GENOMICS 19, 137–144 (1994)

Assignment of 30 Microsatellite Loci to the Linkage Map of *Arabidopsis*

CALLUM J. BELL AND JOSEPH R. ECKER[1]

Plant Science Institute, Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6018

Received July 6, 1993; revised September 22, 1993

Thirty microsatellite loci were assigned to the *Arabidopsis* linkage map. Several microsatellite sequences in *Arabidopsis* DNA were found by searching the EMBL and GenBank databases, and a number of these were subsequently found to detect polymorphisms between different *Arabidopsis* strains by the polymerase chain reaction (PCR). After the presence of microsatellites in *Arabidopsis* and their utility for genetic mapping had been demonstrated, systematic screening for $(CA)_n$ and $(GA)_n$ sequences was carried out on marker-selected plasmid libraries and a small-insert genomic library. Positive clones were sequenced, PCR primers flanking the repeats were synthesized, and PCR was carried out on different strains to look for useful polymorphisms. Surprisingly, of 18 $(CA)_n$ repeats ($n > 13$), only one was polymorphic. In contrast, 25 of 30 $(GA)_n$ repeats, 2 of 3 $(AT)_n$ repeats, and 2 of 4 $(A)_n$ repeats were polymorphic. The majority of the $(CA)_n$ repeats were complex, with adjacent short di-, tri-, or tetranucleotide repeats, whereas most of the $(GA)_n$, $(TA)_n$, and $(A)_n$ repeats were simple. The $(CA)_n$ repeats were also refractory to PCR analysis, requiring extensive optimization of PCR conditions, whereas the other repeat classes were mostly amplified with a single set of standard conditions. When polymorphisms were detected, the microsatellites were mapped using a set of recombinant inbred lines originating from a cross between the strains Columbia and Landsberg *erecta*.

INTRODUCTION

Genetic mapping in mammals has undergone a transformation since the discovery of simple sequence length polymorphisms (SSLPs) (Weber and May, 1989; Litt and Luty, 1989; Tautz, 1989) and their exploitation as linkage markers (Hearne *et al.*, 1992; NIH/CEPH Collaborative Mapping Group, 1992; Dietrich *et al.*, 1992). The many benefits of SSLPs should apply equally to plant studies, where there is also a need for abundant, highly informative, randomly distributed markers that can be assayed by the polymerase chain reaction (PCR) and distributed between laboratories as primer sequences. The adoption of *Arabidopsis thaliana* as a model system for plant genetics and molecular biology makes it desirable to have a dense linkage map of broad utility for this organism. A linkage map of SSLPs would have three obvious uses in *Arabidopsis*.

The first of these uses would be the rapid mapping of mutations, which is currently carried out using classical markers (Koornneef, 1990), restriction fragment length polymorphisms (RFLPs) (Chang *et al.*, 1988; Nam *et al.*, 1989), or random amplified polymorphic DNAs (RAPDs) (Reiter *et al.*, 1992). Classical markers are simple to use and require no use of molecular biology but can suffer from ambiguous scoring and interference between the marker phenotype and the phenotype to be mapped. In addition, only a few markers can be reliably followed in a single cross, meaning that many crosses have to be made to arrive at a location for the gene of interest. RAPDs are easily generated, simple to score, and amenable to automation, but they are generally dominant in nature, meaning that they have limited use in the F2 or backcross populations that are commonly used for mapping. For these reasons, RFLPs are more commonly used. SSLPs represent a considerable methodological advance over RFLPs in that mapping can be accomplished with small preparations of DNA made from single seedlings or leaf pieces, and polymorphisms are visualized by electrophoresis rather than blotting and hybridization. DNA preparation, PCR, analysis of the amplification products, and determination of map position can be accomplished in 2 days. Codominant cleaved amplified polymorphic sequences (CAPS; Konieczny and Ausubel, 1993) are a logical extension of RFLPs that use PCR technology but their generation requires the prior existence of an RFLP and the complete sequence of the RFLP probe. For these reasons, they have so far been limited to cloned genes.

The second use of an SSLP map would be as an ordered set of sequence-tagged sites (STSs; Olson *et al.*, 1989) for construction of a physical map by STS content mapping (Green and Olson, 1990). The assembly of yeast artificial chromosomes (YACs) into contiguous physical maps can be complicated by false positive and

[1] To whom correspondence should be addressed. Telephone: (215) 898-9384. Fax: (215) 898-8780. E-mail: jecker@atgenome.bio.upenn.edu.

0888-7543/94 $6.00

negative results, by the chimeric nature of some YACs, and by STSs that detect sequences at multiple locations in the genome. Prior knowledge of the relative order of the STSs provides a means of detecting some of these errors.

Finally, the existence of multiple alleles and probable selective neutrality make these ideal markers for population and evolutionary studies.

Microsatellite repeat sequences have been found in several plant species. Beckmann and Soller (1989) showed the existence in potato of $(AT)_n$, $(GA)_n$, and $(CG)_n$ repeats in the sequence databases. In addition, an estimate of the abundance of $(CA)_n$ and $(GA)_n$ repeats in corn and in four species of tropical trees has been made (Condit and Hubbell, 1991). $(AT)_n$ and $(ATT)_n$ repeats have been shown to be present in soybean and polymorphic between different strains and also were the first microsatellite loci in a plant species to be mapped (Akkaya *et al.*, 1992). Lagercrantz *et al.* (1993) and Morgante and Olivieri (1993) estimated the frequency of microsatellites in the sequence databases, showing that these elements are less frequent in plant genomes than in mammals, with, on average, one repeat longer than 20 bp every 29 kb, compared to a figure of 6 kb in mammals (Beckmann and Weber, 1992). The most abundant plant microsatellite was found to be $(A)_n$, followed by $(AT)_n$ and then $(GA)_n$, with $(CA)_n$ repeats being relatively scarce compared to mammalian genomes.

In this study, we investigate the utility of microsatellites as tools for genetic mapping in *A. thaliana*, assign 30 microsatellites of various types to the linkage map, and provide polymorphism data for these 30 repeats in six strains.

