A Bifunctional Locus (*BIO3-BIO1*) Required for Biotin Biosynthesis in Arabidopsis[1][W][OA]

Rosanna Muralla, Elve Chen, Colleen Sweeney, Jennifer A. Gray, Allan Dickerman, Basil J. Nikolau, and David Meinke*

Department of Botany, Oklahoma State University, Stillwater, Oklahoma 74078 (R.M., C.S., D.M.); Department of Biochemistry, Biophysics, and Molecular Biology, Iowa State University, Ames, Iowa 50011 (E.C., J.A.G., B.J.N.); and Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061 (A.D.)

We identify here the Arabidopsis (*Arabidopsis thaliana*) gene encoding the third enzyme in the biotin biosynthetic pathway, dethiobiotin synthetase (*BIO3*; At5g57600). This gene is positioned immediately upstream of *BIO1*, which is known to be associated with the second reaction in the pathway. Reverse genetic analysis demonstrates that *bio3* insertion mutants have a similar phenotype to the *bio1* and *bio2* auxotrophs identified using forward genetic screens for arrested embryos rescued on enriched nutrient medium. Unexpectedly, *bio3* and *bio1* mutants define a single genetic complementation group. Reverse transcription-polymerase chain reaction analysis demonstrates that separate *BIO3* and *BIO1* transcripts and two different types of chimeric *BIO3-BIO1* transcripts are produced. Consistent with genetic data, one of the fused transcripts is monocistronic and encodes a bifunctional fusion protein. A splice variant is bicistronic, with distinct but overlapping reading frames. The dual functionality of the monocistronic transcript was confirmed by complementing the orthologous auxotrophs of *Escherichia coli* (*bioD* and *bioA*). *BIO3-BIO1* transcripts from other plants provide further evidence for differential splicing, existence of a fusion protein, and localization of both enzymatic reactions to mitochondria. In contrast to most biosynthetic enzymes in eukaryotes, which are encoded by genes dispersed throughout the genome, biotin biosynthesis in Arabidopsis provides an intriguing example of a bifunctional locus that catalyzes two sequential reactions in the same metabolic pathway. This complex locus exhibits several unusual features that distinguish it from biotin operons in bacteria and from other genes known to encode bifunctional enzymes in plants.

Biotin is a vitamin that functions as an enzyme cofactor in cellular metabolism to facilitate CO_2 transfer during carboxylation and decarboxylation reactions. Biosynthesis of biotin from pimeloyl-CoA and Ala, first elucidated in bacteria more than 40 years ago, occurs through four reactions that result in the sequential production of 7-keto-8-aminopelargonic acid (KAPA), 7,8-diaminopelargonic acid (DAPA), dethiobiotin (DTB), and ultimately biotin. In *Escherichia coli*, four genes that encode these enzymes (*bioF*, *bioA*, *bioD*, *bioB*) are clustered into an operon whose structure and function has been examined in detail (DeMoll, 1996). Biosynthesis of biotin in plants occurs through a similar pathway but is divided between two compartments. The initial production of KAPA occurs in the cytosol (Pinon et al., 2005), whereas the final conversion of DTB to biotin occurs in mitochondria (Weaver et al., 1996; Baldet et al., 1997; Picciocchi et al., 2003; Arnal et al., 2006). Intracellular localization of the intermediate reactions remains unresolved (Rébeillé et al., 2007). Metabolic enzymes that require biotin as a cofactor are located in four different compartments: chloroplasts, mitochondria, protein bodies, and the cytosol (Nikolau et al., 2003). Plants must therefore possess transport mechanisms for delivering biotin and related intermediates to their proper locations in the cell.

Two auxotrophic mutants of Arabidopsis (*Arabidopsis thaliana*) have played an important role in the analysis of biotin biosynthesis in plants. The *bio1-1* mutant was isolated following a forward genetic screen designed to identify embryo-defective (*emb*) mutants in which arrested embryos were rescued on an enriched nutrient medium (Schneider et al., 1989). Aborted seeds from heterozygous siliques contain reduced levels of biotin, consistent with a defect in biotin synthesis (Shellhammer and Meinke, 1990). Embryo rescue experiments and subsequent complementation with the *bioA* ortholog from *E. coli* demonstrated that mutant embryos are defective in the conversion of KAPA to DAPA (Schneider et al., 1989; Patton et al., 1996). The *bio2-1* mutant was isolated through a similar genetic screen for embryo defectives and was shown to be disrupted in the final reaction of the pathway (Patton et al., 1998). A second allele (*bio2-2*) with an insertion identified through reverse genetics has recently been described (Arnal et al., 2006). Because the original *bio2-1* mutant contains a

[1] This work was supported by the National Science Foundation Metabolic Biochemistry Program (grant no. 0416730 to B.J.N.) and Arabidopsis 2010 Program (grants no. 0114866 and no. 0618166 to D.M.).

* Corresponding author; e-mail meinke@okstate.edu.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: David Meinke (meinke@okstate.edu).

[W] The online version of this article contains Web-only data.

[OA] Open Access articles can be viewed online without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.107.107409

deletion that includes an adjacent gene (*FPA*) required for flowering (Schomburg et al., 2001), rescued *bio2-1* homozygotes produce giant rosettes under long days. Rescued *bio1-1* and *bio2-2* homozygotes, in contrast, appear normal when supplemented with biotin. A complete list of Arabidopsis genes and mutants involved in biotin synthesis is presented in Table I.

Several years ago, we initiated a large-scale T-DNA insertional mutagenesis project with colleagues at Syngenta that was designed to identify genes required for embryo development in Arabidopsis (McElver et al., 2001; Tzafrir et al., 2004). Two additional alleles of *bio1* uncovered through that forward genetic screen are described in this article. We also began to pursue reverse genetic approaches to identify *EMB* genes missed through forward genetics. One approach was to focus on nonredundant genes associated with metabolic pathways that were known to be required for embryo development. The effectiveness of this strategy is illustrated by the recent identification of multiple His auxotrophs defective in embryo development (Muralla et al., 2007). In addition, we focused again on the biotin pathway and the *bioD* ortholog (*BIO3*) required for the conversion of DAPA to DTB. This eventually led to the unexpected discovery that *BIO3* and *BIO1* are positioned adjacent to each other on the chromosome, in the same orientation as found in a variety of microorganisms, and that differential splicing results in production of two types of full-length transcripts, one with the potential to encode separate proteins and the other capable of producing a bifunctional fusion protein. We document here the structure and function of this unusual locus, provide indirect evidence that both of the corresponding enzymatic reactions take place within mitochondria, examine related genes and transcripts from other plants and fungi, and present the results of a genome-wide scan for similar types of complex loci associated with metabolic pathways in Arabidopsis.

RESULTS

Molecular Identification of the *BIO1* Locus

Three candidate *BIO1* genes were identified based on the genetic map location of the *bio1-1* mutant allele (Patton et al., 1991) and BLASTP searches of the Arabidopsis genome queried with the BioA ortholog from *E. coli*. Two of these candidates, At5g46180 and At5g63570, were eliminated from consideration because they encode known enzymes, Orn-δ-aminotransferase (Roosens et al., 1998) and chloroplastic Glu-1-semialdehyde 2,1-aminomutase (Ilag et al., 1994). The remaining candidate, At5g57590, is predicted to encode an aminotransferase class III protein that shares 26% sequence identity with BioA. Experimental confirmation that this gene corresponds to *BIO1* was obtained by sequencing PCR-amplified genomic fragments from rescued *bio1-1* homozygotes grown on biotin. A single nucleotide substitution (G to A) that modifies the 3′ acceptor site of the final intron was found in mutant plants. To ensure that this polymorphism was not due to a sequencing error, we amplified and sequenced this genomic region from progeny plants derived from a single heterozygote. Three different genotypes of plants were found at the expected frequencies in this population. Heterozygotes and homozygotes exhibited the predicted polymorphism (A and G) at the mutation site (Fig. 1, A–C). These results confirmed that At5g57590 corresponds to the *BIO1* gene.

Characterization of Additional *bio1* Mutant Alleles

Two embryo-defective mutants identified through a forward genetic screen of T-DNA insertion lines generated at Syngenta (McElver et al., 2001) were found to contain insertions in the *BIO1* region. Genetic complementation tests revealed that both mutants (*bio1-2* and *bio1-3*) were allelic to the original ethyl methanesulfonate (EMS) allele (*bio1-1*). Mutant embryos from immature siliques of *bio1-2* heterozygotes were rescued in culture on DAPA, DTB, or biotin, consistent with the results of previous experiments with *bio1-1* (Schneider et al., 1989; Shellhammer, 1991). Mutant embryos from both insertion lines exhibited a range of phenotypes similar to *bio1-1* and less severe than either *bio2-1* or *bio2-2*.

