ei-2, and Nd-1, 1632; L*er*, Cvi-0, and Ak-2, 1248; in most cases, the first plate contained only 96 clones). All cDNA clones are available from the RZPD (see www.rzpd.de). The number of clones analyzed is smaller than the number of EST sequence reads produced because some cDNAs were sequenced not only from the 5′-end but also from the 3′-end.

EST Sequencing

cDNA inserts were PCR amplified using M13uni (5′-CGTAA AACGACGGCCAGT-3′) and M13rev (5′-GGAAACAGCTAT GACCATG-3′) primers at 0.2 µM final concentration in a reaction containing 1.5 mM $MgCl_2$, 0.4 mM each dNTP, 20 mM Tris/HCl (pH 8.3), 50 mM KCl, 2 U Taq DNA-Polymerase (Invitrogen), and 1 µL bacteria culture as template. PCR reactions were carried out in PTC-225 (MJ Research) or Geneamp9600 (Applied Biosystems) thermocyclers using the following program: 5 min at 95°C, 35 cycles of 30 sec at 94°C, 45 sec at 58°C, 150 sec at 72°C, and a final extension for 5 min at 72°C. PCR products were purified over Sephadex G50 Superfine (Amersham Biosciences) using MultiScreen 96-well filtration plates (Millipore). DNA sequences were determined on an AbiPrism3700 sequencer (Applied Biosystem) using BigDye terminator chemistry. The sequencing primers were T7R (5′-CTAATACGACTCACTATAGGGA-3′) and SP6r (5′-ATTT AGGTGACACTATAGAAGA-3′). EST sequences longer than 50 bp have been submitted to EMBL/GenBank under accession numbers CB255604–CB265223.

STS Selection

A total of 606 primer pairs were designed and tested for amplification in all 12 accessions. Eight sets of STS regions were chosen by the following criteria (Table 3). (1) Publicly available CAPS loci (www.arabidopsis.org/aboutcaps.html) (CAPS set, $n = 122$). (2) Shotgun contigs derived from genomic sequences from the L*er* accession (available from www.tigr.org/tdb/at/atgenome/Ler.html) were aligned with the Col-0 genomic sequence. STS loci were chosen from the pairwise alignment if the alignment was at least 100 bp in length, was different by at least one nucleotide, had a sequence divergence of <5%, was single copy in the Col-0 Sequence, and did not contain microsatellites or any other repetitive elements (TIGR set, $n = 89$). (3) A set of randomly chosen Cys-rich genes identified by a pattern match of conceptual translations of the genome sequence (CYS set, $n = 56$). (4) A random set of regions from the top and bottom of chromosome 4 that are separated by a distance of ~100 kb (CHR4 set, $n = 48$). (5) A random set of regions from the whole genome separated by a distance of ~1 Mb (Je-1MB set, $n = 100$). (6) A random set of regions from the whole genome separated by a distance of ~1 Mb and optimized for intergenic regions (Go-1MB set, $n = 111$). (7) A set derived from polymorphic EST sequences. Loci were chosen to close larger gaps not covered by the other primer sets (ATHA_EST set, $n = 48$). (8) A random set derived from EST sequences that are conserved between *A. thaliana* and the related species *Boechera drummondii* (BDRUM_EST set, $n = 32$).

With the exception of CAPS primers, primer pairs were designed automatically with the Primer3 program (S. Rozen and H.J. Skaletsky, unpubl.) by use of the Col-0 genome sequence as a template. Parameters were chosen such that the primers had an annealing temperature of 60°C and resulted in amplicons of ~600 bp, so that the resulting PCR fragments could be sequenced completely from both directions using the forward and reverse primers.

Genomic DNA Preparation, STS PCR Amplification, and Sequencing

Plants were grown for 6–10 wk in a greenhouse under long-day conditions. DNA was prepared from 2 g of young leaves with Genomic-tips 100/G columns (QIAGEN). PCR reactions were carried out in 20 µL containing 5 ng of genomic DNA, 1 µM each of forward and reverse primers, 0.4 mM each of dNTP, 1.5 mM $MgCl_2$, 1 U Taq Polymerase (QIAGEN) in a Geneamp9600 Thermocycler (Applied Biosystems). Reactions were performed as follows: 2 min at 94°C, followed by 35 cycles of 30 sec at 94°C, 30 sec at 59°C, 1 min at 72°C, and a final extension for 19 min at 72°C. PCR products were purified and sequenced as described above. PCR was used to amplify and sequence genomic regions from 12 accessions. An opportunistic approach to SNP detection was taken, that is, failed PCR experiments and sequencing reactions were not repeated. STS sequences longer than 50 bp have been submitted to EMBL/GenBank under accession numbers BV007447–BV012320.