MATERIALS AND METHODS

Database search. To identify microsatellites in previously sequenced *Arabidopsis* DNA, the GenBank (release 76.0) and EMBL (release 23.0) nucleic acid databases were searched using 20-nucleotide queries corresponding to all possible di- and mononucleotides. Searches were carried out on a Sun SPARC2 workstation by FASTA (Pearson and Lipman, 1988) under the GCG package (Genetics Computer Group, 1991) and by regular expressions as part of DNA WorkBench, an interactive DNA and protein analysis program (Tisdall, 1993).

Construction of a plasmid library. Five micrograms of genomic DNA of the Columbia strain was digested to completion with *Alu*I, *Rsa*I, *Taq*I, and *Eco*RV in 1× KGB (potassium glutamate buffer; Sambrook *et al.*, 1989) and ends of the resulting fragments were rendered blunt by treatment with the Klenow fragment of *Escherichia coli* DNA polymerase I. After phenol/chloroform extraction and ethanol precipitation, the DNA was separated on 2% agarose, and the 200- to 500-bp fraction (representing 15–25% of the genome) was purified using GlasPac (National Scientific). In two subsequent steps, cohesive *Not*I/*Eco*RI adaptors were ligated to the sized fragments and the 5′ ends were phosphorylated by T4 polynucleotide kinase. The DNA was separated from excess adaptors by chromatography through a Sephacryl S-300 cDNA spun column (Pharmacia) according to the manufacturer's protocol. To remove any remaining adaptors, the DNA was run out for a short distance into a 2% agarose gel and purified using GlasPac. The inserts were ligated to *Eco*RI-treated and dephosphorylated pBluescript KS+, and portions of the ligation reactions were introduced into *E. coli* strain CJ236 (*dut*-1, *ung*-1, *thi*-1, *rel*A1; pCJ105 (Cm^r)) by electroporation and plated on LB plates containing ampicillin. Approximately 100,000 colonies were pooled, suspended in 10 ml LB broth containing 7% dimethyl sulfoxide (DMSO), frozen in 200-μl aliquots in liquid nitrogen, and stored at −80°C.

Construction of marker-selected libraries. $(CA)_n$ and $(GA)_n$ marker-selected libraries were constructed essentially according to Ostrander *et al.* (1992) as follows. Single-stranded phage were prepared by inoculating 2 ml of 2× YT broth (Sambrook *et al.*, 1989) containing ampicillin with 1 μl of the pooled library bacteria from above, superinfecting with the helper phage VCSM13, and selecting for infected bacteria by kanamycin selection during overnight incubation. The uracil-containing single-stranded DNA (ssDNA) was purified from culture supernatant by standard methods (Sambrook *et al.*, 1989). Approximately 500 ng of uracil-containing ssDNA was mixed with 5 pmol of the phosphorylated oligonucleotide $(CT)_{10}$ or $(GT)_{10}$ in a 100-μl reaction mixture of 1× *Taq* polymerase buffer (Promega) containing 1.5 m*M* $MgCl_2$ and 200 μ*M* deoxyribonucleotides. This mixture was heated to 95°C for 5 min, cooled to 60°C for 2 min (during which 1 unit of *Taq* polymerase was added), and then incubated at 72°C for 30 min. After phenol/chloroform extraction, ethanol precipitation, and drying, the DNA was taken up in 50 μl of 1× ligation buffer (Promega) containing 1 m*M* ATP and 1 unit of T4 DNA ligase and incubated for 2 h at room temperature to repair the single-strand nicks remaining after the primer extension. The DNA was concentrated by ethanol precipitation and resuspended in water, and aliquots were electroporated into *E. coli* strain DH5α (*supE*44, Δ*lacU*169 (ϕ80 *lacZ*ΔM15), *hsdR*17, *recA*1, *endA*1, *gyrA*96, *thi*-1, *rel*A1) to generate libraries enriched for clones containing $(CA)_n$ and $(GA)_n$ repeats.

Construction of a small-insert λZapII library. To generate a library fully representative of the genome that combined the efficiency of bacteriophage λ cloning and the convenience of plasmids with small inserts, DNA that was randomly digested with DNase was cloned into λZapII as follows: genomic DNA (10 μg) was partially digested with DNase I in the presence of 10 m*M* manganese chloride. After repair of the ends with T4 DNA polymerase, the DNA was run out on a 2% agarose gel, and the 300- to 700-bp fraction was cut out. Purification of the size-selected DNA and ligation of adaptors were as described above. The DNA was ligated to dephosphorylated λZapII vector arms and the ligation was packaged using Gigapack Gold packaging extract (Stratagene). Phage (2×10^6) were amplified by plating on *E. coli* strain LE392 (*supE*44, *supF*58, *hsdR*514, *galK*2, *galT*22, *metB*1, *trpR*55, *lacY*1) and eluting in SM buffer (Sambrook *et al.*, 1989). As determined by PCR of random clones using T7 and T3 primers, the library contains 70% recombinants with inserts averaging 500 bp.

Hybridization screening for $(CA)_n$ and $(GA)_n$ microsatellites. The marker-selected plasmid and λZapII libraries were screened by colony and plaque hybridization (Sambrook *et al.*, 1989), respectively, using random hexamer-labeled poly(dA·dC)/poly(dG·dT), and poly(dA·dG)/poly(dC·dT) as probes (Feinberg and Vogelstein, 1983). Prehybridization of nitrocellulose (Schleicher and Schuell) or nylon (Magna, Micron Separations Inc.) filters was done in 7% sodium dodecyl sulfate (SDS), 0.5 *M* sodium phosphate, pH 7.2, 1% BSA (Sigma) overnight at 60°C. Hybridization was done overnight in the same solution containing $1–2 \times 10^6$ cpm/ml of probe. The filters were washed in 2× SSPE, 0.5% SDS (Sambrook *et al.*, 1989), once for 20 min at room temperature and twice for 30 min each at 55°C, and positive plaques or colonies were identified by autoradiography. For ZapII phage clones, pBluescript plasmids were recovered by *in vivo* excision using the Stratagene Exassist/SOLR system. Miniprep plasmid DNA was sequenced using modified T7 DNA polymerase (Sequenase version 2) and autoradiography or with an Applied Biosystems 373A instrument.

Plant material. Genomic DNA of various *Arabidopsis* strains was prepared according to Ausubel *et al.* (1987) from bulked plant material or from leaf pieces or individual seedlings by the method of Edwards *et al.* (1991).