The locations of mutation sites in *bio1* mutant alleles in relation to different annotated versions of the *BIO3-BIO1* locus are presented in Figure 2. The *bio1-2* mutant

Table I. *Biotin biosynthetic genes and auxotrophic mutants of Arabidopsis*[a]

Arabidopsis Gene	Bacterial Ortholog	Arabidopsis Locus	Enzymatic Product	Allele	Mutagen	Line No.	Reference on Mutant
BIO4	*bioF*	At5g04620	KAPA	NA[b]	NA[b]	NA[b]	None identified
BIO1	*bioA*	At5g57590	DAPA	*bio1-1*	EMS	122G-E	Schneider et al. (1989)
				bio1-2	T-DNA	36172	www.SeedGenes.org[c]
				bio1-3	T-DNA	46455	www.SeedGenes.org[c]
BIO3	*bioD*	At5g57600	DTB	*bio3-1*	T-DNA	RATM53-2665-1G	This article
				bio3-2	T-DNA	RATM53-3000-1G	This article
				bio3-3	T-DNA	SALK_023399	This article
BIO2	*bioB*	At2g43360	Biotin	*bio2-1*	EMS	*emb49*	Patton et al. (1998)
				bio2-2	T-DNA	GABI_100C11	Arnal et al. (2006)

[a] Genes are listed in order of their function in the pathway. [b] Not applicable (NA) because no mutants have been identified. [c] Refer also to McElver et al. (2001) and Tzafrir et al. (2004).

represents a putative null allele because the insertion is located within an exon in the middle of the *BIO1* coding region. Flanking sequences obtained from both sides of the insert revealed a small deletion associated with the insertion (Supplemental Fig. S1). The point mutation in *bio1-1* results in a longer transcript but a shorter open reading frame (ORF), which leads to a defective protein lacking the normal C terminus (Fig. 1D). Based on a comparison of mutant phenotypes, this altered protein appears to retain little BIO1 function. The *bio1-3* insertion is located downstream of the *BIO1* coding region, but the precise location remains unresolved because flanking sequences obtained from both sides of the insert gave contradictory information (Supplemental Fig. S1). The failure of *bio1-3* to complement either *bio1-1* or *bio1-2* in genetic crosses nevertheless confirms that *BIO1* function in this mutant is disrupted.

Isolation and Characterization of *bio3* Mutant Alleles

A candidate *BIO3* gene (At5g57600) was identified in the Arabidopsis genome based on sequence homology to the BioD protein of *E. coli*. Three insertion lines that disrupted the coding region were obtained from the Arabidopsis Biological Resource Center (*bio3-3*) and the RIKEN Bioresource Center in Japan (*bio3-1* and *bio3-2*). All three lines generated heterozygous plants that produced siliques with approximately 25% aborted seeds. Linkage between the T-DNA insert and mutant phenotype was confirmed by PCR genotyping of individual plants. Genetic complementation tests demonstrated that all three mutants are allelic (Table II). Phenotypes of *bio3*-arrested embryos are similar to *bio1* alleles and less severe than *bio2* alleles (Table III). Mutant embryos are pale and typically block at the transition to cotyledon stages of development. All three *bio3* mutants are likely to be nulls based on insert locations. It therefore appears that interfering with the initial reactions in biotin synthesis, catalyzed by BIO1 and BIO3, is less detrimental to embryo development than elimination of the final (BIO2) step. One possible explanation is that maternal supplies of DAPA and DTB may contribute somewhat to continued development of *bio1*/*bio1* and *bio3*/*bio3* embryos.

Mutant embryos from parental *bio1* and *bio3* heterozygotes were rescued by watering plants with biotin (Table IV). Progeny seedlings derived from rescued siliques exhibited the expected 1:2:1 ratio of genotypes (wild type, heterozygote, and homozygote). Responses of mutant embryos in culture are illustrated in Figure 3 and Table V. Immature embryos from *bio3* heterozygotes were fully rescued on DTB and biotin but not on DAPA, consistent with the predicted role of BIO3 in biotin synthesis. Biological activity of the DAPA used in culture experiments was confirmed by successful rescue of *bio1* mutant embryos. Failure to rescue *bio2* embryos indicated that DAPA stocks were not contaminated with biotin. We conclude from these experiments that *bio3* mutants of Arabidopsis are defective in the conversion of DAPA to DTB.

Allelism between *bio1* and *bio3* Heterozygotes

An unexpected result was obtained when genetic complementation tests were performed between *bio1* and *bio3* heterozygotes (Table II). In every combination examined, mutants failed to complement, suggesting that a single gene was disrupted. Because these results were initially analyzed without knowledge of the types of transcripts produced, we reasoned that T-DNA insertions in *BIO3* might be reducing expression of the downstream *BIO1* gene. We then attempted to locate EMS mutations in the *BIO3* coding region by searching the Arabidopsis TILLING database (Henikoff et al., 2004) with the hope that point mutations would have

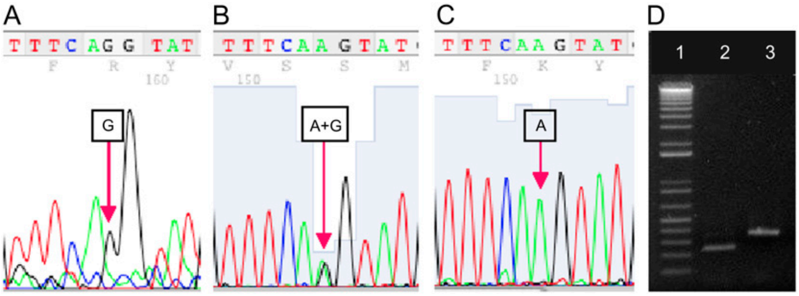

Figure 1. *BIO1* gene identification and nature of the point mutation in *bio1-1*. The 3′ end of the last intron of At5g57590 in wild-type plants (TTTCAG) is modified in *bio1-1* homozygotes (TTTCAA). The 5′ end of the last exon (GTAT) remains unchanged. Refer to Supplemental Figure S1 for additional details on the location of this sequence polymorphism. A, Sequencing of genomic DNA from wild-type plants reveals a G nucleotide at the mutation site. B, Genomic DNA from heterozygotes yields a doublet peak that results from the expected mixture of A and G nucleotides at the mutation site. C, Rescued homozygotes exhibit a single peak, consistent with the G to A substitution. D, RT-PCR products obtained from this region demonstrate that transcripts from rescued homozygotes (lane 3) are longer than normal because they include the final intron (confirmed by sequencing) not found in transcripts from wild-type plants (lane 2). The five smallest bands in the DNA ladder (lane 1) range from 100 to 500 nucleotides.

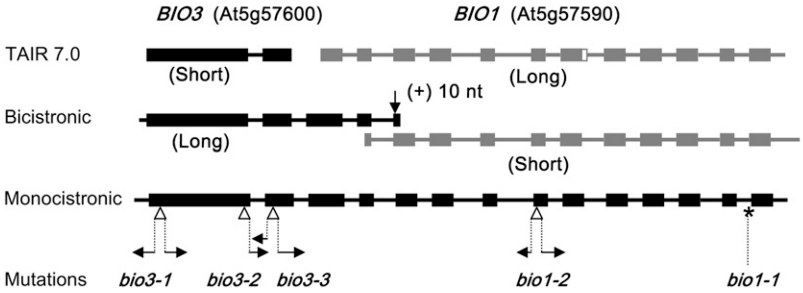

Figure 2. Genome annotation and mutation sites for the *BIO3-BIO1* region. Two separate genes (*BIO3* short and *BIO1* long) are predicted at TAIR (www.arabidopsis.org). Bicistronic cDNA has the potential to encode two distinct proteins (BIO3 long and BIO1 short). Monocistronic cDNA contains a single ORF that encodes a bifunctional fusion protein. Monocistronic (−10) and bicistronic (+10) transcripts differ with respect to the presence or absence of 10 nucleotides (arrow) at the end of intron 4. White triangles represent insertion sites for T-DNA mutant alleles. Horizontal arrows designate the locations of flanking sequences obtained. The asterisk marks a single nucleotide substitution that disrupts splicing of the last intron in the *bio1-1* allele. The final 33 nucleotides of *BIO1* exon 7 in the TAIR model (white rectangle) are differentially spliced and are not present in the full-length cDNAs.

more limited effects. Unfortunately, no candidate mutations in the appropriate region were identified. We therefore concluded, based on genetic evidence alone, that *BIO1* and *BIO3* define a single genetic locus.

