Arabidopsis *PAI* Gene Arrangements, Cytosine Methylation and Expression

Stacey Melquist, Bradley Luff and Judith Bender

Department of Biochemistry, Johns Hopkins University School of Public Health, Baltimore, Maryland 21205

Manuscript received April 1, 1999
Accepted for publication May 10, 1999

ABSTRACT

Previous analysis of the *PAI* tryptophan biosynthetic gene family in *Arabidopsis thaliana* revealed that the Wassilewskija (WS) ecotype has four *PAI* genes at three unlinked sites: a tail-to-tail inverted repeat at one locus (*PAI1-PAI4*) plus singlet genes at two other loci (*PAI2* and *PAI3*). The four WS *PAI* genes are densely cytosine methylated over their regions of DNA identity. In contrast, the Columbia (Col) ecotype has three singlet *PAI* genes at the analogous loci (*PAI1*, *PAI2*, and *PAI3*) and no cytosine methylation. To understand the mechanism of *PAI* gene duplication at the polymorphic *PAI1* locus, and to investigate the relationship between *PAI* gene arrangement and *PAI* gene methylation, we analyzed 39 additional ecotypes of Arabidopsis. Six ecotypes had *PAI* arrangements similar to WS, with an inverted repeat and dense *PAI* methylation. All other ecotypes had *PAI* arrangements similar to Col, with no *PAI* methylation. The novel *PAI*-methylated ecotypes provide insights into the mechanisms underlying *PAI* gene duplication and methylation, as well as the relationship between methylation and gene expression.

THE model higher plant *Arabidopsis thaliana* has a small genome size of $\sim 10^8$ bp. Despite the relative simplicity of the Arabidopsis genome, many functions in this plant are encoded by gene families rather than by a single gene. For example, many of the enzymes in the tryptophan biosynthetic pathway are encoded by two- (Last *et al.* 1991; Niyogi and Fink 1992) or three-member (Niyogi *et al.* 1993; Li *et al.* 1995) gene families. Important questions pertaining to these gene families are how their gene duplications arose and how they are stably maintained in the genome.

In particular, the gene family encoding the enzyme that catalyzes the third step of the tryptophan biosynthetic pathway, *p*hosphoribosyl*a*nthranilate *i*somerase (PAI), displays a number of intriguing features. The *PAI* gene family has been characterized previously in detail in two standard laboratory strains of Arabidopsis, Columbia (Col), and Wassilewskija (WS; Bender and Fink 1995; Li *et al.* 1995). These analyses revealed that in both strains there is an unusually high degree of sequence identity among the *PAI* family members, including untranslated portions of the genes, such as intron and promoter sequences. In Col, the *PAI* family is encoded by three unlinked genes: *PAI1* on the upper arm of chromosome 1, *PAI2* on the upper arm of chromosome 5, and *PAI3* in the middle of chromosome 1. The *PAI1* and *PAI2* genes are almost perfectly identical to each other over the 2307 bp extending from 355 bp upstream of the ATG translational start codon to 470 bp downstream of the TAA translational stop codon, including five exons and four introns. Within this region, the Col *PAI1* and *PAI2* genes differ by only 26 scattered single-base differences (99% identity). In contrast, the *PAI3* gene is 90% identical to *PAI1* and *PAI2* due to many scattered base differences. The base differences in the translated parts of the *PAI3* gene are predicted to yield a protein with 18 amino acid differences from Col *PAI1*.

The high degree of sequence identity across *PAI* gene family members is unusual in comparison to other Arabidopsis tryptophan gene families. For example, the duplicated tryptophan synthase β-subunit genes *TSB1* and *TSB2* are 85% identical to each other in their exon sequences (exclusive of presumed chloroplast target sequences), but they are highly divergent in their transcribed untranslated sequences (Last *et al.* 1991). Furthermore, the duplicated anthranilate synthase α-subunit genes *ASA1* and *ASA2* are 67% identical to each other in their predicted amino acid sequences, but they are divergent in their exon and intron nucleic acid sequences and in their intron structures (Niyogi and Fink 1992). Therefore, the high degree of identity among the Col *PAI* genes, particularly between *PAI1* and *PAI2*, suggests that they evolved relatively recently.

In WS, there are two more unusual features of the *PAI* gene family. First, whereas WS carries *PAI2* and *PAI3* genes that are almost identical to the Col *PAI2* and *PAI3* genes, at the *PAI1* locus, WS carries an inverted-repeat duplication of the *PAI1* gene *PAI1-PAI4* (Bender and Fink 1995). The *PAI1-PAI4* inverted repeat is flanked by $\sim$2.9 kb of perfect direct-repeat sequences. Furthermore, the *PAI1*-proximal direct repeat extends into the first 731 bp of a duplicated *PAI4* gene *pai4 5′* (Bender and Fink 1995; Figure 3).

The second unusual feature of the WS *PAI* genes is

Corresponding author: Judith Bender, Department of Biochemistry, Johns Hopkins University School of Public Health, 615 N. Wolfe St., Baltimore, MD 21205. E-mail: jbender@welchlink.welch.jhu.edu

that all four full-length genes and the partial *pai4 5′* gene are densely cytosine methylated (Bender and Fink 1995; Luff *et al.* 1999). This methylation covers the regions of the genes that have sequence identity to each other, but it does not spread more than a few hundred bases beyond the boundaries of *PAI* sequence identity. In contrast, the three Col *PAI* genes do not display detectable methylation (Bender and Fink 1995). *PAI* methylation in WS correlates with silencing of at least the *PAI2* gene, but there is sufficient total *PAI* expression in this strain for a normal plant phenotype (Bender and Fink 1995; Jeddeloh *et al.* 1998).

Because the *PAI1* locus is polymorphic between WS and Col, we reasoned that an analysis of *PAI* gene copy number, arrangements, and methylation in other Arabidopsis ecotypes might provide new insights into the correlation between gene structure and the onset of cytosine methylation, as well as provide new genetic tools for understanding methylation and its relationship to gene expression. In addition, the *PAI* structures observed in other ecotypes could elucidate the mechanism of *PAI* gene duplication. To investigate this possibility, we screened *PAI* gene structure and methylation by Southern blot analysis for 39 additional ecotypes of Arabidopsis isolated from around the world. We found that whereas most ecotypes had *PAI* gene structures identical to that observed for Col (three unmethylated genes) and two ecotypes had *PAI* gene structures nearly identical to that observed for WS (four methylated genes), four ecotypes had novel *PAI* arrangements, with a variation of the inverted-repeat *PAI* gene structure at the *PAI1* locus and *PAI* cytosine methylation. On the basis of our results, we discuss possible models for the differential methylation of *PAI* gene arrangements, the evolution of the polymorphic *PAI* gene family, and the effects of methylation on *PAI* gene expression.

MATERIALS AND METHODS

Arabidopsis strains and growth conditions: Arabidopsis ecotypes were obtained from the Arabidopsis Biological Resource Center at Ohio State University, with the exception of C24, which was obtained from Patrick Masson (University of Wisconsin, Madison, WI). Plants were grown under continuous light in Fafard Growing Mix 2 (Griffin Greenhouse Supplies).

Plant genomic DNA preparation and Southern blot analysis: Plant genomic DNA was prepared as follows. Four-week-old plants were frozen in liquid nitrogen and ground into a fine powder with a mortar and pestle. Ground tissue (~10 g) was thawed in 20 ml lysis buffer (100 mm Tris-OH, pH 9.5, 1.4 m NaCl, 20 mm EDTA, pH 8.0, 2% hexadeclytrimethylammonium bromide, 1% polyethylene glycol, M_r 8000), plus 50 μl 2-mercaptoethanol and then heated at 74° for 20 min. The lysed tissue was cooled to room temperature and extracted with an equal volume of $CHCl_3$. The aqueous phase was transferred to a fresh tube and precipitated with an equal volume of isopropanol at room temperature for 30 min, followed by centrifugation at 5000 × *g* for 20 min. The pellet was resuspended in 1 ml 0.75 M NaCl, incubated with 5 μl 10 mg/ml RNase A at 37° for 30 min, mixed with 0.25 ml of water and 0.75 ml of JetStar (Genomed) column equilibration buffer E4, and centrifuged for 10 min at 5000 × *g* to pellet insoluble material. The DNA sample supernatant was loaded on an equilibrated JetStar minicolumn and washed, eluted, and isopropanol precipitated as described by the manufacturer. DNA pellets were resuspended to a final concentration of ~0.5 μg/μl in water. For Southern blot analysis, ~1 μg of genomic DNA was digested with the restriction endonuclease of interest, electrophoresed through an 0.7% TBE agarose gel, transferred to a Hybond-N (Amersham, Arlington Heights, IL) membrane, and fixed by UV cross-linking. Purified probe DNA fragments were labeled using the MegaPrime kit (Amersham) as described by the manufacturer and hybridized to Southern blots in Church buffer (Church and Gilbert 1984) at 65°, followed by several washes with 0.2× SSC/0.1% SDS at 65°. The *PAI1* cDNA probe is an internal 0.7-kb *Pst*I fragment (Bender and Fink 1995; Li *et al.* 1995). The direct-repeat probe is a 2.3-kb *Hinc*II-to-*Xho*I fragment derived from the sequences upstream of the WS *PAI4* gene.

Construction and screening of genomic DNA libraries: Genomic DNA isolated from Kas-1, C24, Ita-0, or Cvi-0 was partially digested with *Sau*3AI to an average fragment size of 15–20 kb, ligated with λDASH (Stratagene, La Jolla, CA) arms, and packaged into phage particles using Gigapack Gold (Stratagene) *in vitro* phage packaging extracts as described by the manufacturer. Each library contained 100,000–200,000 primary clones. *PAI*-positive plaques were isolated by transferring to Hybond-N membranes and hybridizing with a *PAI* cDNA or direct-repeat probe as described for Southern blot analysis above.