Polymerase chain reaction and polymorphism determination. PCR primers flanking microsatellite repeat sequences were selected using the PRIMER program (Eric Lander, Whitehead Institute) and either synthesized in house on an Applied Biosystems 380B or purchased from Research Genetics Inc. (Huntsville, AL). Microsatellites were amplified from genomic DNA in 20-μl reactions containing 1–10 ng

TABLE 1

Mono- and Dinucleotide Repeats Greater Than 20 Nucleotides Long in Previously Cloned *Arabidopsis* DNA

Locus	Accession No.	Repeat	Reference
ATHCHIB	M38240	$(AT)_{14}$	Samac *et al.*, 1990
ATEAT1	X66719	$(AT)_{11}$	Gomez-Lim *et al.*[a]
S45384S1	S45384	$(AT)_{44}$	Wilkinson and Crawford, 1991
ATHATPC1	M61741	$(AT)_{32}$	Inohara *et al.*, 1991
ATATSG	X14565	$(AT)_3AA(TA)_{10}(GT)_5$	Krebbers *et al.*, 1988
ATCRB	X14313	$(AT)_{10}$	Pang *et al.*, 1988
ATHCTR1A	L08789	$(AG)_{16}$	Kieber *et al.*, 1993
ATHATPASE	J04185	$(AG)_5GG(AG)_3GG(AG)_3GG(AG)_3$	Manolson *et al.*, 1988
ATGBF3	X63896	$(AG)_{11}$	Schindler *et al.*, 1992
ATHMYBO	M79448	$(TC)_5(CA)_8$	Oppenheimer *et al.*, 1991
ATHACS	M95594	$(A)_{36}$	Liang *et al.*, 1992
ATHGENEA	M21021	$(A)_{39}$	Simoens *et al.*, 1988
ATHPRECA	M58381	$(A)_{22}G(A)_6$	Intapruk *et al.*, 1991

[a] Gomez-Lim, M. A., Valdez-Lopez, V. M., and Saucedo-Arias, L. J. (1992). Isolation and characterization of an ethylene-related gene from *Arabidopsis thaliana*. Unpublished results.

genomic DNA, 5 pmol of each primer, 200 μM deoxyribonucleotides, 50 mM KCl, 10 mM Tris–Cl, pH 9, 0.01% gelatin, 0.1% Triton X-100, and 2 units of *Taq* polymerase. The final concentration of magnesium chloride was usually 2 mM, but was varied for some primer pairs. The DNA in a 10-μl volume of water was heated to 100°C for 5 min along with a 12-μl pellet of paraffin wax and then cooled to room temperature. After the wax had solidified over the DNA, the remaining reagents were added in a 10-μl volume, and the reaction was heated to 94°C for 3 min to melt the wax, providing a hot start. Standard cycling conditions were 94°C for 15 s, 55°C for 15 s, and 72°C for 30 s, repeated 40 times. The annealing temperature was modified for some primer pairs as described in the results. Amplification was done in a Perkin–Elmer–Cetus 480 or in a Bios Biosycler oven. Length variation between PCR products from different strains was assessed by analyzing 4 μl of PCR reactions on 4% agarose gels. When no polymorphisms were detected in this way, one of the primers was 5′ end-labeled with [γ^{32}P]ATP using T4 polynucleotide kinase and the radioactive PCR products were analyzed by 6% denaturing polyacrylamide gel electrophoresis followed by autoradiography.

Linkage mapping. A set of recombinant inbred strains derived from a cross between Columbia (Col-0) and Landsberg *erecta* (Ler) was used for mapping (Lister and Dean, 1993). These strains are F_8 by single-seed descent and so are expected to be greater than 99% homozygous. Primer pairs detecting polymorphisms between Ler and Col-0 were used to amplify genomic DNA from subsets of 48 or 96 of the recombinant inbreds, and each strain was scored for the parental alleles. The data were entered into the program RI Plant Manager 2.4 (Manly, 1993), which assigned linkage positions for the microsatellites in relation to an existing set of approximately 60 RFLP markers (Lister and Dean, 1993) Two-, three-, and multipoint linkage analyses were carried out using the program MAPMAKER 3.0 (Lander *et al.*, 1987; Lincoln *et al.*, 1992) running on a Sun SPARC2 workstation.

RESULTS

Microsatellite Sequences in Previously Cloned DNA

Searches of the GenBank and EMBL databases revealed 13 *Arabidopsis* entries with mono- or dinucleotide repeats greater than 20 nucleotides long. The locus identifications, accession numbers, and repeating units are shown in Table 1. The most common motif is $(AT)_n$ with seven entries, followed by $(AG)_n$ and $(A)_n$ with three entries each and $(CA)_n$ with one entry. PCR primers flanking the repeats were synthesized for all of these, with the exception of ATCRB, and ATGBF3, and ATHMYBO. Genomic DNA of the Columbia strain was successfully amplified using all 10 of the primer pairs tested; however, the results for ATATSG were inconsistent and this locus was not studied further. In the cases of ATHATPC1 and S45384S1, a Landsberg allele could not be amplified even after attempts were made to optimize the PCR conditions. In theory, these loci could be mapped as dominant markers, but in the absence of an internal control, lack of amplification cannot unequivocally be taken to mean a true negative result, so no attempt was made to map these loci. Of the remaining seven microsatellites, all but ATHPRECA were found to be polymorphic between Columbia and Landsberg *erecta*, permitting assignment of a linkage position to these loci using a set of recombinant inbred lines derived from a cross between these two strains (Lister and Dean, 1993).

Isolation of $(CA)_n$ and $(GA)_n$ Containing Plasmid and Lambda Clones

The marker selection procedure provided approximately 10-fold enrichment for $(CA)_n$ and $(GA)_n$ containing plasmid clones, as estimated from the frequency of positive hybridization signals in the primary plasmid library and in the marker-selected libraries. This level of enrichment was sufficient to make large-scale isolation

TABLE 2

Summary of Microsatellite Classes Studied with Polymorphism Data between the Columbia and Landsberg *erecta* Strains

Class	Total	No amplification[a]	Polymorphic	%[b]
GA	37	7	25	83
CA	22	3	1	5
AT	6	3	2	67
A	4	0	2	50

[a] No amplification or amplification in one strain only.

[b] Percentage of clones giving amplification in both Columbia and Landsberg strains that were polymorphic.