BIO1 and *BIO3* Define a Single Locus That Exhibits Differential Splicing

Molecular evidence in support of a single chimeric locus was obtained by characterizing cDNAs derived from this region of the genome. The Arabidopsis EST clone RZ128g09R (GenBank accession no. AV551591) was first sequenced and found to match the exon structure of At5g57590. Sequencing of a 3′-RACE product derived from this cDNA identified a 121-nucleotide 3′-untranslated region (UTR). A 5′-RACE product was then sequenced and found to contain a single ORF that included both *BIO3* (At5g57600) and *BIO1* (At5g57590). The 78-nucleotide 5′-UTR was later confirmed in a cap-dependent RACE experiment (Maruyama and Sugano, 1994). These combined results demonstrate the existence of monocistronic, full-length cDNA (GenBank accession no. EU089963) capable of encoding a single fusion protein (833 amino acids) with potentially two different catalytic activities.

Two additional full-length cDNAs spanning the *BIO3-BIO1* locus were found in public databases: RAFL22-07-J07 (Seki et al., 2002) and BX842298 (Castelli et al., 2004). The RAFL22 clone is bicistronic and contains separate *BIO3* and *BIO1* ORFs. The BX842298 clone includes small insertions and deletions that disrupt the *BIO1* ORF. Whether these single nucleotide polymorphisms reflect true differences in transcripts or represent artifacts of sequencing remains unresolved. Another bicistronic cDNA was identified in the Nikolau laboratory (GenBank accession no. EU090805). Sequence alignments revealed a 10-nucleotide region in the bicistronic clones that is missing in the monocistronic clone (Supplemental Fig. S2). This short sequence (5′-GCTGTTTCAG-3′) provides an alternative 3′-splice acceptor site that corresponds to the end of intron 4 in monocistronic (−10) transcripts and the start of exon 5 in bicistronic (+10) transcripts.

Four different ORFs can be identified within this region based on gene models and cDNA sequences (Figs. 2 and 4). The *BIO3* (long) ORF found in the bicistronic clones terminates right after the (+10) sequence. The Arabidopsis Information Resource (TAIR) 7.0 annotation of the *BIO3* (short) ORF utilizes an upstream stop codon that is predicted in other models to be part of intron 2. The resulting protein is not likely to be functional because it lacks a region conserved in orthologs from a variety of microorganisms. The TAIR 7.0 annotation of *BIO1* (long) requires that a (−10) transcript be produced. In contrast, the BIO1 (short) protein encoded by the bicistronic transcript requires a (+10) transcript. This shortened protein also lacks conserved regions shared among microorganisms. The long BIO1 and BIO3 proteins are therefore most likely to be functional, if they are indeed produced.

Table II. *Results of genetic complementation tests*[a]

Female Parent	Male Parent	Siliques Screened	Results Obtained
bio1-2	*bio1-3*	5	Allelic
bio1-1	*bio3-1*	4	Allelic
bio1-1	*bio3-2*	4	Allelic
bio1-1	*bio3-3*	2	Allelic
bio3-3	*bio1-1*	3	Allelic
bio1-2	*bio3-1*	3	Allelic
bio1-2	*bio3-2*	3	Allelic
bio1-2	*bio3-3*	5	Allelic
bio3-3	*bio1-2*	7	Allelic
bio3-1	*bio3-2*	4	Allelic
bio3-2	*bio3-1*	4	Allelic
bio3-3	*bio3-1*	2	Allelic
bio3-3	*bio3-2*	2	Allelic

[a] Crosses were performed between heterozygotes and the resulting siliques were screened for aborted seeds prior to maturity.

Table III. *Phenotypes of arrested embryos recovered from mutant seeds*[a]

Mutant Allele	Mutant Seeds	Average Embryo Size	Phenotypic Classes of Mutant Embryos			
			Preglobular	Globular	Transition	Cotyledon
	%	$\mu m \pm$ (SD)				
bio1-1	23.1	270 (80)	0	1	2	97
bio1-2	27.0	300 (100)	2	9	0	89
bio1-3	34.8	320 (90)	0	9	4	87
bio3-1	23.5	210 (80)	0	6	29	65
bio3-2	23.8	190 (60)	0	3	36	61
bio3-3	24.3	230 (110)	9	12	13	66
bio2-1	27.6	50 (40)	22	64	5	9

[a]Embryos were removed from 100 aborted seeds and classified as described at www.seedgenes.org.

Because evidence of a monocistronic, full-length transcript was at first limited to a single 5′-RACE experiment, we designed additional reverse transcription (RT)-PCR primers capable of distinguishing between (+10) and (−10) mRNAs. The results obtained (Fig. 5) confirmed that significant amounts of both types of transcripts are present. The (+10) version is somewhat more abundant than the (−10) version in most parts of the plant. Although the (+10) sequence can be found in both the bicistronic full-length transcript and *BIO1* single gene transcripts, the (−10) sequence appears to be limited to monocistronic full-length transcripts. Arabidopsis therefore has the potential to produce a full-length *BIO3-BIO1* transcript that encodes a bifunctional fusion protein capable of catalyzing two sequential reactions in biotin biosynthesis.

Organization of *BIO1* and *BIO3* Orthologs in Flowering Plants

Further evidence of differential splicing and the presence of a monocistronic transcript encoding a bifunctional protein was obtained by searching GenBank for homologous sequences from other plant species that spanned the junction region. Two different types of rice (*Oryza sativa*) transcripts were identified. A full-length cDNA (accession no. AK100945) and EST (accession no. AU0033128) confirm the presence of a monocistronic transcript. Another full-length cDNA (accession no. AK241284) and EST (accession no. CT857795) provide evidence for an alternative splice variant that does not encode either a fusion protein or a functional BIO1 protein. The only source of BIO1 activity in rice therefore appears to be the bifunctional protein. The main difference between the two types of transcripts is a region 37 nucleotides in length that provides an alternative 3′-acceptor site for splicing. This rice sequence (5′-gcaatttttgtagcctaaatttctctttgctcattag-3′) aligns in part with the (+10) region from Arabidopsis. Monocistronic transcripts were also identified from EST databases with a TBLASTN search using a query (WWTQGPDPTFQAELAREMGY) based on the junction region of the predicted Arabidopsis bifunctional protein. This search identified ESTs from snapdragon (*Antirrhinum majus*; accession no. AJ788704), Jerusalem artichoke (*Helianthus tuberosus*; accession no. EL452781), and barley (*Hordeum vulgare*; accession no. CA029744) that appeared to encode a bifunctional protein. We have therefore found evidence to support the widespread occurrence of transcripts capable of producing a bifunctional DAPA synthase/DTB synthetase protein in a variety of plants.

Organization of Biotin Biosynthetic Genes in Microorganisms

Several recent studies have surveyed the organization of biotin biosynthetic genes in microorganisms (Rodionov et al., 2002; Streit and Entcheva, 2003; Hall and Dietrich, 2007). We focused on the identification of *BIO3* and *BIO1* orthologs in bacteria and fungi to search for additional evidence of a bifunctional fusion

Table IV. *Rescue of mutant embryos in siliques of heterozygous plants supplemented with 1 mM biotin*[a]

Mutant	Plants Rescued	Progeny Seedlings Transplanted	Total Seedlings Genotyped	WT	HET	HMZ
bio3-1	7	111	71	14	40	17
bio3-2	6	121	90	23	38	29
bio1-2	6	131	89	17	46	26
Total	19	363	250	54	124	72

[a]Rescued heterozygotes were identified by the absence of aborted seeds in developing siliques. Mature seeds from three rescued siliques (per mutant) were germinated on agar plates containing 0.1 μM biotin. The resulting seedlings, which all appeared normal, were transplanted to pots watered with 100 μM biotin and PCR genotyped. WT, Wild type; HET, heterozygote; HMZ, homozygote.