DNA sequencing: Genomic DNA fragments were subcloned from λ library isolates into pBluescript II KS+ (Stratagene) and were sequenced either from standard T3 and T7 primer sites in the vector or from custom-designed internal primers by the Johns Hopkins University Department of Biological Chemistry Biosynthesis and Sequencing Facility.

PCR analysis of sequences flanking the *PAI1*-proximal direct repeat: PCR primers based on sequences lying just outside the 731-bp *pai4 5′* duplication in WS and the heterologous 944-bp sequence present at this region in Col (see Figure 3) were used to amplify the analogous regions from Kas-1, C24, and Ita-0 genomic DNA using standard PCR reagents and conditions. These three ecotypes gave fragments that were identical in size and in *Pst*I or *Hinc*II restriction patterns to the Col fragment. Also, genomic clones of this region from Kas-1 and Cvi-0 were explicitly sequenced and found to be nearly identical to Col.

Isolation and analysis of *PAI* cDNAs: Standard Col (Elledge *et al.* 1991), Landsberg erecta (Ler, Minet *et al.* 1992), and WS (gift of L. Castle and D. Meinke, obtained from the Arabidopsis Biological Resource Center) cDNA libraries were screened by hybridization with the *PAI1* cDNA probe to identify *PAI* clones. Inserts were subcloned into pBluescript II KS+ and sequenced to determine their 5′ and 3′ end structures and to distinguish from which *PAI* gene they arose. PAI function was tested by determining whether cDNAs expressed from the *lac* promoter on pBluescript II KS+ in a *pai*-deficient *Escherichia coli* strain W3110 *trpC9830* could complement the mutant grown on minimal M9 medium, as described previously (Bender and Fink 1995; Li *et al.* 1995). Site-directed mutageneses to remove the 25-bp intron insertion in the *PAI3* cDNA, to introduce the WS *PAI4* 9-bp deletion into a WS *PAI1* cDNA, or to introduce the C24 *PAI4* 6-bp deletion into a WS *PAI1* cDNA were performed using standard methods (Kunkel *et al.* 1987).

Reverse transcriptase PCR analysis of *PAI* cDNAs: Total Arabidopsis RNA isolated as described (Nagy *et al.* 1988) was treated with RNase-free DNase (Promega, Madison, WI) and used as a template for murine Maloney leukemia virus reverse

transcriptase (M MLVRT, GIBCO-BRL, Bethesda, MD). Specifically, 20-μl reactions containing 2.0 μg total RNA, 0.5 mm of each dNTP, the appropriate reverse primer, and buffer conditions recommended by the manufacturer were incubated at 65° for 5 min, then chilled to 4°, at which point 200 units of M MLVRT was added. Reactions were then incubated at 42° for 1 hr, heated to 95° for 5 min, and chilled to 4°. For each RT reaction, 0.5 μl was used as a template for PCR amplification in a 100-μl volume using standard reagents and an amplification program of 40 cycles of (94° 30 sec, 55° 30 sec, 72° 1 min). Detailed information about primer sequences used in this analysis is available upon request from J. Bender. RT-PCR products <300 bp were resolved by electrophoresis through 3% agarose/1% NuSieve (FMC, Rockland, ME) TBE gels, and RT-PCR products >300 bp were resolved on 1.5% agarose TBE gels.

Northern blot analysis: RNA was prepared from whole 4-wk-old plants, electrophoresed, transferred to Hybond-N (Amersham) membranes using standard procedures (Ausubel *et al.* 1989), and hybridized with a *PAI1* cDNA probe as described for Southern blot analysis above. Blots were subsequently stripped and reprobed with an α-tubulin probe to normalize for differences in loading.

RESULTS

Determination of *PAI* gene structures and methylation in Arabidopsis ecotypes: Because the *PAI1* locus is structurally polymorphic between the previously characterized Col and WS Arabidopsis ecotypes (Bender and Fink 1995), we screened 39 new ecotypes for *PAI* gene structures using Southern blot assays (Table 1, Figures 1 and 2). This analysis revealed that most ecotypes had *PAI* structures similar to Col, with a singlet gene at the *PAI1* locus, but six ecotypes displayed band patterns diagnostic of novel arrangements at the *PAI1* locus. Furthermore, because *PAI* sequences are not cytosine methylated in the Col ecotype but are densely methylated in the WS ecotype, we also screened the 39 new ecotypes for evidence of *PAI* gene methylation using a Southern blot assay. This analysis revealed that the ecotypes with a singlet *PAI1* gene structure at the *PAI1* locus (the Col arrangement) displayed no detectable *PAI* methylation. In contrast, like WS, the six ecotypes with novel structures at the *PAI1* locus displayed band patterns diagnostic of dense *PAI* methylation (Table 1, Figures 2 and 3). These observations suggest that the acquisition of an unusual structure at the *PAI1* locus causes *PAI* gene methylation.

As a preliminary screen of both *PAI* structure and methylation, we digested genomic DNA from 39 new ecotypes with the methylation-sensitive restriction endonuclease isoschizomers *Hpa*II and *Msp*I, followed by Southern blot analysis with a *PAI1* cDNA probe. Each *PAI* locus carries a single conserved *Hpa*II/*Msp*I site in the second *PAI* intron with different flanking *Hpa*II/*Msp*I sites (Figure 3). Therefore, each *PAI* locus yields distinct *Hpa*II/*Msp*I fragments, diagnostic of either fully cleaved unmethylated or partially cleaved methylated sequences. *PAI* methylation detected by *Hpa*II/*Msp*I digestion was previously shown to correlate with *PAI* methylation across the rest of the WS *PAI* genes detected by several other restriction endonucleases and by a geno-

TABLE 1

Arabidopsis ecotype *PAI* gene arrangements detected by *Hpa*II/*Msp*I Southern blots

Ecotype name[a]	Habitat[b]	*PAI* gene arrangement
Aa-0	Germany	= Col[c]
Ag-0	France	= Col
An-1	Belgium	= Col
Ba-1	United Kingdom	= Col
Be-0	Germany	= Col
Br-0	Czechoslovakia	= Col
Bur-0	Ireland	= WS[d]
C24	?	Novel, methylated
Can-0	Canary Islands	= Col
Col-0	USA	= Col
Ct-1	Italy	= Col
Cvi-0	Cape Verde Islands	Novel, methylated
Di-0	France	= Col
Edi-0	Scotland	= Col
En-1	Germany	= Col
Es-0	Finland	= Col
Est-0	Russia	= Col
Ge-0	Switzerland	= Col
In-0	Austria	= Col
Ita-0	Morocco	Novel, methylated
Jm-0	Czechoslovakia	= Col
Kas-1	India	Novel, methylated
Kil-0	UK	= Col
Lc-0	Scotland	= Col
Le-0	Netherlands	= Col
Ler	Germany	= Col
Ll-0	Spain	= Col with deletion of *PAI1*
Lu-1	Sweden	= Col with *PAI3* polymorphism
Mh-0	Poland	= Col
Ms-0	Russia	= Col
Mt-0	Libya	= Col
Mv-0	USA	= Col
Nd-0	Germany	= WS with *PAI3* *Msp*I polymorphism
No-0	Germany	= Col
Oy-0	Norway	= Col
Pog-0	Canada	= Col
RLD	?	= Col
Sei-0	Spain	= Col
Tsu-0	Japan	= Col
Tul-0	USA	= Col
WS-0	Russia	= WS

[a] Ecotype name abbreviations as given by the Arabidopsis Biological Resource Center.

[b] Habitat as reported by the Arabidopsis Biological Resource Center.

[c] Identical to the Col *Hpa*II/*Msp*I *PAI* pattern shown in Figure 2.

[d] Identical to the WS *Hpa*II/*Msp*I *PAI* pattern shown in Figure 2.

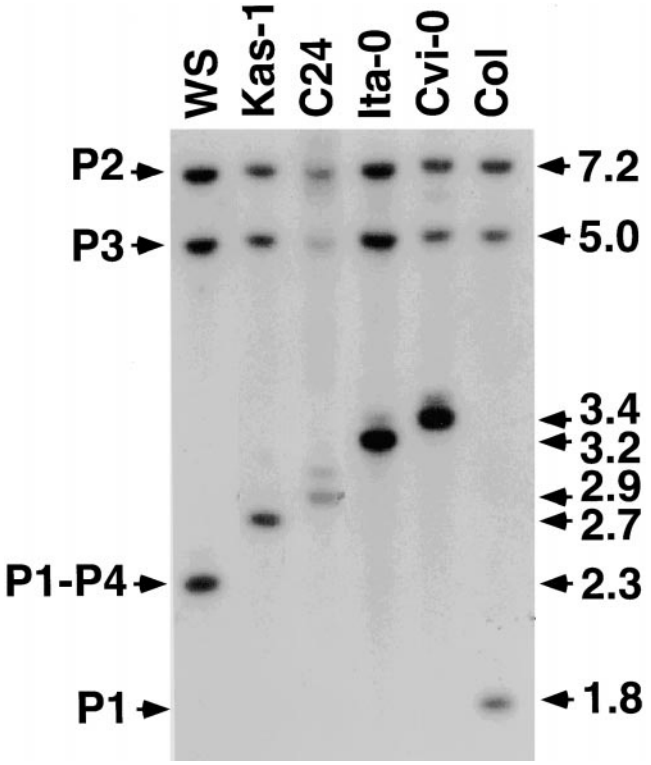

Figure 1.—The *PAI1* locus is polymorphic. Genomic DNA prepared from the indicated ecotypes was cleaved with *Hin*dIII and probed with a *PAI*-internal cDNA fragment. *Hin*dIII cleaves each gene twice 170 bp apart in the first and second introns; both sites lie upstream of the probe fragment so that each singlet gene yields a single hybridizing band. *PAI* inverted repeats also yield a single hybridizing band because *Hin*dIII does not cleave between *PAI1* and *PAI4*. In the Cvi-0 *PAI1* gene, the second intron *Hin*dIII site is missing because of a single-base polymorphism, so this gene yields a fragment with an additional 170 bp relative to *Hin*dIII digests of other *PAI1* genes. Note that in Cvi-0, the partial direct-repeat *PAI4** duplication (Figure 3) yields a *PAI*-hybridizing *Hin*dIII band of the same size as the full *PAI1-PAI4* inverted repeat so that the two bands are superimposed on this blot. P1 is Col *PAI1*, P1-P4 is WS *PAI1-PAI4*, P2 is *PAI2*, and P3 is *PAI3*. The molecular weight of each band is indicated along the right margin in kilobases. The molecular weight of the *PAI3* band is estimated relative to size standards. The molecular weights of all other bands are based on genomic sequences.