TABLE 3

Primer Sequences

Locus	Forward primer	Reverse primer
ATHCHIB	CTCATATATACAAAGAACTACTATAC	ATGAGAAGCTATAATTTTTTCAATA
ATEAT1	GCCACTGCGTGAATGATATG	CGAACAGCCAACATTAATTCCC
ATHACS	AGAAGTTTAGACAGGTAC	AAATGTGCAATTGCCTTC
ATHGENEA	ACCATGCATAGCTTAAACTTCTTG	ACATAACCACAAATAGGGGTGC
ca72:	AATCCCAGTAACCAAACACACA	CCCAGTCTAACCACGACCAC
ATHATPASE	CTGGGAACGGTTCGATTCGAGC	GTTCACAGAGAGACTCATAAACCA
ATHCTR1A	TATCAACAGAAACGCACCGAG	CCACTTGTTTCTCTCTCTAG
nga6	TGGATTTCTTCCTCTCTTCAC	ATGGAGAAGCTTACACTGATC
nga8	GAGGGCAAATCTTTATTTCGG	TGGCTTTCGTTTATAAACATCC
nga12	AATGTTGTCCTCCCCTCCTC	TGATGCTCTCTGAAACAAGAGC
nga32	GGAGACTTTTTGAGATTGGCC	CCAAAACAATTAGCTCCCCA
nga59	GCATCTGTGTTCACTCGCC	TTAATACATTAGCCCAGACCCG
nga63	AACCAAGGCACAGAAGCG	ACCCAAGTGATCGCCACC
nga76	GGAGAAAATGTCACTCTCCACC	AGGCATGGGAGACATTTACG
nga106	GTTATGGAGTTTCTAGGGCACG	TGCCCCATTTTGTTCTTCTC
nga111	CTCCAGTTGGAAGCTAAAGGG	TGTTTTTTAGGACAAATGGCG
nga112	TAATCACGTGTATGCAGCTGC	CTCTCCACCTCCTCCAGTACC
nga126	GAAAAAACGCTACTTTCGTGG	CAAGAGCAATATCAAGAGCAGC
nga128	GGTCTGTTGATGTCGTAAGTCG	ATCTTGAAACCTTTAGGGAGGG
nga129	TCAGGAGGAACTAAAGTGAGGG	CACACTGAAGATGGTCTTGAGG
nga139	AGAGCTACCAGATCCGATGG	GGTTTCGTTTCACTATCCAGG
nga151	GTTTTGGGAAGTTTTGCTGG	CAGTCTAAAAGCGAGAGTATGATG
nga158	TCATTTTGGCCGACTTAGC	ACCTGAACCATCCTCCGTC
nga162	CATGCAATTTGCATCTGAGG	CTCTGTCACTCTTTTCCTCTGG
nga168	TCGTCTACTGCACTGCCG	GAGGACATGTATAGGAGCCTCG
nga172	AGCTGCTTCCTTATAGCGTCC	CATCCGAATGCCATTGTTC
nga225	GAAATCCAAATCCCAGAGAGG	TCTCCCCACTAGTTTTGTGTCC
nga248	TACCGAACCAAAACACAAAGG	TCTGTATCTCGGTGAATTCTCC
nga249	TACCGTCAATTTCATCGCC	GGATCCCTAACTGTAAAATCCC
nga280	CTGATCTCACGGACAATAGTGC	GGCTCCATAAAAAGTGCACC

of these clones straightforward; however, the enrichment was also accompanied by bias in the distribution of clones in the marker-selected libraries. Sequencing of 79 $(CA)_n$-containing independently picked clones revealed only 34 unique sequences, and several of these were sequenced four, five, or six times. A smaller sample from the $(GA)_n$ marker-selected library was examined, but a similar pattern was noted. Since the enrichment by the marker selection procedure was only modest and accompanied by considerable redundant sequencing, the small insert λZapII library was used as the source of the majority of microsatellites.

PCR Amplification and Polymorphism Determination

After false-positive clones and those containing microsatellites less than 20 nucleotides long were discarded, primers for 22 $(CA)_n$, 6 $(AT)_n$, 4 $(A)_n$ and 37 $(GA)_n$ sequences were selected. Amplification was initially carried out on genomic DNA from Columbia and Landsberg under the standard PCR conditions and analyzing the products on 4% agarose gels to check for amplification and the presence of polymorphisms. In cases where agarose gels revealed no polymorphism, the PCR was repeated with one of the primers end-labeled with ^{32}P, and the products were analyzed on denaturing 6% polyacrylamide gels to check for polymorphisms. When amplification was seen in only one of the strains or not at all, the PCR conditions were varied by altering the annealing temperature and/or the magnesium concentration in an effort to determine optimum conditions.

The first set of clones to be studied contained $(CA)_n$ repeats that, almost without exception, were very difficult to amplify, requiring extensive optimization of the PCR conditions. In 3 of 22 cases, no amplification could be achieved, while in the remaining 18, multiple amplification products were mostly obtained. The number of bands was reduced to two or three after the conditions were optimized, but only in 3 cases was a single band obtained. Of the 18 primer pairs that gave amplification, only 1 detected a polymorphism between Columbia and Landsberg. Of five $(AT)_n$ sequences studied, ATEAT1 and ATHCHIB detected polymorphisms between Columbia and Landsberg, whereas S45384S1 and ATHATPC1 were amplified only from Columbia DNA, and amplification of ATATSG was unreliable. Of four $(A)_n$ sequences studied, all were successfully amplified, with those in ATHACS and ATHGENEA being polymorphic, while those in ATHPRECA and nga78 (a clone originally identified as putatively containing a $(GA)_n$ tract) were not.

As the $(CA)_n$ class of repeats was mostly uninformative, the $(GA)_n$ class was examined in more detail, and 37 clones were studied. Of these, 7 were unamplifiable, 5 were amplifiable but nonpolymorphic between Columbia and Landsberg, and the remaining 25 were polymorphic. These results are summarized in Table 2.