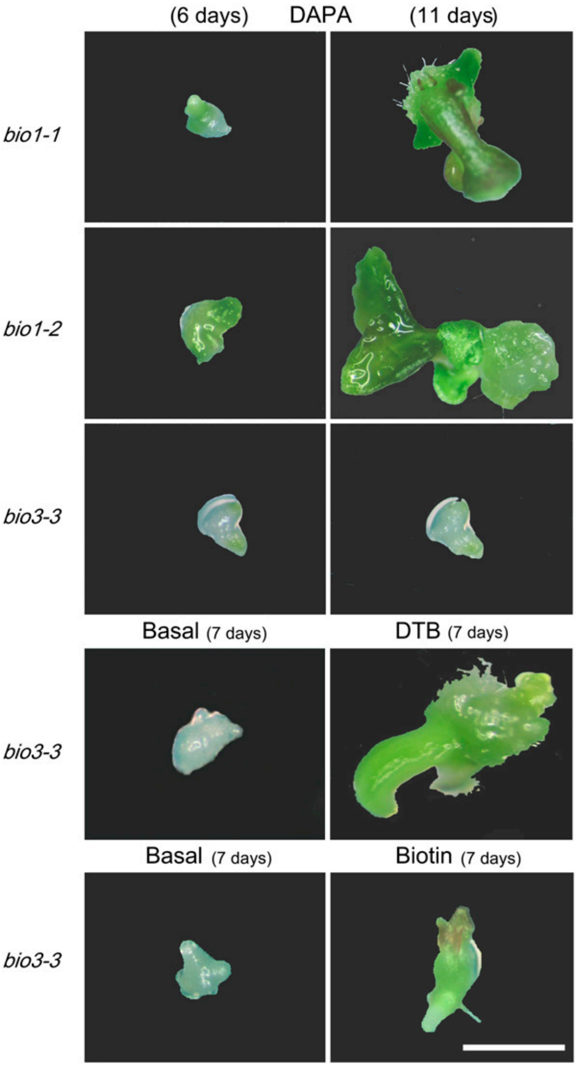

Figure 3. Responses of mutant embryos in culture. Immature embryos were removed from heterozygous siliques, plated on agar medium containing the supplements noted, and observed after the specified number of days in culture. Top, Dual images of three embryos at two different time points. DAPA, a biotin intermediate, was expected to rescue *bio1* embryos but not *bio3* embryos. DTB, a later intermediate, was expected to rescue *bio3* embryos. See Table V for additional details. Scale bar = 1 mm.

protein and determine whether this region of the Arabidopsis genome might represent a novel remnant of a bacterial operon. Figure 6 illustrates some of the different types of gene organization encountered. The *BIO3* ortholog (*bioD*) is located just upstream of *bioA* (*BIO1*) and in the same orientation in a wide range of bacteria, including *Agrobacterium tumefaciens*, *Staphylococcus aureus*, and *Bacillus sphaericus*. Twenty examples of this pattern of organization are included among the 90 eubacterial and archaeal genomes characterized by Rodionov et al. (2002). However, no evidence of a BioD-BioA fusion protein can be found among these sequenced genomes or in public databases (www.igs.cnrs-mrs.fr/FusionDB) of prokaryotic gene fusion events (Suhre and Claverie, 2004). This novel feature of biotin gene organization and enzyme function therefore appears to be limited to eukaryotes.

BIO3 and *BIO1* orthologs are adjacent but oriented in opposite directions in yeast (*Saccharomyces cerevisiae*). Separate orthologs are also found in a variety of hemiascomycete fungi, including *Candida albicans* but not *Yarrowia lipolytica*. A different situation is encountered in the basidiomycetes and filamentous fungi. Evidence of a single ORF encoding a bifunctional protein can be found in at least 18 different species, including *Aspergillus nidulans*, *Ustilago maydis*, and *Cryptococcus neoformans*. This list includes data from Hall and Dietrich (2007) and additional examples obtained from BLASTP searches at GenBank using Arabidopsis and *Aspergillus* fusion proteins as queries. A significant match was also found with a protein from the sequenced genome of the green alga, *Ostreococcus tauri*. The existence of a bifunctional BIO3-BIO1 fusion protein in flowering plants is therefore supported by extensive sequence data from lower eukaryotes.

Functional Complementation of *E. coli* Biotin Auxotrophs

To assess the functions of different *BIO3-BIO1* gene products, Arabidopsis proteins encoded by the monocistronic (−10) and bicistronic (+10) full-length cDNAs were produced in *E. coli* using a Gateway expression vector (pDEST17) that fused an N-terminal 6×-His tag to each recombinant protein. This added about 3 kD to the molecular mass of each product. The (−10) construct was therefore expected to produce a 95-kD BIO3-BIO1 fusion protein and the (+10) version a 48-kD BIO3 protein. Plasmid DNA from the expression clones was transformed into *E. coli* and evaluated for expression of targeted proteins (Fig. 7, A and B). Expression of the (+10) construct in *E. coli* strain BL21(λDE3) resulted in accumulation of the BIO3 protein (48 ± 4 kD). Expression of the (−10) construct in strain C41(λDE3) (Miroux and Walker, 1996) generated a fusion protein (95 ± 8 kD). Both polypeptides uniquely reacted with anti-His tag antibodies (data not shown). The failure of *E. coli* cells carrying the (+10) construct to accumulate BIO1 (short) protein suggests that bacterial ribosomes are unable to initiate translation effectively at the required internal AUG site.

Functional properties of recombinant proteins were evaluated by introducing each construct into *E. coli* strains *bioD* (JW0761) and *bioA* (JW0757) obtained from the Keio collection of single gene knockouts (Baba et al., 2006). We also examined the *ynfK* knockout (JW5264) because this gene encodes a DTB synthase-like protein that shares 50% sequence identity with BioD. The *bioD* and *bioA* mutants exhibited the expected biotin requirement for growth, whereas the *ynfK* mutant grew on basal medium (Fig. 7, C and D). We therefore used only the *bioD* and *bioA* mutants for subsequent complementation studies.

Table V. *Response of mutant embryos cultured on DAPA*[a]

Genotype	Embryos Cultured	Extent of Embryo Response in Culture				
		A	B	C	D	F
bio1-1	60	44	6	10	0	0
bio1-2	60	36	8	12	0	4
bio3-3	60	0	0	1	52	7
bio2-1	20	0	0	0	0	20
Wild type	40	25	5	2	8	0

[a] Embryos were removed from immature siliques of heterozygotes and cultured on 1 to 2 μM DAPA. Embryo stages were noted at the time of culture. Responses were ranked after 21 d: A, extensive green callus with shoots; B, green callus with small shoots; C, green callus without shoots; D, trace amount of callus (unpigmented); F, no change in culture.

Because the expression vector used for complementation utilizes the T7 RNA polymerase promoter to control expression of the targeted sequence, auxotrophic *E. coli* strains were first lysogenized with λ(DE3) to introduce the required T7 RNA polymerase. The resulting strains were then transformed with (+10) and (−10) constructs and with a negative control (pDEST17) and evaluated for growth in the absence of biotin. Both the (−10) and (+10) constructs complemented the *bioD* mutant and supported growth on basal medium, although the (+10) construct was less effective than the (−10) construct (Fig. 7, E and G). In contrast, only the (−10) construct complemented the *bioA* mutant (Fig. 7, F and H). As expected, transformation with the control pDEST17 vector resulted in no growth on basal medium. We therefore conclude that the monocistronic (−10) transcript encodes a 92-kD fusion protein that is bifunctional, catalyzing both the DTB synthetase (BioD/BIO3) and DAPA aminotransferase (BioA/BIO1) reactions. The (+10) transcript, in contrast, encodes a smaller protein that exhibits only DTB synthase (BioD/BIO3) activity.