Sequence Analysis

We established an automated sequence analysis pipeline consisting of a database, publicly available programs, and a set of Python scripts that link the database with the analysis programs. Base calling was carried out using Phred (Ewing et al. 1998). The raw EST sequence data were trimmed for vector and adaptor sequences (using cross_match; P. Green, unpubl.), and poly(A) or poly(T) stretches were clipped. Slippage reads were removed as described (Telles and da Silva 2001). The sequences were quality trimmed (sliding window of 50 bp with a minimal average Phred score of 20), and filtered for a minimum length of 80 bp. In addition, the sequences were filtered for similarity to bacterial, mitochondrial, and chloroplast sequences, and for matches to duplicated genes (e.g., photosynthesis-related genes) for which the true paralog could not be identified. STS sequences were clustered with Phrap separately for every accession and STS locus, whereas the filtered EST sequences were clustered separately for every accession. After clustering, bases in the consensus sequences that had a Phred score <30 and the neighboring bases a score <20 (see Altshuler et al. 2000) were masked. The consensus sequence was then compared with the Col-0 genome sequence using BLASTN (NCBI Blast version 2.0) to identify the homologous sequences in the Col-0 accession.

SNP Detection

SNPs were detected by finding sequence differences in pairwise alignments with the Col-0 genome sequence. Only sequences with at least 100 (for STS data) and 80 (for EST data) high-quality bases and <3% sequence divergence to the reference sequence were included in the SNP detection to exclude potentially paralogous sequences. EST sequences were aligned with the program sim4 (Florea et al. 1998), and STS sequences with cross_match. To differentiate between silent and replacement polymorphisms, protein-coding genes tagged by the clustered EST or STS sequences were identified by obtaining the coordinates of pairwise alignments with the best matching BAC clone sequence and retrieving the annotated genes covered by the alignment from the MIPS annotation (version 171102; available at ftp://ftpmips.gsf.de/cress; Schoof et al. 2002). The conceptual protein sequence of covered genes was then aligned to the matching clustered EST or STS sequences with the Wise2 program (Birney et al. 1996). The automated analysis of the Wise2 output and the classification of polymorphisms into silent and replacement polymorphisms was done with Python scripts. If more than one nucleotide substitution occurred in a codon, classification was done using the parsimony approach as described in the manual of the DnaSP program (Rozas and Rozas 1999).

Polymorphisms were mapped physically onto the pseudochromosome and given a unique identifier (MASC number) after merging the EST and STS datasets. SNPs that were sequenced from one side only were scored as hypothetical and SNPs sequenced from both directions in the same accession (STS only) as confirmed. All SNP markers generated by this project are available from www.mpiz-koeln.mpg.de/masc/. Data will also be available from TAIR.

Phylogenetic Analysis

To calculate distance relationships between accessions, all SNPs that were sequenced from all 12 individuals were extracted from the database and combined into a single alignment. The subsequent analyses were carried out with the PHYLIP package (Version 3.6; Felsenstein 1989). Calculation of DNA distance was done with the program dnadist using Kimura correction, bootstrap sampling with *bootdist*, calculation of the phylogenetic trees with neighbor (using the neighbor-joining algorithm), and the consensus tree was generated with consense using a majority rule.

ACKNOWLEDGMENTS

We thank Henriette Ringys-Beckstein and the DNA core facility teams at MPI-Z and MPI-CE for excellent technical assistance, and Jürgen Kroymann and Dan Kliebenstein for primer design and discussions. Many thanks to Svenja Meyer (GABI Primary Database at RZPD) for handling the data submission to EMBL/GenBank, to Heiko Schoof (GABI-Info, MIPS) for providing MAtDB annotation data, and to Dieter Berger who provided the C24 cDNA library. This work was funded by an Emmy-Noether-Fellowship of the Deutsche Forschungsgemeinschaft to K.J.S., by grants of the BMBF-GABI project to T.A., T.M.-O., and B.W., and by the Max-Planck-Society. Partial support was also provided by grant number QLG2-CT-2001-01097 NATURaL from the European Union.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Alonso-Blanco, C. and Koornneef, M. 2000. Naturally occurring variation in *Arabidopsis*: An underexploited resource for plant genetics. *Trends Plant Sci.* **5:** 22–29.

Alonso-Blanco, C., El-Assal, S.E.-D., Coupland, G., and Koornneef, M. 1998a. Analysis of natural allelic variation at flowering time loci in the Landsberg *erecta* and Cape Verde islands ecotypes of *Arabidopsis thaliana*. *Genetics* **149:** 749–764.

Alonso-Blanco, C., Peeters, A., Koornneef, M., Lister, C., Dean, C., van den Bosch, N., Pot, J., and Kuiper, M. 1998b. Development of an AFLP based linkage map of L*er*, Col and Cvi *Arabidopsis thaliana* ecotypes and construction of a L*er*/Cvi recombinant inbred line population. *Plant J.* **14:** 259–271.

The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.

Altshuler, D., Pollara, V., Van Etten, C.C.W., Baldwin, J., Linton, L., and Lander, E. 2000. An SNP map of the human genome generated by reduced representation sequencing. *Nature* **407:** 513–516.

Bergelson, J., Stahl, E., Dudeck, S., and Kreitman, M. 1998. Genetic variation between and within populations of *Arabidopsis thaliana*. *Genetics* **148:** 1311–1323.

Birney, E., Thompson, J., and Gibson, T. 1996. PairWise and SearchWise: Finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucleic Acids. Res.* **24:** 2730–2739.