The primer pairs that detected polymorphisms are

TABLE 4

Allele Sizes of PCR Products Amplified from Six Strains for 30 Microsatellites

Locus	Repeat	Col-0	Ler	Ws-0	No-0	Nd-0	RLD	No. of alleles
ATHCHIB	$(AT)_{14}$	84	74	82	72	66	90	6
ATEAT1	$(AT)_{11}$	172	162	162	164	164	162	3
ATHACS	$(A)_{36}$	259	256	262	259	259	259	3
ATHGENEA	$(A)_{39}$	209	205	211	221	213	217	6
ca72[a]	$(CA)_{18}$	124	110	110	106	106	106	3
ATHATPASE	$(AG)_{18}$	85	69	69	69	69	77	3
ATHCTR1[b]	$(AG)_{16}$	159	143	145	143	147	143	4
nga6	$(GA)_{31}$	143	123	131	147	123	133	5
nga8	$(GA)_{27}$	154	198	166	168	188	160	6
nga12	$(GA)_{16}$	247	234	247	n.a.	232	n.a.	3
nga32	$(GA)_{13}$	260	256	260	250	260	252	4
nga59	$(CT)_{19}$	111	115	83	141	111	111	4
nga63	$(GA)_{23}$	111	89	91	89	91	89	3
nga76	$(GA)_{22}$	231	>250	199	n.a.	203	n.a.	4
nga106	$(GA)_{26}$	157	123	123	131	123	123	3
nga111	$(GA)_{16}$	128	162	146	140	128	130	5
nga112[c]	$(GA)_{16}$	197	189	189	189	n.a.	189	2
nga126	$(AG)_{31}$	119	147	119	131	149	103	5
nga128	$(AG)_{16}$	180	190	172	180	186	188	5
nga129	$(GA)_{20}$	177	179	165	165	165	165	3
nga139	$(AG)_{29}$	174	132	132	132	182	136	4
nga151	$(CT)_{31}$	150	120	102	150	110	120	4
nga158	$(GA)_{13}$	108	104	120	106	112	124	6
nga162	$(GA)_{21}$	107	89	85	87	97	91	6
nga168	$(GA)_{25}$	151	135	135	135	135	135	2
nga172	$(GA)_{29}$	162	136	138	162	164	180	5
nga225	$(CT)_{18}$	119	189	119	123	131	97	5
nga248	$(CT)_{24}$	143	129	133	125	133	135/115	6
nga249	$(TC)_{15}$	125	115	115	115	135	115	3
nga280	$(AG)_{15}$	105	85	85	85	85	85	2

[a] Annealing temperature is 61°C.
[b] Annealing temperature is 56°C.
[c] Magnesium concentration is 3 m*M*.

shown in Table 3. These were used to amplify genomic DNA from six commonly used laboratory strains using the end-labeled PCR primer method. The amplification products were separated on 6% denaturing polyacrylamide gels and their sizes estimated by comparison with a sequencing ladder. Table 4 shows PCR product sizes for these six strains, amplified using primer pairs for all 30 microsatellites. In the great majority of cases, amplification was successful with all six strains, failing only five times, twice each with Niederzanz and RLD and once with Nossen. In one case, nga248 amplified from RLD, two alleles were detected; all other loci were homozygous in all strains. The number of alleles detected ranged from 2 to 6 with a mean of 4. The primer pairs flanking the $(AT)_n$ repeat in the basic chitinase gene intron (ATHCHIB) were used to assess polymorphisms across a larger sample of 20 strains. The results, shown in Fig. 1, show 12 alleles in 19 samples that were amplified. One strain, Ei-5, was heterozygous at this locus.

The abundance of $(CA)_n$ and $(GA)_n$ sequences in the genome was estimated by plaque hybridization. Discounting the 30% nonrecombinants in the 1× amplified ZapII library, $(CA)_n$- and $(GA)_n$-containing clones were detected at frequencies of 1 in 860 and 1 in 488, respectively. With an average insert size of 500 bp, this indicates that these sequences are found, on average, every 430 and 244 kb, respectively.

Linkage Mapping

Each of the primer pairs in Table 3 was used to amplify DNA from 48 or, in some cases, 96 recombinant inbred strains derived from a Columbia × Landsberg *erecta* cross (Lister and Dean, 1993). When the size difference between the Landsberg and Columbia alleles was large, the amplification products were analyzed on 4% agarose gels; otherwise, 6% denaturing polyacrylamide was used. Figure 2 shows an example of one such experiment with segregation of Landsberg and Columbia alleles of microsatellite nga172 in 46 RI lines.

The strain distribution patterns were analyzed using the program RI Plant Manager 2.4 (Manly, 1993) to make initial linkage assignments relative to a set of approximately 60 RFLP markers (Lister and Dean, 1993). Two-, three-, and multipoint linkage analyses were then carried out using the program MAPMAKER 3.0 (Lander *et al.*, 1987; Lincoln *et al.*, 1992). All of the microsatellite markers were found to be linked to at least 2 other markers at greater than LOD 3.0, by two-point analysis, and were unequivocally assigned to a single

chromosome. Multipoint analysis established single linkage groups for chromosomes 2, 3, 4, and 5 plus two linkage groups on chromosome 1. The maximum likelihood position of marker GAP-B is between the two chromosome 1 linkage groups, in agreement with Lister and Dean (1993). Figure 3 shows the maximum-likelihood linkage maps of all five chromosomes.

DISCUSSION

The first 30 polymorphic microsatellite loci have been assigned to the *Arabidopsis* linkage map. Of these, 6 were obtained from previously cloned sequences and the remainder were obtained by screening genomic DNA libraries. The most abundant class of microsatellites longer than 20 nucleotides found by database searching consisted of $(AT)_n$ repeats, 7 of which were detected. $(GA)_n$ and $(A)_n$ repeats were approximately half as abundant with 3 each, whereas only 1 $(CA)_n$ repeat was found. Prevalence of $(AT)_n$ repeats seems to be a general feature of plant genomes (Lagercrantz *et al.*, 1993; Morgante and Olivieri, 1993), as does relative paucity of $(CA)_n$ repeats, which are the most common dinucleotides in mammalian DNA. The frequencies of $(CA)_n$ and $(GA)_n$ repeats in the *Arabidopsis* genome were estimated by plaque hybridization to be 1 every 430 and 244 kb, respectively. The reported frequencies of these repeats in a number of other higher plant species range from 1 every 86–300 kb for $(CA)_n$ to 1 every 17–125 kb for $(GA)_n$ (Condit and Hubbell, 1991; Lagercrantz *et al.*, 1993), making the *Arabidopsis* genome the least rich in these repeats. It is likely that our estimates of $(CA)_n$ and $(GA)_n$ repeat frequencies are low, since they were made from fairly stringent hybridization experiments that excluded most repeats of $n < 15$ from detection. Also, an amplified library was used for the analysis, which raises the possibility of a biased distribution of clones. A more confident assessment of their abundance, therefore, will require estimation from unamplified libraries and/or reconstruction experiments. The greater abundance of $(GA)_n$ repeats than $(CA)_n$ repeats appears to be a consistent feature of plant genomes (Lagercrantz *et al.*, 1993).