Evidence for Distinct *BIO1* and *BIO3* Transcripts in Arabidopsis

Having established that an Arabidopsis fusion protein produced from the monocistronic transcript is bifunctional in *E. coli* and that similar proteins should be present in a variety of plants and fungi, we next sought to determine whether single gene transcripts capable of producing distinct BIO1 and BIO3 proteins are produced in Arabidopsis. We were initially surprised by the striking differences in expression levels for *BIO1* and *BIO3* in public microarray databases. If full-length transcripts alone are produced, then the

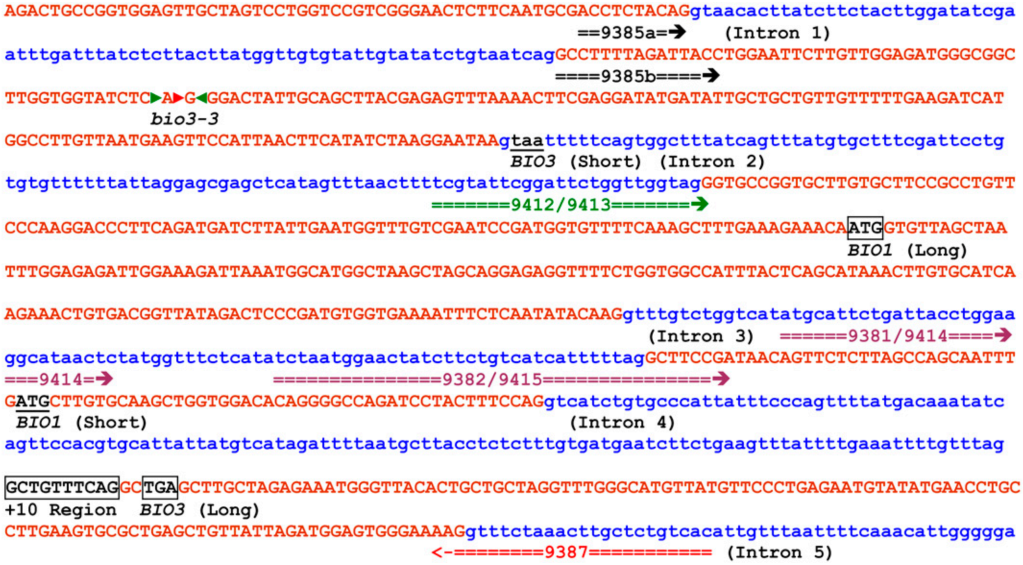
AGACTGCCGGTGGAGTTGCTAGTCCTGGTCCGTCGGGAACTCTTCAATGCGACCTCTACAGgtaacacttatcttctacttggatatcga
==9385a=➔ (Intron 1)
atttgatttatctcttacttatggttgtgtattgtatatctgtaatcagGCCTTTTAGATTACCTGGAATTCTTGTTGGAGATGGGCGGC
====9385b====➔
TTGGTGGTATCTC▶A▶G◀GGACTATTGCAGCTTACGAGAGTTTAAAACTTCGAGGATATGATATTGCTGCTGTTGTTTTTGAAGATCAT
bio3-3
GGCCTTGTTAATGAAGTTCCATTAACTTCATATCTAAGGAATAAgtaattttcagtggctttatcagtttatgtgctttcgattcctg
BIO3 (Short) (Intron 2)
tgtgtttttattaggagcgagctcatagtttaacttttcgtattcggattctggttggtagGGTGCCGGTGCTTGTGCTTCCGCCTGTT
=======9412/9413=======➔
CCCAAGGACCCTTCAGATGATCTTATTGAATGGTTTGTCGAATCCGATGGTGTTTTCAAAGCTTTGAAAGAAACAATGGTGTTAGCTAA
BIO1 (Long)
TTTGGAGAGATTGGAAAGATTAAATGGCATGGCTAAGCTAGCAGGAGAGGTTTTCTGGTGGCCATTTACTCAGCATAAACTTGTGCATCA
AGAAACTGTGACGGTTATAGACTCCCGATGTGGTGAAAATTTCTCAATATACAAGgtttgtctggtcatatgcattctgattacctggaa
(Intron 3) ======9381/9414====➔
ggcataactctatggtttctcatatctaatggaactatcttctgtcatcattttagGCTTCCGATAACAGTTCTCTTAGCCAGCAATTT
===9414=➔ ===============9382/9415===============➔
GATGCTTGTGCAAGCTGGTGGACACAGGGGCCAGATCCTACTTTCCAGgtcatctgtgcccattatttcccagttttatgacaaatatc
BIO1 (Short) (Intron 4)
agttccacgtgcattattatgtcatagattttaatgcttacctctctttgtgatgaatcttctgaagtttattttgaaattttgtttag
GCTGTTTCAGGCTGAGCTTGCTAGAGAAATGGGTTACACTGCTGCTAGGTTTGGGCATGTTATGTTCCCTGAGAATGTATATGAACCTGC
+10 Region BIO3 (Long)
CTTGAAGTGCGCTGAGCTGTTATTAGATGGAGTGGGAAAAGgtttctaaacttgctctgtcacattgtttaattttcaaacattgggga
<-========9387=========== (Intron 5)

Figure 4. Region of the Arabidopsis genome spanning the *BIO3-BIO1* junction. Exons are shown in capital letters (orange) and introns in lower case (blue). Boxed and underlined sequences identify potential start and stop codons for bicistronic and single gene transcripts. The long *BIO1* start site is located upstream of the long *BIO3* stop site. Alternative *BIO3* stop and *BIO1* start sites for translation are underlined. The large boxed (+10) region is differentially spliced from the full-length monocistronic transcript (GenBank accession no. EU089963). RT-PCR primers are noted beneath the sequence with black (*BIO3*, forward), green (*BIO1*, long), violet (*BIO1*, short), and red (*BIO3*, reverse) arrows. Flanking sequences from *bio3-3* begin at green (Oklahoma State sequence) and red (Salk sequence) arrowheads located within the second exon.

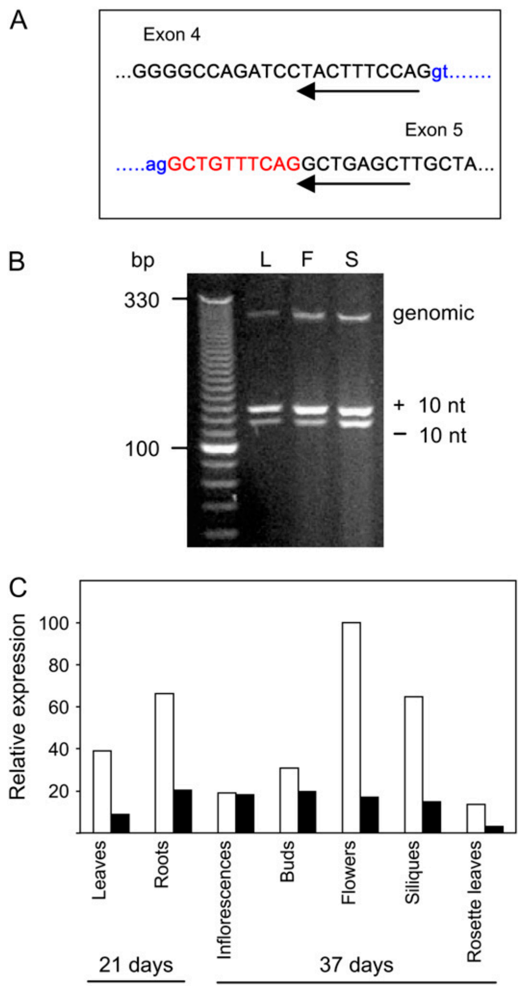

Figure 5. RT-PCR confirmation of (+10) and (−10) transcripts. A, One strategy used a reverse primer (underlined) that spanned the fourth and fifth exons and skipped the 10 nucleotides (red) that are alternatively spliced. Sequencing confirmed that the single product obtained was derived from the (−10) transcript. B, A second strategy used a reverse primer located in a downstream exon. As expected, two products that differed in length by 10 nucleotides were obtained from leaves (L), flowers (F), and siliques (S). Sequencing confirmed that these products differed with respect to the 10 nucleotides in question. A small amount of genomic DNA was also amplified. C, Semiquantitative RT-PCR analysis of the (+10; white rectangles) and (−10; black rectangles) products obtained using a reverse primer that spanned two downstream exons. Ubiquitin served as an internal standard. Band intensities were quantified using ImageJ (http://rsb.info.nih.gov/ij) and normalized relative to the maximal intensity in the flower sample.

relative levels of transcripts identified using primers localized to different regions of this locus should be the same. However, as shown in Figure 8, multiple microarray experiments indicate that *BIO1* expression is consistently above *BIO3* levels.