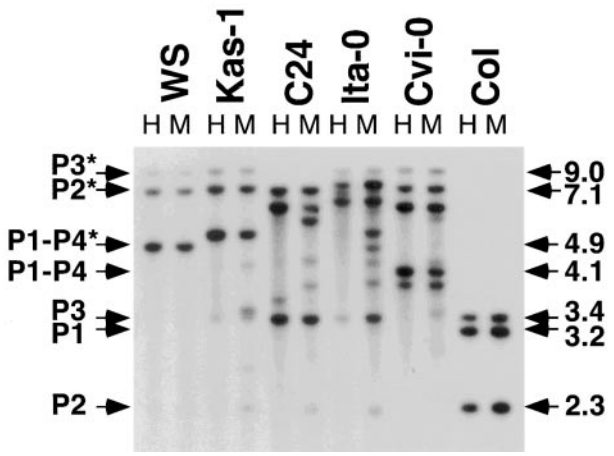

Figure 2.—Ecotypes with novel *PAI* gene structures and cytosine methylation. Genomic DNA prepared from the indicated Arabidopsis ecotypes was cleaved with either *Hpa*II (H) or *Msp*I (M) and probed with a *PAI*-internal cDNA fragment. *Hpa*II / *Msp*I restriction maps and the region covered by the probe are shown in Figure 3. *Hpa*II and *Msp*I are both blocked from cleaving by cytosine methylation of their recognition sequence, 5′ CCGG 3′, although *Hpa*II is sensitive to methylation of either cytosine in the sequence and *Msp*I is sensitive only to methylation of the outer cytosine. Bands diagnostic of methylation are marked in the margin with asterisks. P1 is Col *PAI1*, P1-P4 is WS *PAI1-PAI4*, P2 is *PAI2*, and P3 is *PAI3*. The molecular weights of these bands are indicated along the right margin in kilobases. The molecular weights of the *PAI3* bands are estimated relative to size standards. The molecular weights of all other bands are based on genomic sequences.

mic methylation sequencing technique (Bender and Fink 1995; Jeddeloh *et al.* 1998; Luff *et al.* 1999).

Out of the 39 new ecotypes examined, 31 had *Hpa*II / *Msp*I *PAI* patterns identical to those observed in Col, diagnostic of three unmethylated genes (Table 1, Figure 2). Two other ecotypes had Col-related patterns. Lu-1 had *PAI1* and *PAI2* patterns identical to those observed for Col, but a different *PAI3* pattern. This pattern was most likely caused by a *Hpa*II / *Msp*I *PAI3* polymorphism because when Lu-1 DNA was cleaved with *Xho*I, it yielded a band pattern identical to Col (data not shown). Ll-0 had *PAI2* and *PAI3* patterns identical to those observed for Col, but no detectable *PAI1* fragment. This pattern was verified with a *Xho*I digest (data not shown).

Two ecotypes had band patterns related to those observed in WS, diagnostic of a complex *PAI* arrangement at the *PAI1* locus plus dense cytosine methylation across all three *PAI* loci (Table 1). One *PAI*-methylated ecotype, Bur-0, had *PAI* patterns identical to those observed in WS. Another *PAI*-methylated ecotype, Nd-0, had *PAI1-PAI4* and *PAI2* patterns identical to those observed in WS, but a different *PAI3* pattern. This pattern was most likely caused by a *Hpa*II / *Msp*I *PAI3* polymorphism because when Nd-0 DNA was cleaved with *Xho*I, it yielded a band pattern identical to WS (data not shown). Four other *PAI*-methylated ecotypes, Kas-1, C24, Ita-0, and Cvi-0, displayed band patterns consistent with novel complex structures at the *PAI1* locus (Figures 1 and 2).

Southern blot analysis of the four novel *PAI*-methylated ecotypes using the enzyme *Hin*dIII, which is relatively insensitive to cytosine methylation, confirmed that there were only three *PAI* loci in each of these strains (Figure 1). The *PAI2* and *PAI3* loci were structurally invariant relative to Col and WS, whereas the *PAI1* locus was polymorphic. The sizes of the *PAI1* locus bands were consistent with either singlet genes or inverted-repeat genes spaced further apart than in WS. Southern blot analysis with the enzyme *Xho*I (restriction map shown in Figure 3), and subsequent cloning and sequencing of the *PAI1* locus from each ecotype (see below), showed that each of the novel *PAI*-methylated ecotypes in fact

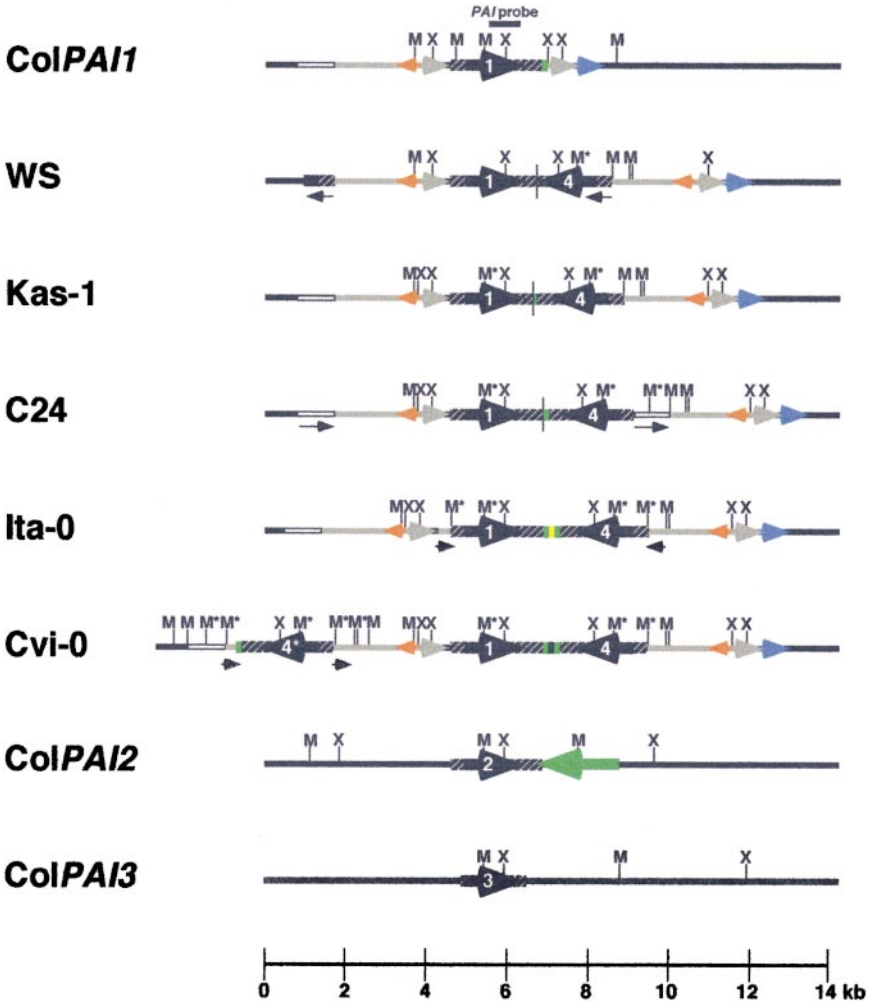

Figure 3.—Gene structures and restriction maps for the Arabidopsis *PAI* genes. The *Xho*I (X) and *Msp*I (M) restriction maps for the *PAI1* loci from the indicated ecotypes are shown with maps for the Col *PAI2* and *PAI3* loci. Thin gray arrows indicate the direct-repeat sequences flanking each *PAI1* locus. Wide colored arrows represent genes, with the start of the arrow at the ATG translational start codon and the tip of the arrowhead at the translational stop codon: *PAI* genes are in black, *S15a* is in blue, histone *H2B* is in red, and *FAD8* is in green. For the *PAI* genes, the additional upstream and downstream regions of shared sequence identity are indicated by underlying hatched black lines. In the regions between inverted-repeat *PAI* genes for WS, Kas-1, and C24, the junctions of *PAI1* and *PAI4* downstream sequences are marked with a vertical line. For Kas-1, C24, Ita-0, and Cvi-0, central sequences that include the *FAD8* coding region are marked in green. In Ita-0, the unique central sequences are shown in yellow, and in Cvi-0, the central sequences derived from *PAI* sequences are shown in black. See Figure 4 for a close-up view of the structures between inverted-repeat *PAI1* and *PAI4* genes. *PAI1-PAI4* duplications are oriented with the *PAI1* gene on the left and the *PAI4* gene on the right. In Cvi-0, the leftmost black arrow represents the duplicated *PAI4** gene, and in WS, the short leftmost black box represents the partial *pai4 5′* duplication. The white box upstream of the *PAI1* gene in most ecotypes indicates the 944 bp of heterologous sequence relative to WS *pai4 5′*. Black arrows drawn under short duplicated sequences in Cvi-0, Ita-0, C24, and WS indicate the relative orientations and extent of the duplications. Asterisks indicate methylation. The methylated MspI sites around the Cvi-0 *PAI4** duplication could not be determined precisely because of the density of sites flanking this region; the marked sites around this locus are therefore an approximation. Sequences for all loci except Col *PAI3* are available from GenBank: Col *PAI1*, AF130878; Col *PAI2*, AB005241; WS *PAI1-PAI4* (updated), U34757; Cvi-0 *PAI1-PAI4*, AF130876; Cvi-0 *PAI4**, AF130877; Ita-0 *PAI1-PAI4*, AF130875; C24 *PAI1-PAI4*, AF130874; and Kas-1 *PAI1-PAI4*, AF130873.

carried a variation of the *PAI1-PAI4* inverted-repeat structure.