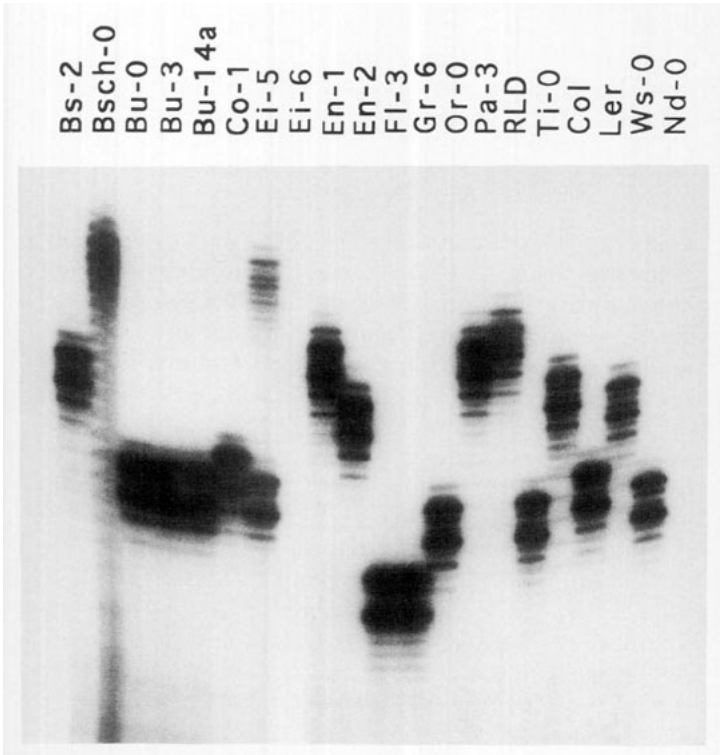

FIG. 1. Amplification of the $(AT)_n$ sequence in the intron of the gene encoding basic chitinase (ATHCHIB) from 20 strains of *Arabidopsis thaliana*.

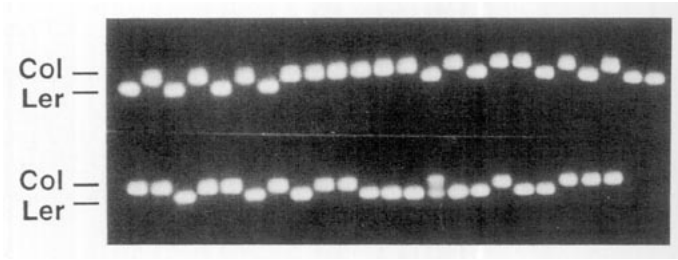

FIG. 2. Amplification of locus nga172 from a subset of 46 recombinant inbred strains. Each lane contains PCR products amplified from genomic DNA of one recombinant inbred line. The samples are arranged in two rows corresponding to two gels. Plant 38 is heterozygous at this locus. Ler and Col indicate the Landsberg *erecta* and Columbia alleles, respectively.

Attempts to use poly(AT) as a hybridization probe were largely unsuccessful, probably due to the self-complementarity of this sequence and also to the high background resulting from the low-stringency conditions used to accommodate the instability of the AT basepairs. No estimates of $(AT)_n$ or $(A)_n$ microsatellite frequency were made, but given the frequent occurrence of these sequences in database entries of plant DNA, they may represent a large untapped pool of polymorphisms.

Initial efforts were inspired by the success of $(CA)_n$ repeats as polymorphic markers in mammalian studies; therefore, the discovery that these sequences are very conserved in length between the Columbia and Landsberg strains of *Arabidopsis* was very surprising. Interestingly, lack of polymorphism was correlated with complex repeat structure and difficult PCR amplification. All but three of the $(CA)_n$ elements studied are compound repeats, with short di-, tri-, or, in a few cases, tetranucleotide repeats adjacent to the major run of $(CA)_n$. The majority of $(CA)_n$ also required extensive optimization of the PCR conditions, requiring annealing temperatures of 60–64°C. The optimum conditions for each primer pair also differed between strains, meaning that comparison of allele sizes across a range of strains was unfeasible.

In contrast to $(CA)_n$, $(GA)_n$ repeats were found to be highly polymorphic. Eighty-three percent of primer pairs giving amplification with both Landsberg and Columbia strains also detected a polymorphism between them. Unlike the $(CA)_n$ repeats, the $(GA)_n$ class was without exception simple in structure and mostly amplified with a single set of conditions, requiring no optimization. Why repeat class, complexity of structure, and ease of amplification should be correlated with polymorphism is unclear.