RT-PCR primers were therefore designed to distinguish between single gene and full-length transcripts based on the assumption that portions of the 3′-UTR for *BIO3* transcripts and the 5′-UTR for *BIO1* transcripts were positioned within introns of full-length transcripts. The locations of primers used in these experiments are illustrated in Figure 4. A PCR product of expected size (approximately 0.7 kb) was obtained when a *BIO3* forward primer (9385) joining the first two exons was used in combination with a reverse primer (9387) located in the fifth intron of the *BIO3-BIO1* locus, downstream of the *BIO3* (long) ORF and within the putative 3′-UTR of the *BIO3* (long) transcript (Fig. 9, lanes 6 and 7). Sequencing of this product (lower band) confirmed that a *BIO3* (long) single gene transcript is produced. We did not attempt to identify *BIO3* (short) single gene transcripts because the resulting protein would not likely be functional.

No product was obtained when *BIO1* forward primers (9381 and 9414) located in *BIO3-BIO1* intron 3 and far upstream of the predicted ATG for *BIO1* (short) were used in combination with a reverse primer that spanned two downstream exons (e.g. Fig. 9, lane 3). We concluded that these forward primers were located beyond the start of the *BIO1* (short) 5′-UTR. Additional primers (9415 and 9382) located further downstream gave a small amount of product of expected size (approximately 1.2 kb) that was confirmed by sequencing to represent *BIO1* (short) single gene (+10) transcripts (Fig. 9, lanes 4 and 5). A more dramatic result was obtained when forward primers (9412 and 9413) located in the predicted 5′-UTR for *BIO1* (long) were combined with the same reverse primer spanning downstream exons (Fig. 9, lanes 1 and 2). Sequencing of this product (approximately 1.4 kb) of expected size confirmed the presence of a (+10) *BIO1* (long) single gene transcript. The abundance of this transcript may explain in part the high *BIO1* signal observed in microarray experiments. What remains surprising is the consistently low *BIO3* signal observed in multiple microarrays. This appears to indicate that much of the *BIO1* signal observed in microarrays corresponds to single gene transcripts and that only

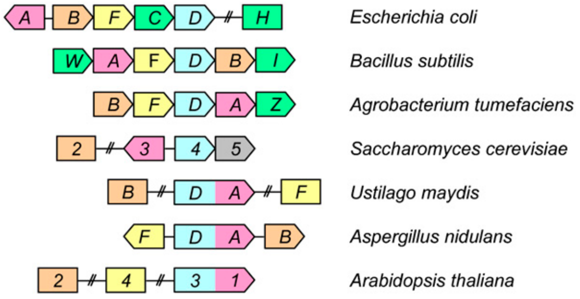

Figure 6. Genomic organization of biotin biosynthetic genes in microorganisms. Orthologs are depicted using the same color. Gene names (*bio*; *BIO*) and directions of transcription (pointed edge) are noted. Adjacent genes are abutted, genes that are close but not adjacent are joined by a thin line, and unlinked genes are separated by a hatched line. Green rectangles represent dissimilar genes involved in the biosynthesis of the initial biotin precursors. The *BIO5* gene of yeast is involved in transport. A *BIO3-BIO1* fusion protein is found in some fungi but not in yeasts.

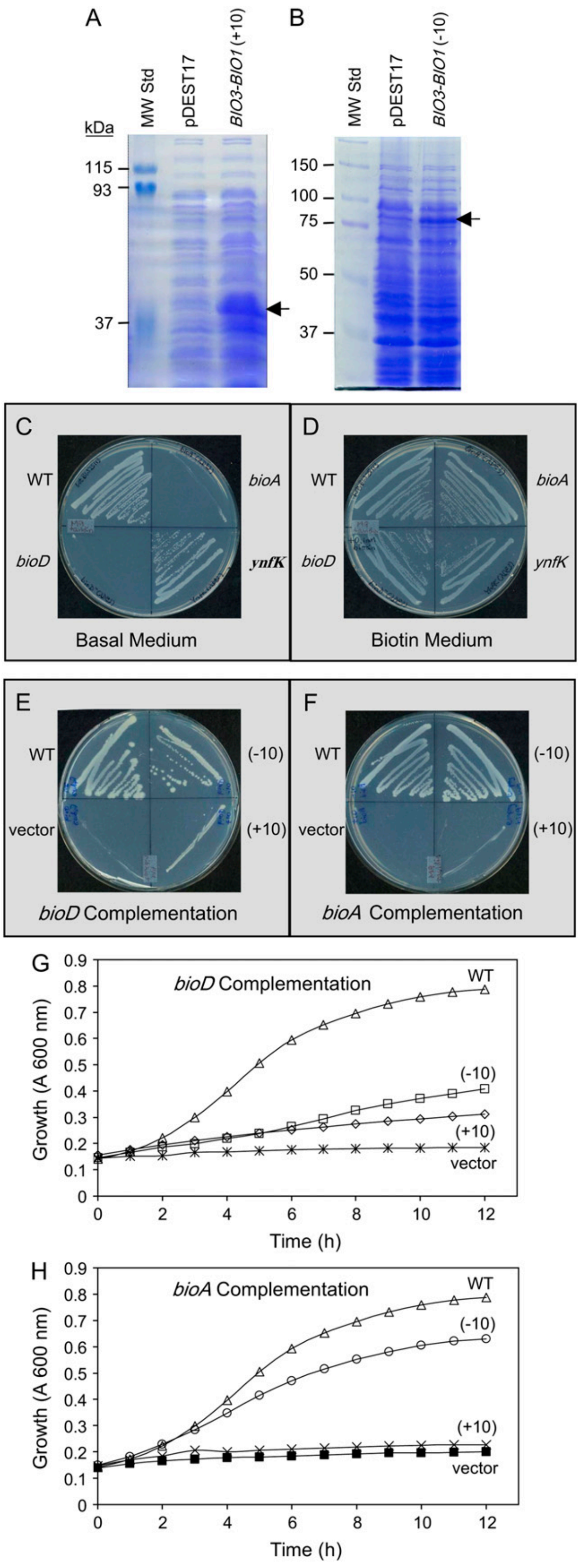

Figure 7. Heterologous expression (A and B) and functional characterization (C–H) of *BIO3-BIO1* gene products in *E. coli*. A and B, Protein extracts from isopropylthio-β-galactoside-induced *E. coli* cultures harboring either the empty vector (pDEST17) or the *BIO3-BIO1* bicistronic (+10) or monocistronic (−10) full-length cDNA were subjected to SDS-PAGE and stained with Coomassie Blue. Putative BIO3 (A) and fusion (B) proteins are marked with arrows. C and D, Expected responses of wild-type (WT) and mutant strains of *E. coli* in the presence and absence of biotin. E and F, Responses of a wild-type control strain and *bioD* (E) and *bioA* (F) mutants transformed with either the pDEST17 (empty) vector or recombinant vectors containing the (−10) or (+10) cDNA. G and H, Responses of wild-type and transformed *bioD* (G) and *bioA* (H) strains in liquid cultures. E to H, Strains were lysogenized with λDE3 and cultured on a kanamycin medium without biotin.

trace levels of the *BIO3-BIO1* monocistronic transcript and bifunctional protein are present in Arabidopsis plants.

Predicted Intracellular Localization of Biotin Biosynthetic Enzymes

The BIO3 protein contains an N-terminal sequence that is predicted to target the protein to mitochondria. All of the major prediction programs for intracellular localization of plant proteins (Emanuelsson et al., 2007) support this conclusion. It therefore appears that the bifunctional protein produced from the monocistronic, full-length transcript and whatever BIO3 (long) protein is produced from the single gene transcript all function in mitochondria. In contrast, the BIO1 (long) protein encoded by the single gene transcript does not appear to contain a mitochondrial localization signal. The short version of this protein encoded by the predominant (+10) single gene transcript is also missing conserved N-terminal sequences found in microorganisms. BIO1 activity required for biotin biosynthesis in Arabidopsis therefore appears to be associated primarily with the bifunctional protein. Whether alternative pathways exist for production of small amounts of cytosolic DAPA utilizing related *S*-adenosyl Met transaminases remains to be explored.

These results help to explain the failure of *bio3* and *bio1* mutants to complement. BIO1 activity is disrupted by insertion mutations in either the *BIO3* or *BIO1* coding regions. Following genetic crosses between *BIO3*/*bio3* and *BIO1*/*bio1* heterozygotes, embryos with a *bio3-BIO1*/*BIO3-bio1* genotype become arrested because the *bio3* insertion mutation disrupts the function of the adjacent copy of *BIO1*, whereas the second copy (*bio1*) located on the homologous chromosome is altered by mutation. The BIO3 (long) protein produced from single gene transcripts should be functional based on bacterial complementation experiments. But this protein cannot rescue mutant embryos devoid of BIO1 activity in complementation tests. Whether a majority of BIO3 activity in Arabidopsis is associated with the bifunctional protein or with the BIO3 (long) protein remains an open question. Comparative studies with fungi nevertheless suggest that the bifunctional protein is ancestral and may therefore be predominant.