The *PAI1-PAI4* locus in WS is flanked by ~3 kb of direct-repeat sequences (Figure 3; Bender and Fink 1995). Analysis of the novel *PAI*-methylated ecotypes with a direct-repeat probe indicated that like WS, all four ecotypes carried direct-repeat sequences flanking the inverted-repeat *PAI* genes (data not shown). Subsequent cloning and sequencing confirmed this result.

Sequence analysis of *PAI1* locus structures: To obtain a detailed understanding of the novel *PAI1* loci detected by Southern blot in Kas-1, C24, Ita-0, and Cvi-0, we constructed genomic libraries from these strains and screened for *PAI1* locus clones using direct-repeat and/or *PAI1* cDNA probes. Relevant portions of the clones were subcloned and sequenced. We also extended the existing sequences around the cloned WS and Col *PAI1* loci. These analyses, as well as the complete sequence of a P1 clone carrying the *PAI2* region of the genome (Sato *et al.* 1997), revealed a series of structural relationships among the six different *PAI* arrangements for WS, Kas-1, C24, Ita-0, Cvi-0, and Col, as well as a structural relationship between the unlinked *PAI2* gene and the *PAI1-PAI4* duplication (Figure 3).

Sequencing of the WS *PAI1-PAI4* locus over a continuous region of ~11 kb (Figure 3) showed that the direct repeats flanking *PAI1-PAI4* were nearly identical to each other: the *PAI1*-proximal repeat was 2880 bp long, and the *PAI4*-proximal repeat was 2905 bp long because of the presence of a 25-bp insertion near the end furthest away from *PAI4*. Each of the repeats contained a complete open reading frame for a gene with identity to histone *H2B*, although this gene was not represented in the Arabidopsis Expressed Sequence Tag (EST) database. Immediately upstream of the *PAI1*-proximal direct repeat was a 731-bp partial *pai4 5′* duplication followed

by unique sequences with no significant identity to the sequences in the database. The unique sequences immediately beyond the *PAI4*-proximal direct repeat encoded ribosomal protein S15a (Bonham-Smith and Moloney 1994). Based on an EST sequence for a Col *S15a* cDNA (GenBank accession no. R31305), the promoter, transcriptional start, and translational start of the *S15a* gene are duplicated in both copies of the direct repeat so that the *S15a* promoter lies upstream of both *PAI1* and *S15a* (Figure 3). Furthermore, a comparison of the *S15a* genomic sequence to the cDNA sequence indicated that there is an intron in the upstream untranslated sequence that is included in the direct repeats.

Previous sequence analysis of the Col *PAI1* locus indicated that this locus carries a deleted direct repeat of 813 bp relative to the WS sequence downstream of *PAI1*, followed by the *S15a* gene (Bonham-Smith and Moloney 1994; Bender and Fink 1995; Li *et al.* 1995; Figure 3). Southern blot analysis was used to determine that upstream of *PAI1* there is at least a partial direct-repeat sequence but no *pai4 5′* duplication (Bender and Fink 1995). To better understand the structure in this upstream region, we sequenced through the upstream direct repeat into unique flanking sequences. This analysis revealed that Col carried an almost-complete upstream direct repeat relative to WS, missing only the 5 bp most distal to the *PAI1* gene. Therefore, Col and WS have almost identical *PAI1* promoter sequences over a region of ~3.2 kb upstream of the *PAI1* translational start. Beyond this point, Col carried 944 bp of the heterologous sequence relative to the WS 731-bp *pai4 5′* duplication (Figure 3). Further upstream of this heterologous region, however, both ecotypes had identical sequences. The 944-bp region did not have any significant identity to other sequences in the database.

The novel methylated ecotypes all had overall structures similar to that of WS at the *PAI1* locus, with two direct repeats flanking a *PAI1-PAI4* inverted repeat. However, each ecotype had unique variations on this basic pattern (Figure 3). For example, the Kas-1 and C24 ecotypes had the same structure as Col upstream of *PAI1*. The Ita-0 ecotype also had a Col-like sequence upstream of *PAI1*, except that substituting for the 33 bp of direct-repeat sequence most proximal to *PAI1* was 300 bp of the *PAI1*-distal end of the direct-repeat sequence in an inverted orientation. In Cvi-0, the region upstream of *PAI1* had a more complex structure than any of the other ecotypes examined. Beyond the *PAI1*-proximal direct repeat, there was a duplication consisting of a full-length *PAI* gene oriented in the same direction as *PAI4* (*PAI4**) plus the last 207 bp of a direct-repeat sequence. Beyond this deleted direct-repeat *PAI4** duplication was the same upstream sequence found in Col, Kas-1, C24, and Ita-0.

In the region upstream of *PAI4* the ecotypes Kas-1, Ita-0, and Cvi-0 each had a full direct repeat followed by *S15a* sequences similar to WS (Figure 3). In C24, however, 96 bp around the junction between *PAI4* and its flanking direct repeat was substituted with 929 bp of heterologous sequences. These sequences consisted of a 911-bp duplication of the sequences found immediately adjacent to the upstream end of the *PAI1*-proximal direct repeat, followed by 18 bp of unique sequence at the junction with *PAI4*. Beyond this duplication, C24 carried direct-repeat and *S15a* sequences analogous to those found in the other ecotypes (Figure 3).

Another focus of our analysis was the sequence between the *PAI1-PAI4* inverted repeats, representing the sequences 3′ of each *PAI* gene. In Col, the *PAI1* 3′ sequence extended for 470 bp downstream of the translational stop codon with almost perfect identity to the Col *PAI2* gene 3′ sequence, followed by 10 bp of heterologous sequence and then 813 bp of flanking direct-repeat sequence (Figures 3 and 4). In the ecotypes WS, Kas-1, and C24, the *PAI1-PAI4* inverted-repeat central sequences consisted of deleted palindromes of the 470-bp downstream sequence (Figure 4). In Ita-0 and Cvi-0, *PAI1* and *PAI4* both carried the full 470-bp downstream sequence with a short novel sequence sandwiched in between. In Ita-0, this central sequence carried 24-bp inverted-repeat ends plus another repeat of the 24-bp sequence and a repeat of the last 43 bp of the *PAI2*-identical downstream sequences in the middle. In Cvi-0,the central sequence carried a short duplication of a *PAI* fifth-exon sequence flanked by novel sequences. There was no identity between Ita-0 and Cvi-0 central sequences.

With a combination of sequencing and restriction mapping analysis, we were able to construct *Hpa*II/*Msp*I restriction maps for the novel *PAI1* loci. This information plus the observed fragment sizes on a *Hpa*II/*Msp*I Southern blot (Figure 2) allowed a determination of which sites in each *PAI1-PAI4* structure were cytosine methylated (marked with asterisks in Figure 3). Methylation was almost entirely contained within the *PAI*-identical sequences, with little or no spread into flanking direct repeats.

***PAI* cDNA abundance and function studies:** Using a deletion mutant derivative of WS that lacks the *PAI1-PAI4* inverted-repeat genes, we previously showed that cytosine methylation correlates with a loss of expression for the *PAI2* gene (Bender and Fink 1995; Jeddeloh *et al.* 1998). Genetic backcross experiments indicated that *PAI2* is also silenced in the parental WS strain (Bender and Fink 1995). Nonetheless, parental WS is phenotypically normal, with levels of total *PAI* transcripts and PAI enzyme activity comparable to levels measured in Col, suggesting that there is a relatively high level of *PAI* expression from the methylated *PAI1* and/or *PAI4* genes (Bender and Fink 1995).