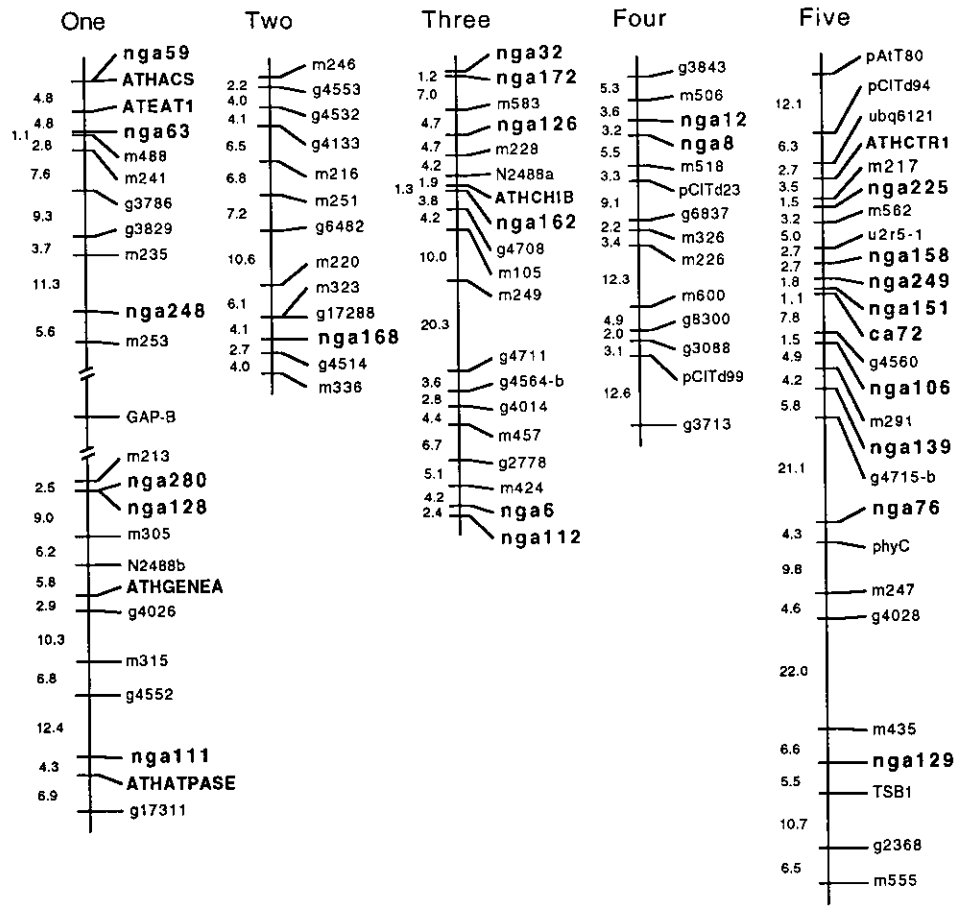

FIG. 3. Maximum-likelihood linkage map of *Arabidopsis* based on the linkage data generated by Lister and Dean (1993) and this study. The microsatellite loci assigned in this study are in boldface.

Amplification of DNA from six common strains showed that the microsatellites in this study are highly polymorphic. There was no obvious correlation of polymorphism information content with repeat length up to 50 nucleotides (nt); some of the shorter repeats (26–32 nt) had 4–6 alleles, while some longer repeats (n = 46–50 nt) had only 2–3 alleles. However, repeats longer than 52 nucleotides had a mean of 5 alleles, and none had fewer than 4 alleles. The mean number of alleles for all markers was 4. These results indicate that randomly selected microsatellites are likely to be informative in any given mapping population and will be especially useful for studying the evolutionary relationships between the many strains of *A. thaliana.*

Using the set of recombinant inbred lines developed by Lister and Dean (1993), all 30 markers were assigned unequivocally to one chromosome, and linkage for each was established to neighboring markers at greater than LOD 3.0. Mapping the microsatellites was straightforward since in most cases the Columbia/Landberg polymorphism could be confidently scored using agarose gel electrophoresis, and the 48 PCR reactions normally used for mapping could be accommodated in one gel. Ten, one, seven, two, and ten markers were assigned to the five chromosomes, respectively. Given the size of the sample, this distribution appears biased against chromosomes 2 and 4, but more markers will need to be mapped before this can be stated unequivocally. Also, since $(GA)_n$ clones were the majority in this study, very little information on the chromosomal distribution of $(AT)_n$ and $(A)_n$ clones is available.

In this study, we have determined that mono- and dinucleotide simple sequence length polymorphisms are present in *Arabidopsis,* estimated their abundance, and demonstrated a high probability of finding polymorphisms between different strains. Thirty new markers were assigned to the *Arabidopsis* linkage map. It is clear that these markers will be very useful for both linkage mapping of mutations and population studies, due to their high rate of polymorphism and distribution among the chromosomes. They may also have potential use as a dense STS set for the construction of a physical map of the *Arabidopsis* genome (Ewens *et al.,* 1991).

ACKNOWLEDGMENTS

We thank Clare Lister and Caroline Dean for providing the recombinant inbred lines and linkage information prior to publication, Athanasios Theologis for primers flanking the $(A)_{36}$ repeat in ATHACS, and James Tisdall for the use of DNA WorkBench. This work was supported in part by a grant from the University of Pennsylvania Research Foundation.

REFERENCES

Akkaya, M. S., Bhagwat, A. A., and Cregan, P. B. (1992). Length polymorphisms of simple sequence DNA in soybean. *Genetics* **132:** 1131–1139.

Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. J., Smith, J. A., and Struhl, K. (1987). "Current Protocols in Molecular Biology," Greene Publishing Associates/Wiley Interscience, New York.

Beckmann, J. S., and Soller, M. (1989). Toward a unified approach to genetic mapping of eukaryotes based on sequence tagged microsatellite sites. *Biotechnology* **8:** 930–932.

Beckmann, J. S., and Weber, J. L. (1992). Survey of human and rat microsatellites. *Genomics* **12:** 627–631.

Chang, C., Bowman, J. L., Dejohn, A. W., Lander, E. S., and Meyerowitz, E. M. (1988). Restriction fragment length polymorphism map for *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **85:** 6856–6860.

Condit, R., and Hubbell, S. P. (1991). Abundance and DNA sequence of two-base repeat regions in tropical tree genomes. *Genome* **34:** 66–72.

Dietrich, W., Katz, H., Lincoln, S. E., Shin, H., Friedman, J., Dracopoli, N. C., and Lander, E. S. (1992). A genetic map of the mouse suitable for typing intraspecific crosses. *Genetics* **131:** 423–447.

Edwards, K., Johnstone, C., and Thompson, C. (1991). A simple and rapid method for the preparation of plant genomic DNA for PCR analysis. *Nucleic Acids Res.* **19:** 1349.

Ewens, W. J., Bell, C. J., Donnelly, P. J., Dunn, P., Matallana, E., and Ecker, J. R. (1991). Genome mapping with anchored clones: Theoretical aspects. *Genomics* **11:** 799–805.

Feinberg, A. P., and Vogelstein, B. (1983). A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.* **132:** 6–13.

Genetics Computer Group (1991). Program manual for the GCG package, version 7, April, 1991, Madison, WI.

Green, E. D., and Olson, M. V. (1990). Chromosomal region of the cystic fibrosis gene in yeast artificial chromosomes: A model for human genome mapping. *Science* **250:** 94–98.