Breyne, P., Rombaut, D., Van Gysel, A., Van Montagnu, M., and Gerats, T. 1999. AFLP analysis of genetic diversity within and between *Arabidopsis thaliana* ecotypes. *Mol. Gen. Genet.* **261:** 627–634.

Cardon, L. and Bell, J. 2001. Association study designs for complex diseases. *Nat. Rev. Genet.* **2:** 91–98.

Cho, R., Mindrinos, M., Richards, D., Sapolsky, R., Anderson, M., Drenkard, E., Dewdney, J., Reuber, T., Stammers, M., Federspiel, N., et al. 1999. Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. *Nat. Genet.* **23:** 203–207.

Ewing, B., Hillier, L., Wendl, M., and Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8:** 175–185.

Felsenstein, J. 1989. PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics* **5:** 164–166.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G., and Miller, W. 1998. A computer program for aligning a cDNA sequence with a genomic sequence. *Genome Res.* **8:** 967–974.

Hagenblad, J. and Nordborg, M. 2002. Sequence variation and haplotype structure surrounding the flowering time locus *FRI* in *Arabidopsis thaliana*. *Genetics* **161:** 289–298.

Haubold, B., Kroymann, J., Ratzka, A., Mitchell-Olds, T., and Wiehe, T. 2002. Recombination and gene conversion in a 170 kb genomic region of *Arabidopsis thaliana*. *Genetics* **161:** 1269–1278.

Jander, G., Norris, S., Rounsley, S., Bush, D., Levin, I., and Last, R.

2002. *Arabidopsis* map-based cloning in the post-genome era. *Plant Phys.* **129:** 440–450.

Kuittinen, H. and Aguadé, M. 2000. Nucleotide variation at the *CHALCONE ISOMERASE* locus in *Arabidopsis thaliana*. *Genetics* **155:** 863–872.

Lister, C. and Dean, C. 1993. Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. *Plant J.* **4:** 745–750.

Lukowitz, W., Gillmor, C., and Scheible, W.-R. 2000. Positional cloning in *Arabidopsis*. Why it feels good to have a genome initiative working for you. *Plant Phys.* **123:** 795–805.

Meinke, D.W., Cherry, J.M., Dean, C., Rounsley, S.D., and Koornneef, M. 1998. *Arabidopsis thaliana*: A model plant for genome analysis. *Science* **282:** 662–682.

Miyashita, N.T., Kawabe, A., Innan, H., and Terauchi, R. 1998. Intra- and interspecific DNA variation and codon bias of the alcohol dehydrogenase (*Adh*) locus in *Arabis* and *Arabidopsis* species. *Mol. Biol. Evol.* **15:** 1420–1429.

Miyashita, N.T., Kawabe, A., and Innan, H. 1999. DNA variation in the wild plant *Arabidopsis thaliana* revealed by amplified random fragment length polymorphism analysis. *Genetics* **152:** 1723–1731.

Nordborg, M. 2000. Linkage disequilibrium, gene trees and selfing: An ancestral recombination graph with partial self-fertilization. *Genetics* **154:** 923–929.

Nordborg, M. and Tavaré, S. 2002. Linkage diseuqilibrium: What history has to tell us. *Trends Genet.* **18:** 83–90.

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

Rozas, J. and Rozas, R. 1999. DnaSP version 3: An integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15:** 174–175.

Schoof, H., Zaccaria, P., Gundlach, H., Lemcke, K., Rudd, S., Kolesov, G., Arnold, R., Mewes, H.W., and Mayer, K.F. 2002. MIPS *Arabidopsis thaliana* Database (MAtDB): An integrated biological knowledge resource based on the first complete plant genome. *Nucleic Acids Res.* **30:** 91–93.

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

Somerville, C. and Dangl, J. 2000. Genomics—Plant biology in 2010. *Science* **290:** 2077–2078.

Steinmetz, L., Mindrinos, M., and Oefner, P. 2000. Combining genome sequences and new technologies for dissecting the genetics of complex phenotypes. *Trends Plant Sci.* **5:** 397–401.

Tabor, H., Risch, N., and Myers, R. 2002. Candidate-gene approaches for studying complex genetic traits: Practical considerations. *Nat. Genet. Rev.* **3:** 1–7.

Telles, G. and da Silva, F. 2001. Trimming and clustering sugarcane ESTs. *Gen. Mol. Biol.* **24:** 17–23.

WEB SITE REFERENCES

ftp://ftpmips.gsf.de/cress; MIPS *Arabidopsis thaliana* annotation files (download).

http://www.mpiz-koeln.mpg.de/masc/; GABI-MASC SNP database.

www.arabidopsis.org/aboutcaps.html; *Arabidopsis* CAPS marker table.

www.arabidopsis.org/Cereon/index.html; Cereon *Arabidopsis* polymorphism and L*er* sequence collection.

www.rzpd.de; German Resource center/Primary Database for genomic research.

www.tigr.org/tdb/at/atgenome/Ler.html; L*er* Sequence Database.

Received August 22, 2002; accepted in revised form March 19, 2003.