To better understand which *PAI* genes are expressed in *PAI*-methylated ecotypes, we analyzed *PAI* transcripts in WS by cloning and sequencing *PAI* cDNAs from a standard WS library. This approach revealed that *PAI1*

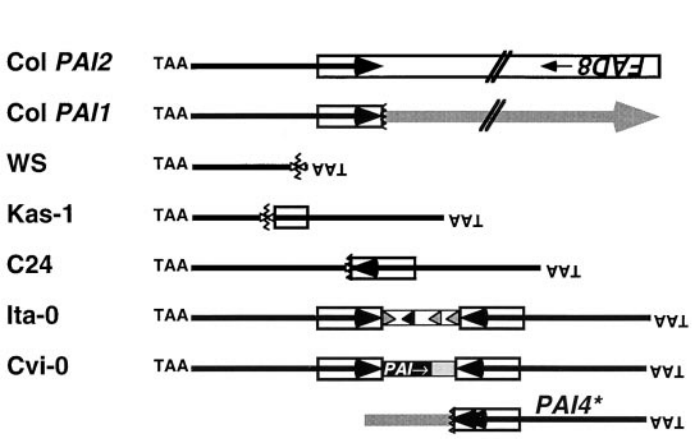

Figure 4.—Structure of central regions in *PAI1-PAI4* loci. The regions downstream of the *PAI1* and *PAI4* translational stop codons (TAA) are shown for the indicated ecotypes, with *PAI1* on the left and *PAI4* on the right. Black arrows represent the 470 bp of conserved *PAI* 3′ sequence. White arrowheads indicate the termini of deletions within this sequence. Vertical zigzag lines indicate deletion junctions. The double black arrowheads near the end of the Cvi-0 *PAI4** sequence represent a 21-bp tandem duplication in this region. The divergently transcribed *FAD8* coding sequence (2014 bp from translational start to translational stop) downstream of *PAI2* is shown as an open box, with hatched lines indicating that this region is not drawn to scale. The boxed regions at the ends of some *PAI1* and *PAI4* 3′ sequences indicate *FAD8* duplications (shown in green in Figure 3). The light gray arrow downstream of Col *PAI1* represents the 813-bp flanking direct-repeat sequence, with hatched lines indicating that this region is not drawn to scale. Similarly, the light gray line downstream of Cvi-0 *PAI4** represents the 207-bp deleted direct-repeat sequence. The partial *PAI* fifth-exon duplication in Cvi-0 is indicated by a black box, and the flanking heterologous sequence is indicated by a gray box. The novel sequence between the Ita-0 *PAI1* and *PAI4* genes is indicated by a white box. The black arrowhead in this box represents the partial duplication of *PAI* 3′ sequences, and the gray arrowheads in this box represent novel 24-bp duplications. The table at the right indicates the number of base pairs of sequence included in the *PAI1* and *PAI4* 3′ ends and additional center material relative to the conserved 470-bp sequence.

was the only detectable species (Table 2). In contrast, screening of Col and Ler cDNA libraries revealed that *PAI1* and *PAI2* were equally abundant in either of these unmethylated *PAI* ecotypes, with *PAI3* approximately fourfold less abundant than either of its sister genes. In both Ler and WS, rare *PAI1* cDNAs were found that had upstream direct-repeat sequences spliced to *PAI1* first-exon sequences (Table 2). These rare cDNAs are best explained as transcripts that initiate from upstream *S15a* sequences in the direct-repeat rather than more proximal *PAI1* sequences. The material that is deleted between the rare transcript start sites and the *PAI1* first-exon site corresponds to predicted intron material in the 5′ untranslated region of the *S15a* transcript on the basis of a comparison between an *S15a* cDNA sequence (GenBank accession no. R31305) and the genomic sequence. In the Ler 5′ spliced transcript, the 3′ junction of the *S15a* first upstream exon is spliced directly to a *PAI1* upstream site. In the WS 5′ spliced transcript, the 3′ junction of the *S15a* first upstream exon is correctly spliced to the second *S15a* exon, and then a cryptic site in the second exon is spliced to the same *PAI1* upstream site used in the Ler transcript (Table 2).

cDNA analysis also suggested that *PAI4* and *PAI3*, besides being poorly expressed, do not encode functional PAI enzymes. WS *PAI4* contains a 9-bp deletion in the fifth exon relative to *PAI1* and *PAI2* (Bender and Fink 1995). We explicitly mutagenized a WS *PAI1* cDNA in an *E. coli* expression vector to generate this deletion and then tested the ability of the *PAI1* parental construct *vs.* the 9-bp-deleted *PAI1* construct to complement the tryptophan auxotrophy of an *E. coli* PAI-deficient mutant. The WS *PAI1* cDNA complemented the mutant, as reported previously for Col *PAI1* and *PAI2* cDNAs (Table 2; Bender and Fink 1995; Li *et al.* 1995). However, the 9-bp deletion rendered the cDNA unable to complement the mutant. Therefore, WS *PAI4* is not likely to produce active PAI enzyme because of this deletion.

Sequencing of *PAI3* cDNAs isolated from Col and Ler libraries indicated that the *PAI3* transcript was incorrectly spliced to yield a 25-bp insertion of intron sequences between the fourth and fifth exons (Table 2). The insertion is most likely caused by a single-base polymorphism that changes the consensus 3′ intron site in the fourth *PAI* intron from AG to TG. *PAI3* cDNAs are predicted to encode a protein with 12 novel amino acids downstream of the fourth exon followed by a premature termination codon. When expressed in a PAI-deficient *E. coli* strain, the Col *PAI3* cDNA failed to complement tryptophan auxotrophy. Therefore, the *PAI3* gene is not likely to produce PAI activity in any ecotype that carries the *PAI3* splice site polymorphism, including Col, WS, C24, Kas-1, Ita-0, and Cvi-0 (as assessed by sequencing cloned *PAI3* genes and / or by monitoring a *Pst*I polymorphism created by the *PAI3* splice junction mutation).

RT-PCR analysis of *PAI* gene expression: As for WS, the novel *PAI*-methylated ecotypes Kas-1, C24, Ita-0, and Cvi-0 all had steady-state levels of total *PAI* transcripts slightly higher than the levels measured in Col (Figure 5A). Each *PAI*-methylated ecotype also displayed low levels of higher-molecular-weight *PAI* transcripts, as previously observed in WS (Bender and Fink 1995). The higher-molecular-weight species were largest in Ita-0 and Cvi-0, consistent with these transcripts arising by readthrough of the inverted repeat. *PAI* polymorphisms

TABLE 2

***PAI* cDNA analysis**

Ecotype	*PAI* gene	5′ end[a]	3′ end[a]	PAI function[b]
Col	1	−101	+958	+
Col	1	+334	+1090	ND
Col	1	+285	+1079	ND
Col	1	+85	+958	ND
Col	2	+85	ND	+
Col	2	−102	+958	ND
Col	2	−106	+971[c]	ND
Col	2	+418	+1078	ND
Col	3	−33	+950[d]	−
Ler	1	−43	+1059	+
Ler	1	−68	+958	ND
Ler	1	Spliced[e]	+958	ND
Ler	1	+177	+958	ND
Ler	2	−92	+1000	+
Ler	2	−81	+958	ND
Ler	2	−25	ND	ND
Ler	3	−67	+950[d]	−
WS	1	−42	+973[c]	+[f]
WS	1	Spliced[g]	+958	ND
WS	1	+8	+958	ND
WS	1	+33	+958	ND
WS	1	−36	+961	ND

[a] Numbering is relative to the first base (+1) of the ATG *PAI* translation initiator codon.

[b] PAI function was determined by cDNA complementation of an *E. coli* PAI-deficient mutant, as described in materials and methods.

[c] cDNA had no poly(A) tail.

[d] The *PAI3* cDNA contained an insertion of 25 bp of intron sequences most proximal to the 3′ junction of the fourth intron between the normal fourth- and fifth-exon junctions. The numbering of the 3′ end position includes the 25-bp insertion. Explicit deletion of the extra 25-bp in the Col *PAI3* cDNA rendered the gene able to complement an *E. coli* PAI-deficient mutant.

[e] Upstream sequences from −811 to −776 (based on the Col genomic sequence) were spliced to *PAI1* position −96.

[f] Introduction of the WS *PAI4* 9-bp fifth-exon deletion into the WS *PAII* cDNA rendered the gene unable to complement an *E. coli* PAI-deficient mutant.

[g] Upstream sequences from −835 to −802 and −383 to −360 (based on the WS genomic sequence) were spliced to *PAI1* position −96. The first block of upstream sequences has the same 3′ junction as the block of upstream sequences in the spliced Ler *PAI1* cDNA; the difference in numbering is caused by the length polymorphisms between Col and WS upstream regions.

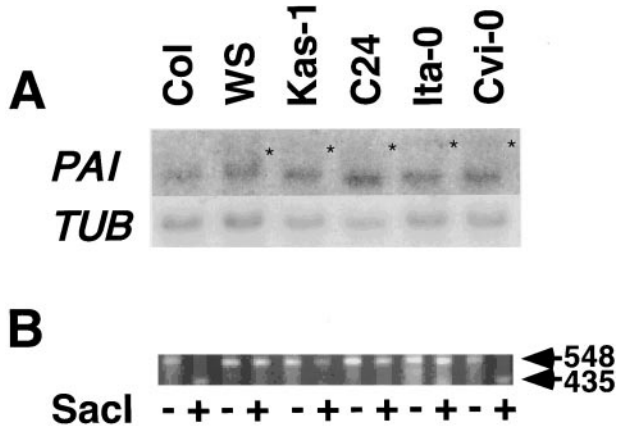

Figure 5.—*PAI1*-methylated ecotypes express *PAI* transcripts from the inverted repeat. (A) Total *PAI* steady-state transcript levels in the indicated ecotypes were determined by Northern blot with the *PAI*-internal cDNA probe. To control for loading differences, the blot was stripped and reprobed with an α-tubulin (*TUB*) probe. The position of the high-molecular-weight *PAI* RNA species in each *PAI*-methylated ecotype is marked with an asterisk. (B) The proportion of *PAI* transcripts carrying second-exon *Sac*I sites was determined by RT-PCR analysis. Total RNA from the indicated ecotypes was reverse transcribed and then amplified by PCR using primers to common sequences in all *PAI* gene first and fourth exons flanking the polymorphic *Sac*I site. These primers yield a 548-bp fragment that is cleaved by *Sac*I into 435 and 113 bp (not shown) fragments in DNA products that contain the site. Equal amounts of uncut (−) and *Sac*I-cut (+) RT-PCR product for each ecotype were resolved on a 1.5% agarose gel and visualized by ethidium bromide staining.

in the *PAI*-methylated ecotypes allowed us to determine via RT-PCR which transcripts are the predominant species detected by Northern blot. In every case, *PAI1* proved to be the only abundantly expressed *PAI* gene (see below).