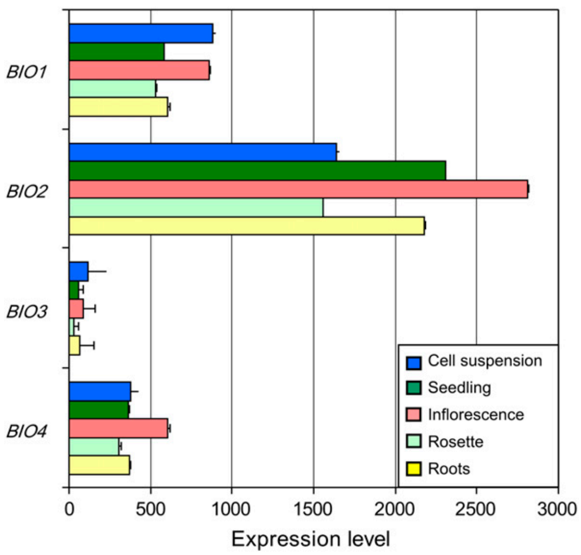

Figure 8. Summary of expression data for *BIO* genes of Arabidopsis. Results of microarray experiments were obtained from https://www.genevestigator.ethz.ch. *BIO3* expression levels are low in all tissues examined.

A Genome-Wide Survey for Gene Clusters with Related Metabolic Functions

Three approaches were taken to search for additional examples of gene clusters encoding proteins with related metabolic functions in Arabidopsis. We first searched for full-length cDNAs that covered two distinct genes based on TAIR 7.0 annotation of the genome, which updated the list of 58 complex loci published by Thimmapuram et al. (2005). We then compared this dataset with genes listed in the AraCyc (Zhang et al., 2005) database of metabolic pathways at TAIR. This approach uncovered only the *BIO3-BIO1* locus described here. Next, we scanned the Arabidopsis genome for examples of adjacent loci that encoded proteins with related metabolic functions but did not produce a full-length transcript that covered both genes. This identified 99 adjacent sets comprising 262 total genes. The vast majority of these cases are tandem duplications involving genes annotated as having the same enzymatic function. Several of these clusters contain more than two adjacent genes. Two clusters of interest were identified in addition to the *BIO3-BIO1* locus. One is required for chorismate biosynthesis in plastids: At1g48850 (chorismate synthase) and At1g48860 (5-enolpyruvylshikimate-3-phosphate synthase). The other involves branched-chain amino acid catabolism in mitochondria: At3g06850 (branched-chain keto-acid dehydrogenase) and At3g06860 (enoyl-CoA hydratase).

Because the success of these initial strategies for identifying clusters of interest is dependent on correct annotation of gene function, we pursued a third approach by looking for Arabidopsis orthologs of clustered yeast genes with related metabolic functions. A recent study by Hall and Dietrich (2007) identified 14 examples of such clusters in the sequenced genome of yeast. BLASTP (Altschul et al., 1997) analyses and Kyoto Encyclopedia of Genes and Genomes database (Ogata et al., 1999) searches failed to identify any definitive clusters of putative orthologs in Arabidopsis. We therefore conclude that with respect to the regulation of metabolic pathways, the *BIO3-BIO1* locus described here provides an interesting and atypical example of gene organization and function in plants.

DISCUSSION

Gene Clusters with Related Metabolic Functions

Clusters of genes with related metabolic functions are a defining feature of prokaryotic genomes. Eukaryotic orthologs of these genes, in contrast, tend to be dispersed throughout the genome and do not typically produce a polycistronic transcript. A significant number of eukaryotic operons have been described over the years (Blumenthal, 2004), particularly in *Caenorhabditis elegans*, but most of these are not involved in basic metabolism. Even in yeast, there are few known examples of gene clusters that produce enzymes with related metabolic functions (Hall and Dietrich, 2007). The locus described here provides an interesting example in Arabidopsis of two adjacent genes involved in sequential reactions of the same pathway that produce a combination of separate and chimeric transcripts. This locus does not appear to be an evolutionary remnant of a prokaryotic operon. Instead, it defines a genomic region in Arabidopsis that represents both a single gene for a bifunctional enzyme and two adjacent genes that produce multiple, distinct types of transcripts.

There are numerous examples of adjacent genes in Arabidopsis that produce a chimeric transcript (Thimmapuram et al., 2005). These transcripts can be either monocistronic, encoding a single bifunctional

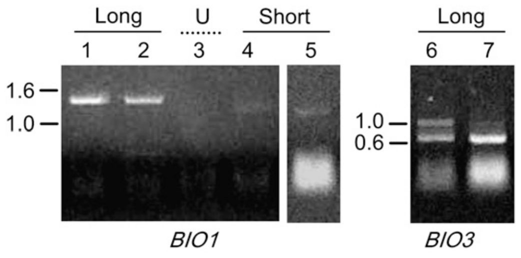

Figure 9. Semiquantitative RT-PCR evidence of *BIO1* and *BIO3* single gene transcripts in extracts from wild-type plants. *BIO1* forward primers were located either in the predicted 5′-UTR for the long transcript (lanes 1 and 2) or short transcript (lanes 4 and 5) or upstream (U) of the 5′-UTR for the short transcript (lane 3), but in all cases within an intron for the *BIO3-BIO1* transcript. The *BIO3* reverse primer was located in the predicted 3′-UTR for the long transcript but in an intron for the *BIO3-BIO1* transcript. The lower band in lanes 6 and 7 was confirmed by sequencing to be the expected product. This band is less abundant than the *BIO1* (long) product (lanes 1 and 2) when gels are run under equivalent conditions. The top band in lanes 6 and 7 represents contaminating genomic DNA. Plant extracts were prepared from leaves (lanes 1–6) and flowers (lane 7).

protein, or bicistronic, encoding two distinct polypeptides. Transcripts limited to individual genes can also be produced from these loci. The *BIO3-BIO1* region is not part of the original list of 60 loci characterized by Thimmapuram et al. (2005) because the corresponding full-length cDNAs were not yet deposited in GenBank. The distinctive feature of the *BIO3-BIO1* locus is the involvement of both gene products in the same metabolic pathway. The updated survey described here failed to identify additional examples of adjacent genes, based on current genome annotation, that produce a chimeric transcript for related but distinct enzymes. Even the broader search for adjacent genes involved in the same metabolic pathway, but not necessarily transcribed as a unit, identified only two new candidates. One of these clusters (At1g48850 and At1g48860), involved in aromatic amino acid biosynthesis, is of special interest because At1g48850 is required for embryo development (*EMB1144*; www.seedgenes.org) and At1g48860 (5-enolpyruvylshikimate-3-phosphate synthase) represents the well-characterized target for glyphosate herbicides. Orthologs of these two genes (*aroA* and *aroC*) are dispersed in the *E. coli* chromosome, whereas in *Saccharomyces* and *Neurospora* they constitute part of a complex (*arom*) locus that encodes a large, pentafunctional polypeptide (Duncan and Coggins, 1986). The second cluster of Arabidopsis genes (At3g06850 and At3g06860), involved in the metabolism of branched-chain amino acids, is more difficult to analyze from a comparative perspective because a definitive ortholog of At3g06860 (enoyl-CoA hydratase) remains to be identified in fungi and bacteria. The general conclusion, however, is that the Arabidopsis genome contains few examples of adjacent genes with sequential roles in metabolism.