Hearne, C. M., Ghosh, S., and Todd, J. A. (1992). Microsatellites for linkage analysis of genetic traits. *Trends Genet.* **8:** 288–294.

Inohara, N., Iwamoto, A., Moriyama, Y., Shimomura, S., Maeda, M., and Futai, M. (1991). Two genes, atpC1 and atpC2, for the gamma subunit of *Arabidopsis thaliana* chloroplast ATP synthase. *J. Biol. Chem.* **226:** 7333–7338.

Intapruk, C., Higashimura, N., Yamamoto, K., Okada, N., Shinmyo, A., and Takano, M. (1991). Nucleotide sequences of two genomic DNAs encoding peroxidase of *Arabidopsis thaliana*. *Gene* **98:** 237–241.

Kieber, J. J., Rothenberg, M., Roman, G., Feldmann, K. A., and Ecker, J. R. (1993). CTR1, a negative regulator of the ethylene response pathway in *Arabidopsis*, encodes a member of the Raf family of protein kinases. *Cell* **72:** 427–441.

Konieczny, A., and Ausubel, F. (1993). A procedure for quick mapping of *Arabidopsis* mutants using ecotype specific markers. *Plant J.* **4:** 403–410.

Koornneef, M. (1990). Linkage map of *Arabidopsis thaliana* ($2n = 10$). *In* "Genetic Maps" (S. J. O'Brien. Ed.), pp. 6.95–6.97, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Krebbers, E., Seurinck, J., Herdies, L., Cashmore, A. R., and Timko, M. P. (1988). Four genes in two diverged subfamilies encode the ribulose 1,5-bisphosphate carboxylase small subunit polypeptides of *Arabidopsis thaliana*. *Plant Mol. Biol.* **11:** 745–760.

Lagercrantz, U., Ellegren, H., and Andersson, L. (1993). The abundance of various polymorphic microsatellite repeats differs between plants and vertebrates. *Nucleic Acids Res.* **21:** 1111–1115.

Lander, E. S., Green, P., Abrahamson, J., Barlow, A., Daly, M. J., Lincoln, S. E., and Newburg, L. (1987). MAPMAKER, an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1:** 174–181.

Liang, X., Keller, J. A., Abel, S., Shen, N. F., and Theologis, A. (1992). The 1-aminocyclopropane 1-carboxylate synthase gene family of *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **89:** 11046–11050.

Lincoln, S., Daly, M., and Lander, E. (1992). Constructing genetic maps with MAPMAKER/EXP 3.0. *Whitehead Institute Technical Report, 2nd ed.*

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

Litt, M., and Luty, J. A. (1989). A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am. J. Hum. Genet.* **44:** 397–401.

Manly, K. F. (1993). A Macintosh program for storage and analysis of experimental genetic-mapping data. *Mamm. Genome* **4:** 303–313.

Manolson, M. F., Ouellette, B. F., Filion, M., and Poole, R. J. (1988). cDNA sequence and homologies of the 57 kDa nucleotide binding subunit of the vacuolar ATPase from *Arabidopsis*. *J. Biol. Chem.* **263:** 17987–17994.

Morgante, M., and Olivieri, A. M. (1993). PCR amplified microsatellites in plant genetics. *Plant J.* **3:** 175–182.

Nam, H.-G., Giraudat, J., Den Boer, B., Moonan, F., Loos, W. B. D., Hauge, B. M., and Goodman, H. M. (1989). Restriction fragment length polymorphism map of *Arabidopsis thaliana*. *Plant Cell* **1:** 953–960.

NIH/CEPH Collaborative Mapping Group (1992). A comprehensive linkage map of the human genome. *Science* **258:** 67–86.

Olson, M. V., Hood, L., Cantor, C., and Botstein, D. (1989). A common language for physical mapping of the human genome. *Science* **245:** 1434–1435.

Oppenheimer, D. G., Herman, P. L., Sivukuraman, S., Esch, J., and Marks, M. D. (1991). A myb gene required for leaf trichome differentiation is expressed in stipules. *Cell* **67:** 483–493.

Ostrander, E. A., Jong, P. M., Rine, J., and Duyk, G. (1992). Construction of small-insert genomic DNA libraries highly enriched for microsatellite repeat sequences. *Proc. Natl. Acad. Sci. USA* **89:** 3419–3423.

Pang, P. P., Pruitt, R. E., and Meyerowitz, E. M. (1988). Molecular cloning, genomic organization and evolution of 12s seed storage protein genes of *Arabidopsis thaliana*. *Plant Mol. Biol.* **11:** 805–820.

Pearson, W. R., and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85:** 2444–2448.

Reiter, R. S., Williams, J. G. K., Feldmann, K. A., Rafalski, J. A., Tingey, S. V., and Scolnik, P. A. (1992). Global and local genome mapping in *Arabidopsis thaliana* by using recombinant inbred lines and random amplified polymorphic DNAs. *Proc. Natl. Acad. Sci. USA* **89:** 1477–1481.

Samac, D. A., Hironaka, C. M., Yallaly, P. E., and Shah, D. M. (1990). Isolation and characterization of the genes encoding basic and acidic chitinase in *Arabidopsis thaliana*. *Plant Physiol.* **93:** 907–914.

Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). "Molecular Cloning, A Laboratory Manual," Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Schindler, U., Menkens, A. E., Beckmann, H., Ecker, J. R., and Cashmore, A. R. (1992). Heterodimerization between light-regulated and ubiquitously expressed *Arabidopsis* GBF bZIP proteins. *EMBO J.* **11:** 1261–1273.

Simoens, C. R., Peleman, J., Valvekens, D., Van Montagu, M. M., and Inze, D. (1988). Isolation of genes expressed in specific tissues of *Arabidopsis* thaliana by differential screening of a genomic library. *Gene* **67:** 1–11.

Tautz, D. (1989). Hypervariability of simple sequences as a general source of polymorphic DNA markers. *Nucleic Acids Res.* **17:** 6463–6471.

Tisdall, J. (1993). DNA WorkBench. Technical report MS-CIS-93-38, Department of Computer Science, University of Pennsylvania.

Weber, J. L., and May, P. E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44:** 388–396.

Wilkinson, J. Q., and Crawford, N. M. (1991). Identification of the *Arabidopsis* CHL3 gene as the nitrate reductase structural gene nia2. *Plant Cell* **3:** 461–471.