PAI transcripts expressed from the *PAI1-PAI4* inverted-repeat genes can be distinguished from transcripts expressed from the singlet *PAI2* and *PAI3* genes in most *PAI*-methylated ecotypes by a restriction site polymorphism. Specifically, the *PAI1* and *PAI4* genes in the ecotypes WS, Kas-1, C24, and Ita-0 lack a conserved second exon *Sac*I site, whereas the *PAI2* and *PAI3* genes carry this site. We performed RT-PCR on total RNA from these ecotypes using primers to common *PAI* sequences in the first and fourth exons that flank the polymorphic *Sac*I site and then cleaved the PCR product with *Sac*I. This analysis showed that in WS, Kas-1, C24, and Ita-0, none of the *PAI* RT-PCR product cleaved with *Sac*I (Figure 5B), indicating that *PAI2* and *PAI3* are not expressed at significant levels in these ecotypes (estimated to be <10% of total transcripts) and that the bulk of expression comes from *PAI1* and/or *PAI4*. As a control for *Sac*I cleavage, we performed the analogous RT-PCR reaction on RNA from the Col and Cvi-0 ecotypes, where all the *PAI* genes carry the *Sac*I site. We found that in these ecotypes, the PCR product was completely cleaved.

In the ecotypes with the *PAI Sac*I polymorphism, we determined that only *PAI1* was being significantly expressed, with additional RT-PCR experiments to distinguish between *PAI1* and *PAI4* transcripts. For WS, *PAI1* and *PAI4* can be distinguished by the 9-bp deletion in the *PAI4* fifth exon. RT-PCR of this region with flanking primers to common *PAI* sequences showed that there is no detectable *PAI4*-sized transcript (Figure 6A). Therefore, this experiment confirms the result of the

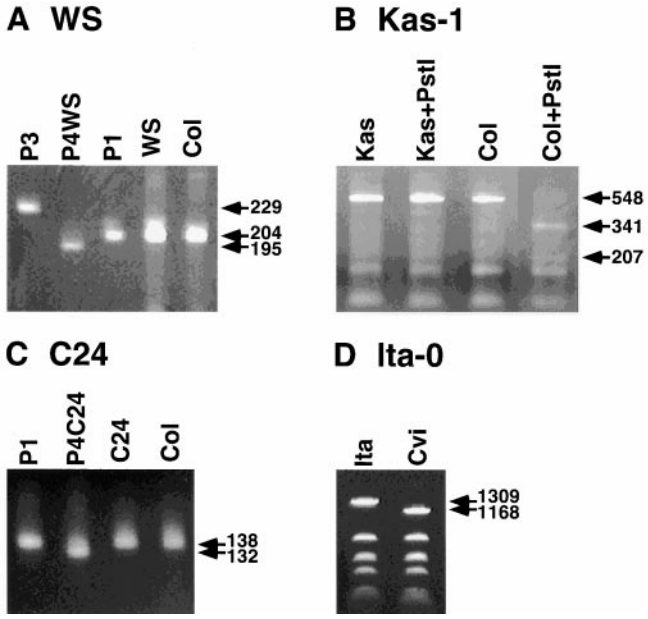

Figure 6.—*PAI1* is the predominant transcript in *PAI*-methylated ecotypes. (A) Total WS RNA was reverse transcribed and amplified by PCR with fourth- and fifth-exon primers to conserved sequences that flank a *PAI4* fifth-exon 9-bp deletion. Control fragments for *PAI1* (P1), *PAI3* (P3), and WS *PAI4* (P4WS) were amplified from cloned cDNAs. (B) The same 548-bp Kas-1 RT-PCR product used for *Sac*I analysis (Figure 5B) was cleaved with *Pst*I, which will generate 341- and 207-bp fragments from *PAI* RT-PCR products containing this site. (C) Total C24 RNA was reverse transcribed and amplified by PCR with third- and fourth-exon primers to conserved sequences that flank a *PAI4* third-exon 6-bp deletion (138 bp in *PAI1 vs.* 132 bp in *PAI4*). Control fragments for PAI1 (P1) and C24 PAI4 (P4C24) were amplified from cloned cDNAs. (D) The same Ita-0 RT-PCR product used for *Sac*I analysis was cloned into pBluescript II KS+. A representative Ita-0 subclone is shown, together with an analogous clone of Cvi-0 *PAI1*, cleaved with *Dde*I. In this assay, if the polymorphic *Dde*I site is missing (diagnostic of Ita-0 *PAI1*), then a 1309-bp band will be generated. If the polymorphic *Dde*I site is present (diagnostic of all other *PAI* genes), however, then a 1168-bp band will be generated. In all panels, DNA was visualized by ethidium bromide staining.

WS cDNA abundance analysis, that *PAI1* is the abundant transcript. The primer set used in this analysis also flanks the *PAI3* 25-bp intron insertion region between the fourth and fifth exons, but no *PAI3*-sized transcript was detected in either WS or Col. Because cDNA analysis shows that *PAI3* can be transcribed in Col (Table 2), our failure to detect this species via RT-PCR of total Col RNA suggests that our assay conditions were not sensitive enough to detect this weakly expressed gene. In particular, the *PAI3* transcript might have been discriminated against during PCR amplification because it is longer than the *PAI1* or *PAI4* species. Nonetheless, this RT-PCR experiment suggests that *PAI3* is not a major transcript in either WS or Col.

In the ecotype Kas-1, *PAI1* can be distinguished from *PAI4* by a polymorphism that destroys a conserved *Pst*I site in the third exon of *PAI1*. RT-PCR analysis of Kas-1 RNA with the same primer set used for the *Sac*I analysis, followed by cleavage of the PCR product with *Pst*I, showed that none of the Kas-1 product was cleaved (Figure 6B). As a control, the same product amplified from Col RNA, where all three *PAI* genes contain the *Pst*I site, was completely cleaved. Therefore, like WS, Kas-1 expresses transcripts primarily from *PAI1*. Furthermore, the Kas-1 *PAI4* gene has a deletion of 7 bp in the second exon, which is predicted to alter the reading frame so that the protein is terminated near the end of the putative chloroplast transit sequence (Li *et al.* 1995). Thus, regardless of expression, no functional PAI enzyme would be produced from the Kas-1 *PAI4* gene.

In the ecotype C24, *PAI1* can be distinguished from *PAI4* by a 6-bp deletion in the *PAI4* third exon. RT-PCR of this region with flanking primers to common *PAI* sequences showed that there is no detectable *PAI4*-sized transcript (Figure 6C). Furthermore, the C24 *PAI4* gene carries a 91-bp deletion extending from the middle of the fourth intron into the middle of the fifth exon, which is predicted to disrupt the correct splicing of the fourth-intron and fifth-exon coding sequences. Thus, regardless of expression, no functional PAI enzyme would be produced from the C24 *PAI4* gene.

In the ecotype Ita-0, *PAI1* can be distinguished from *PAI4* by a polymorphism that destroys a conserved *Dde*I site just upstream of the translational start in the first exon. This polymorphic site is contained in the RT-PCR product from the *Sac*I analysis shown in Figure 5B. However, it was difficult to visualize the polymorphism by direct cleavage of the RT-PCR fragment with *Dde*I followed by gel electrophoresis because the relevant cleavage products were obscured by background nonspecific PCR species. We therefore used the alternative strategy of subcloning the Ita-0 *Sac*I analysis RT-PCR products into a pBluescript II KS+ plasmid vector and analyzing individual clones by *Dde*I digest (Figure 6D). An analogous subclone of a Cvi-0 *PAI1* RT-PCR product was used as a control for the restriction pattern given by a *PAI* gene carrying the upstream *Dde*I site. All of 24 independent Ita-0 subclones tested in this way lacked the upstream *Dde*I site, indicating that in Ita-0, *PAI1* is the only abundant transcript species.

To determine which *PAI* transcripts are expressed in Cvi-0, we cloned the RT-PCR fragments generated in the *Sac*I analysis (Figure 5B) and sequenced 10 independent clones, using single-base polymorphisms unique to each *PAI* gene as a means of identification. All 10 Cvi-0 clones were *PAI1*. Therefore, in all *PAI*-methylated ecotypes, *PAI1* is the only significantly expressed *PAI* gene.

DISCUSSION

Our studies of cytosine methylation, structural polymorphisms, and gene expression indicate that an inverted repeat in the Arabidopsis genome displays a number of unusual behaviors. The *PAI1-PAI4* inverted-repeat structures analyzed here consist of ~2 kb of nearly per-

fect mirror image sequence separated by not more than 247 bp and not less than 90 bp of nonpalindromic sequence (Figure 4). The inverted repeats are densely cytosine methylated over their regions of mirror image identity, without a significant spread into neighboring sequences (Figures 2 and 3). Moreover, the inverted repeats are associated with dense methylation of the unlinked identical sequences *PAI2* and *PAI3* (Figure 2). These observations suggest that inverted repeats provide uniquely favorable substrates for methylation. The wide variation in *PAI* inverted-repeat structures observed across ecotypes (Figures 3 and 4) suggests that these structures are unusually unstable, perhaps because they are difficult to replicate accurately and/or because they are recombinationally very active. Finally, in contrast to the singlet *PAI2* gene, the inverted-repeat *PAI1* gene is not silenced by cytosine methylation, suggesting that it is relatively exposed to the transcription machinery.

Inverted repeats trigger cytosine methylation: Our original observation that the *PAI* genes are methylated in the WS ecotype but not in the Col ecotype could be explained by several models, including the difference in *PAI* gene arrangement and copy number between the two strains or strain-to-strain variation in the efficiency of the methylation machinery. The observations reported here argue that *PAI* gene arrangements, rather than variations in other loci, determine *PAI* cytosine methylation. Specifically, none of the 34 ecotypes with two or three unlinked singlet *PAI* genes displays *PAI* methylation, whereas all 7 ecotypes with a *PAI1-PAI4* inverted repeat at the *PAI1* locus display dense *PAI* methylation (Table 1). Consistent with the model that the inverted repeat locus triggers cytosine methylation, we found that when the WS inverted repeat is combined with the unmethylated Col *PAI* genes in WS × Col hybrid plants, the Col *PAI* genes become methylated *de novo* within a few generations of inbreeding (Luff *et al.* 1999).

The inverted repeat could promote cytosine methylation via DNA/DNA interactions, RNA/DNA interactions, or both. Evidence suggesting that *PAI* methylation is triggered by DNA/DNA interactions comes from two observations. First, methylation is coextensive with *PAI* DNA sequence identity, including intron and promoter sequences (Figure 3; Bender and Fink 1995; Luff *et al.* 1999). Second, a promoterless *pai1-pai4* transgene, when inserted in single copy in either the Col or the WS genome, becomes self-methylated just over its regions of mirror-image sequence within a few generations of inbreeding (Luff *et al.* 1999). We have proposed that the inverted repeat is uniquely prone to methylation because of its ability to form unusual structures, such as a four-stranded DNA "hairpin," that serve as substrates for *de novo* methylation (Bender 1998).

An alternative model is that the inverted repeat gives rise to unusual hairpin RNA molecules as a result of readthrough transcription, and that these molecules promote *PAI* cytosine methylation and silencing, perhaps because they are converted into double-stranded RNA molecules (Montgomery and Fire 1998; Waterhouse *et al.* 1998; Mette *et al.* 1999). Indeed, high-molecular-weight *PAI* transcripts that become proportionally longer with the length of the inverted-repeat region are observed on Northern blots (Figure 5A), consistent with the production of readthrough species that could lead to RNA-directed DNA methylation. Ultimately, the isolation of mutations that alter *PAI* methylation and silencing should identify the important mechanistic components.

Genesis of the *PAI* gene family: The sequence divergence between *PAI3* and its sister *PAI* genes suggests that the two classes of genes are relatively evolutionarily distant from each other. However, the close sequence identity among *PAI1*, *PAI2*, and *PAI4* argues that they arose more recently from a common progenitor. Transposon-mediated rearrangements provide a mechanism for horizontal transfer of sequences within a genome and for generation of tandem-sequence duplications. For example, transposon-generated, inverted-repeat sequence duplications have been characterized previously at the *nivea* locus in *Antirrhinum majus* (Bollmann *et al.* 1991) and at the *amylose extender1* locus in maize (Stinard *et al.* 1993). It is attractive to speculate that the transmission of a common progenitor *PAI* sequence into the *PAI2* and *PAI1-PAI4* genes in Arabidopsis was similarly transposon generated. However, sequence analysis of the regions around these genes in several ecotypes has failed to reveal any obvious transposon-like sequence. Therefore, if the *PAI1-PAI4* and/or the *PAI2* duplicate genes were transposon-generated, the transposon sequences that mediated the process most likely excised after the duplication event. Because little is known about families of transposons in Arabidopsis (Bhatt *et al.* 1998), we cannot determine whether short sequences in the *PAI1-PAI4* or *PAI2* regions are characteristic of transposon excision "footprints" or of transposon deletion derivatives.

Regardless of the mechanism(s) of *PAI* sequence duplication, it is most likely that *PAI2* on chromosome 5 was the progenitor gene that was copied to produce *PAI1* and/or *PAI1-PAI4* on chromosome 1. In particular, the 3′ sequences that are duplicated in *PAI1*, *PAI2*, and *PAI4* (but not *PAI3*) include the 3′ end of the *FAD8* gene (Gibson *et al.* 1994) that lies just downstream of *PAI2* on chromosome 5 (Figure 3). The simplest explanation of this partial *FAD8* duplication is that a region extending from the *PAI2* promoter through the 3′ end of *FAD8* was the basic sequence unit that was transmitted to the *PAI1* locus on chromosome 1. In this case, the flanking direct-repeat sequence duplication might have happened concurrently with the rearrangement event that transmitted *PAI* sequences to the *PAI1* locus.

Our analysis of structural variants of the *PAI1* locus across six ecotypes of Arabidopsis suggests that all the

variants are related and arise from a common progenitor structure. Our data support either of two models for the genesis of the *PAI1* locus. One model is that the structure found in Col and 32 other ecotypes (Table 1) is the progenitor structure and that this structure underwent a duplication and rearrangement event to generate an inverted repeat of *PAI* genes flanked by full direct-repeat sequences in one unusual lineage. This initial inverted-repeat structure then underwent further deletions and rearrangements to generate the variety of inverted-repeat structures observed in *PAI*-methylated ecotypes today. Presumably, the high degree of variation from the progenitor inverted-repeat structure was due to both the instability of the inverted repeat (Henderson and Petes 1993; Ruskin and Fink 1993) and the ability of the duplicated sequences in the structure to undergo intramolecular recombination or unequal crossing over. Molecular evolutionary studies of Arabidopsis ecotypes (Hanfstingl *et al.* 1994; Innan *et al.* 1997) do not indicate clustering among the *PAI*-methylated strains, as might be predicted by this model. However, a detailed study of sequences in the *PAI1* region of the genome across many ecotypes is needed before this point can be determined definitively. It is also possible, but unlikely, that a progenitor Col-like structure rearranged five independent times to give the various inverted-repeat structures found in Cvi-0, Ita-0, Kas-1, C24, and WS.

An alternative model is that the progenitor structure of all ecotypes was an inverted repeat of two *PAI* genes flanked by full-length direct repeats. This common structure could have given rise to the Col structure by deletion of one *PAI* gene and part of a flanking direct repeat, and to the inverted-repeat ecotype structures by other rearrangement events (Figure 3). In this scenario, the predominance of the Col structure over methylated inverted-repeat structures in the wild population (Table 1) might reflect a greater fitness of the Col structure. Because Col expresses PAI enzyme from two unlinked genes, *PAI1* and *PAI2* (Table 2), this redundancy protects Col from deleterious consequences of *PAI* gene mutations. In contrast, the *PAI*-methylated ecotypes express PAI enzyme only from the *PAI1* gene (Table 2, Figures 5 and 6), making them vulnerable to tryptophan auxotrophy via *PAI* gene mutation. This vulnerability might therefore account for the underrepresentation of *PAI*-methylated ecotypes in the wild population.

Although Col and the inverted-repeat ecotypes all have the potential to yield a deletion at the *PAI1* locus because of homologous recombination between the flanking direct-repeat sequences (Bender and Fink 1995), this type of rearrangement was observed only in one ecotype, Ll-0 (Table 1). The lack of *PAI1*-deleted strains in the wild population suggests that the homologous recombination event that generates this structure might be rare. However, given that five independent isolates of a *PAI1-PAI4*-deleted version of WS were isolated from a T-DNA-transformed population (Bender and Fink 1995), and that similar deletions occur at a high frequency in EMS-mutagenized WS populations (J. Bender, unpublished results), homologous recombination between two full-length direct repeats can happen at a relatively high frequency, at least in plants that have undergone mutagenic treatments. Alternatively, *PAI1*-deleted strains might be selected against in a wild population because they are dependent on a single gene, *PAI2*, for PAI activity; like the *PAI*-methylated ecotypes, *PAI1*-deleted variants might be underrepresented in the wild because of their vulnerability to *PAI* gene mutation. Furthermore, the reduced *PAI* expression in *PAI1*-deleted strains, although sufficient for normal plant morphology under laboratory conditions (Bender and Fink 1995), might not be sufficient for optimal survival in the wild.

In the *PAI*-methylated ecotypes Cvi-0, Ita-0, C24, Kas-1, and WS, the major structural difference is the sequence between the inverted-repeat *PAI* genes (Figures 3 and 4). Cvi-0 and Ita-0 are the only ecotypes that carry extra sequences in this central region relative to the sequences downstream of Col *PAI1* and *PAI2* (Figure 4). Other structural differences unique to particular ecotypes can best be explained as secondary events (Figure 3). For example, the WS *pai4 5′* duplication could have been generated by pairing between the direct-repeat sequences, followed by gene conversion of the sequences adjacent to the *PAI1*-proximal repeat to sequences adjacent to the *PAI4*-proximal repeat. Similarly, the C24 *PAI4* promoter rearrangement could have been generated by pairing between the direct-repeat sequences, followed by gene conversion of the sequences adjacent to the *PAI4*-proximal repeat to sequences adjacent to the *PAI1*-proximal repeat. The Ita-0 *PAI1* promoter rearrangement could have been generated by pairing between the two inverted-repeat *PAI* genes, with gene conversion of the *PAI1* promoter sequences to the *PAI4* promoter sequences. The Cvi-0 *PAI4** duplication is most simply explained by unequal crossing over between direct repeats to amplify a structure consisting of direct-repeat 1-*PAI1**–*PAI4**-direct-repeat 2-*PAI1-PAI4*-direct-repeat 3, followed by a deletion of most of direct-repeat 1 and *PAI1**.

Cytosine methylation has been shown to suppress homologous recombination (Maloisel and Rossignol 1998). Nonetheless, at least some of the structural polymorphisms in the methylated *PAI1-PAI4* regions studied here are likely to have arisen via recombination mechanisms. Perhaps even with the negative effects of cytosine methylation, the *PAI* inverted repeats can pair with each other so readily that some recombination still occurs between them. Alternatively, the inverted repeat might be immune to accumulating the factors that would normally block homologous recombination on methylated sequences because of unique structural or sequence characteristics.

Cytosine methylation and *PAI* gene expression: Cytosine methylation is usually correlated with a loss of transcription from methylated sequences (Kass *et al.* 1997). Consistent with this general rule, the singlet *PAI2* gene in WS is silenced by cytosine methylation. Specifically, in a derivative of WS lacking the *PAI1-PAI4* inverted-repeat genes, *PAI2* expression is inversely correlated with the density of methylation on the *PAI2* gene (Bender and Fink 1995; Jeddeloh *et al.* 1998). Genetic backcross experiments show that *PAI2* is also silenced in parental WS (Bender and Fink 1995). However, parental WS is phenotypically normal and expresses levels of *PAI* transcripts and PAI enzyme activity comparable to those measured in Col. Therefore, *PAI1* and/or *PAI4* must account for the bulk of PAI activity in WS despite being densely methylated. In fact, cDNA abundance (Table 2) and RT-PCR analyses (Figures 5 and 6) indicate that *PAI1* is the only detectable *PAI* transcript in WS. Furthermore, *PAI1* is the only detectable transcript in the other *PAI*-methylated ecotypes (Figures 5 and 6).

Why is the methylated *PAI1* gene expressed while the methylated *PAI2* gene is silenced? Detailed analysis of cytosine methylation patterns for WS *PAI1* and *PAI2* have revealed no significant differences in the extent or density of methylation (Luff *et al.* 1999). Therefore, the different effects of methylation on *PAI* gene expression might be caused by differences in the genomic contexts (*cis*-acting sequences, DNA structure, and/or chromosomal domains) that determine different chromatin assembly and expression states. Interestingly, the *PAI4* gene immediately adjacent to *PAI1* is not expressed in *PAI*-methylated ecotypes (Figure 6) despite ~250 bp (or in the case of Ita-0, ~550 bp) of promoter identity. The lack of *PAI4* expression could result from the absence of critical upstream promoter sequences and/or from silencing by methylation. The later case would suggest that the transcriptionally active state at the *PAI1* promoter does not propagate as far as the *PAI4* promoter, 4 kb away.

We thank Laura Pawlowski, Gromoslaw Smolen, and Tomoko Hamma for technical assistance. This work was supported by a Basil O'Connor Starter Scholar Award 5-FY98-0535 from the March of Dimes, a National Institute of Environmental Health Sciences Training Grant to S.M. (ES 07141), and a Searle Scholars Award 97-E-103 to J.B.

LITERATURE CITED

Ausubel, F. M., R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman *et al.*, 1989 *Current Protocols in Molecular Biology*. Greene Publishing Associates and Wiley-Interscience, New York.

Bender, J., 1998 Cytosine methylation of repeated sequences in eukaryotes: the role of DNA pairing. Trends Biochem. Sci. **23:** 252–256.

Bender, J., and G. R. Fink, 1995 Epigenetic control of an endogenous gene family is revealed by a novel blue fluorescent mutant of Arabidopsis. Cell **83:** 725–734.

Bhatt, A. M., C. Lister, N. Crawford and C. Dean, 1998 The transposition frequency of *Tag1* elements is increased in transgenic Arabidopsis lines. Plant Cell **10:** 427–434.

Bollmann, J., R. Carpenter and E. S. Coen, 1991 Allelic interactions at the *nivea* locus of *Antirrhinum*. Plant Cell **3:** 1327–1336.

Bonham-Smith, P. C., and M. M. Moloney, 1994 Nucleotide and protein sequences of a cytoplasmic ribosomal protein *S15a* gene from *Arabidopsis thaliana*. Plant Physiol. **106:** 401–402.

Church, G. M., and W. Gilbert, 1984 Genomic sequencing. Proc. Natl. Acad. Sci. USA **81:** 1991–1995.

Elledge, S. J., J. T. Mulligan, S. W. Ramer, M. Spottswood and R. D. Davis, 1991 λYES: a multifunctional cDNA expression vector for the isolation of genes by complementation of yeast and *Escherichia coli* mutations. Proc. Natl. Acad. Sci. USA **88:** 1731–1735.

Gibson, S., V. Arondel, K. Iba and C. Somerville, 1994 Cloning of a temperature-regulated gene encoding a chloroplast omega-3 desaturase from *Arabidopsis thaliana*. Plant Physiol. **106:** 1615–1621.

Hanfstingl, U., A. Berry, E. A. Kellogg, J. T. Costa III, W. Rudiger *et al.*, 1994 Haplotypic divergence coupled with lack of diversity at the *Arabidopsis thaliana* alcohol dehydrogenase locus: roles for both balancing and directional selection? Genetics **138:** 811–828.

Henderson, S. T., and T. D. Petes, 1993 Instability of a plasmid-borne inverted repeat in *Saccharomyces cerevisiae*. Genetics **134:** 57–62.

Innan, H., R. Terauchi and N. T. Miyashita, 1997 Microsatellite polymorphism in natural populations of the wild plant *Arabidopsis thaliana*. Genetics **146:** 1441–1452.

Jeddeloh, J. A., J. Bender and E. J. Richards, 1998 The DNA methylation locus *DDM1* is required for maintenance of gene silencing in *Arabidopsis*. Genes Dev. **12:** 1714–1725.

Kass, S. U., D. Pruss and A. P. Wolffe, 1997 How does DNA methylation repress transcription? Trends Genet. **13:** 444–449.

Kunkel, T. A., J. D. Roberts and R. A. Zakour, 1987 Rapid and efficient site-specific mutagenesis without phenotypic selection. Methods Enzymol. **154:** 367–382.

Last, R. L., P. H. Bissinger, D. J. Mahoney, E. R. Radwanski and G. R. Fink, 1991 Tryptophan mutants in *Arabidopsis*: the consequences of duplicated tryptophan synthase β genes. Plant Cell **3:** 345–358.

Li, J., J. Zhao, A. B. Rose, R. Schmidt and R. L. Last, 1995 Arabidopsis phosphoribosylanthranilate isomerase: molecular genetic analysis of triplicate tryptophan pathway genes. Plant Cell **7:** 447–461.

Luff, B., L. Pawlowski and J. Bender, 1999 An inverted repeat triggers cytosine methylation of identical sequences in *Arabidopsis*. Mol. Cell **3:** 505–511.

Maloisel, L., and J.-L. Rossignol, 1998 Suppression of crossing-over by DNA methylation in *Ascobolus*. Genes Dev. **12:** 1381–1389.

Mette, M. F., J. Van Der Winden, M. A. Matzke and A. J. M. Matzke, 1999 Production of aberrant promoter transcripts contributes to methylation and silencing of unlinked homologous promoters *in trans*. EMBO J. **18:** 241–248.

Minet, M., M.-E. Dufour and F. Lacroute, 1992 Complementation of *Saccharomyces cerevisiae* auxotrophic mutants by *Arabidopsis thaliana* cDNAs. Plant J. **2:** 417–422.

Montgomery, M. K., and A. Fire, 1998 Double-stranded RNA as a mediator in sequence-specific genetic silencing and co-suppression. Trends Genet. **14:** 255–258.

Nagy, F., S. A. Kay and N.-H. Chua, 1988 Analysis of gene expression in transgenic plants, pp. B4/11–B4/12 in *Plant Molecular Biology Manual*, edited by S. B. Gelvin and R. A. Schilperoort. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Niyogi, K. K., and G. R. Fink, 1992 Two anthranilate synthase genes in Arabidopsis: defense-related regulation of the tryptophan pathway. Plant Cell **4:** 721–733.

Niyogi, K. K., R. L. Last, G. R. Fink and B. Keith, 1993 Suppressors of *trp1* fluorescence identify a new Arabidopsis gene, *TRP4*, encoding the anthranilate synthase β subunit. Plant Cell **5:** 1011–1027.

Ruskin, B., and G. R. Fink, 1993 Mutations in *POL1* increase the mitotic instability of tandem inverted repeats in *Saccharomyces cerevisiae*. Genetics **134:** 43–56.

Sato, S., H. Kotani, Y. Nakamura, T. Kaneko, E. Asamizu *et al.*, 1997 Structural analysis of *Arabidopsis thaliana* chromosome 5. I. Sequence features of the 1.6 Mb regions covered by twenty physically assigned P1 clones. DNA Res. **4:** 215–230.

Stinard, P. S., D. S. Robertson and P. S. Schnable, 1993 Genetic isolation, cloning, and analysis of a *Mutator*-induced, dominant antimorph of the maize *amylose extender1* locus. Plant Cell **5:** 1555–1566.

Waterhouse, P. M., M. W. Graham and M.-B. Wang, 1998 Virus resistance and gene silencing in plants can be induced by simultaneous expression of sense and antisense RNA. Proc. Natl. Acad. Sci. USA **95:** 13959–13964.

Communicating editor: J. A. Birchler