Bifunctional Enzymes in Arabidopsis

At least 19 examples of bifunctional enzymes associated with cellular metabolism have been identified in Arabidopsis (Moore, 2004). In each case, a single gene product catalyzes more than one reaction in a common pathway. Six of these enzymes are involved with amino acid metabolism, six with lipid and carbohydrate metabolism, and the remainder with miscellaneous biochemical pathways. Some of these enzymes have well-characterized, bifunctional counterparts in lower eukaryotes. Two factors argue against the simple interpretation that BIO3-BIO1 should be viewed as another bifunctional plant protein that was incorrectly annotated in the Arabidopsis genome and previously escaped detection in biochemical studies. First, the predominant full-length transcript from this locus is bicistronic and encodes separate BIO3 and BIO1 proteins, not the bifunctional protein. Furthermore, single gene transcripts can also be produced, although the *BIO1* transcript does not appear to encode a functional protein. Some bifunctional Arabidopsis proteins, however, are also encoded by complex loci that produce more than one type of transcript. One intriguing example with notable similarities to the case described here is the bifunctional Lys ketoglutarate reductase-saccharopine dehydrogenase enzyme that catalyzes the initial reactions in Lys degradation. In some plants, including Arabidopsis, this enzyme is encoded by a complex locus with an internal promoter that allows expression of the monofunctional (downstream) saccharopine dehydrogenase, as well as internal polyadenylation sites that result is the production of monofunctional (upstream) Lys ketoglutarate reductase (Tang et al., 2002).

Origin of the BIO3-BIO1 Bifunctional Protein

Although the presence of adjacent genes oriented in the same direction and associated with a single metabolic pathway is reminiscent of gene organization in bacterial operons, the probable origin of the *BIO3-BIO1* locus of Arabidopsis is a gene fusion event that occurred early in the evolution of eukaryotes. This conclusion is supported by evidence of a bifunctional protein from whole-genome sequencing of *O. tauri*, a basal member of the green alga lineage that gave rise to land plants (Derelle et al., 2006) and *Cyanidioschyzon merolae*, a primitive red alga (Matsuzaki et al., 2004) used for comparative studies of plant evolution (Misumi et al., 2005), and from extensive sequence data derived from a wide range of basidiomycetes and filamentous fungi. Hall and Dietrich (2007) propose that much of the biotin pathway was lost in fungal ancestors of *Saccharomyces* and *Candida*, and that *BIO3* and *BIO1* orthologs were reacquired through separate, horizontal gene transfer from an unspecified prokaryotic donor. The ability to produce a bicistronic transcript and separate gene products through differential splicing appears to have been a more recent event because it is limited, based on available sequence data, to selected angiosperms, including Arabidopsis, *Brassica*, and rice. Two examples of adjacent Arabidopsis genes with related functions in amino acid metabolism identified here are different in that rice orthologs of these genes are not physically adjacent. We are therefore unable to point to a single example of adjacent genes with related but distinct metabolic functions in Arabidopsis that remain adjacent in unrelated angiosperms but do not produce a chimeric transcript or encode a fusion protein.

Implications for Biotin Biosynthesis in Plants

The intracellular localization of intermediate reactions in biotin synthesis in plants has remained unresolved despite the demonstration that the first reaction catalyzed by KAPA synthase occurs in the cytosol and the final reaction involving biotin synthase occurs in mitochondria (Rébeillé et al., 2007). Results presented here provide strong evidence that both of the intermediate steps catalyzed by the bifunctional BIO3-BIO1 protein and the monofunctional BIO3 protein also take place in mitochondria. This underscores the central role that mitochondria serve in the biosynthesis of vitamin coenzymes (Rébeillé et al., 2007). The membrane transport system that delivers KAPA into the mitochondrion of plant cells remains to be identified. The proteins involved may be difficult to identify through sequence homology because of differences in the compartmentalization of the biotin biosynthetic pathway in plants and microorganisms. The ability of a single plant protein to convert KAPA into DAPA and then DTB raises interesting questions about enzyme mechanics that remain to be addressed. Improved catalytic efficiency of the bifunctional enzyme may be advantageous in light of the trace amounts of intermediates available. The presence of a bifunctional enzyme also has implications for ongoing efforts to design herbicides that interfere with biotin production (Ashkenazi et al., 2007) and with biotechnological strategies to increase biotin levels in crop plants.

Another issue that needs to be reconciled is the ability of an *E. coli bioA* transgene to rescue the phenotype of the Arabidopsis *bio1-1* point mutant (Patton et al., 1996). Because the bacterial protein introduced into mutant plants did not include a mitochondrial localization signal, the enzymatic conversion of KAPA to DAPA probably took place in the cytosol, with the product transported into mitochondria. BIO3 activity in this case must have been provided either by a defective fusion protein altered only at a site associated with BIO1 activity, or by small amounts of monofunctional BIO3 protein. The incomplete rescue observed in these experiments, despite high levels of transgene expression, is consistent with inefficient reaction and transport mechanisms related to aberrant localization of biotin intermediates.

The presence of a bifunctional BIO3-BIO1 protein in cell extracts from Arabidopsis plants remains to be definitively established. Based on the low transcript levels detected in multiple microarray and RT-PCR experiments, and the small amounts of biotin produced in plant cells, this may be a challenging task. Indeed, our initial efforts to generate and utilize antibodies against purified Arabidopsis BIO3 and BIO1 proteins produced in *E. coli* have not been successful. These Arabidopsis proteins have also not been identified in a broad survey of the plant mitochondrial proteome (Heazlewood and Millar, 2005). Detailed biochemical studies on the enzymatic production of DAPA and DTB in plants may therefore need to focus initially on protein generated in *E. coli* rather than isolated from plant extracts. The work presented here provides an important framework for a variety of future studies on the BIO3-BIO1 protein of Arabidopsis, the regulation of biotin synthesis in plants, and the role of mitochondria in plant growth and development.

MATERIALS AND METHODS

Plant Materials and Growth Conditions

All three *bio1* alleles (Columbia ecotype) and the *bio3-3* insertion line (Columbia ecotype) from the Salk Institute (Alonso et al., 2003) can be obtained through the Arabidopsis Biological Resource Center at Ohio State University. We used internal seed stocks for the *bio1-1* EMS allele identified in the Meinke laboratory (Schneider et al., 1989) and the *bio1-2* and *bio1-3* insertion lines generated at Syngenta (McElver et al., 2001). The Syngenta lines are distinct from the SAIL population (Sessions et al., 2002) designed for reverse genetics. The *bio3-1* and *bio3-2* mutants (No-0 ecotype) were obtained from the RIKEN Bioresource Center in Japan. Plants at Oklahoma State University were grown in a soil mixture and placed in a growth room (24°C ± 2°C) under fluorescent lights (16-h light/8-h dark cycles) as described by Berg et al. (2005). At Iowa State University, seeds were first germinated on Murashige and Skoog agar medium (Invitrogen) containing 0.1% Suc. Seedlings were then transferred to LC1 Sunshine Mix soil (Sun Gro Horticulture) and grown to maturity under continuous illumination (170 μmol m^{-2} s^{-1}) at 22°C. Heterozygous plants were identified by screening siliques for the presence of aborted seeds. Allelism tests were performed by crossing two heterozygotes and screening immature F_1 siliques for the presence of aborted seeds. Detailed information on the methods used to characterize mutant seeds is presented in the tutorial section at www.seedgenes.org.

PCR Genotyping of Plants

Gene-specific primers for each mutant were designed using the SIGnAL iSect Primer Design program at http://signal.salk.edu and were purchased from IDT. Primers for the left T-DNA border in Salk and Syngenta lines and the Ds border in the RIKEN lines were used in combination with the appropriate gene-specific primers to detect and confirm insertions. A complete list of primers used is presented in Supplemental Table S1. Genomic DNA was isolated in the Meinke lab using a modified cetyl-trimethyl-ammonium bromide protocol (Lukowitz et al., 2000) and in the Nikolau lab using an SDS-phenol-chloroform extraction protocol. Two different PCR parameters were used: 94°C for 2 min, followed by 30 cycles of 94°C for 30 s, 56°C for 40 s, 72°C for 80 s, and a final elongation step of 72°C for 10 min (Meinke); and 96°C for 10 min followed by 35 cycles at 94°C for 15 s, 65°C for 30 s, and then at 72°C for 4 min for final extension (Nikolau). Reactions were performed with a Biometra Uno II (Meinke) or an MJ Research (Nikolau) thermocycler. Amplified products were separated in 1.0% agarose gels, stained with ethidium bromide, gel purified (Qiagen), and sequenced at the Oklahoma State University Recombinant DNA/Protein Resource Facility or the DNA Facility at Iowa State University to confirm insert locations.

Biotin Rescue Experiments

Biotin rescue of mutant seeds in heterozygous plants grown in soil was accomplished by daily watering of plants (20–40 mL/pot) with a solution of fertilizer (Berg et al., 2005) supplemented with 1 mM biotin. The solution was refrigerated between applications to limit microbial contamination. Heterozygous plants chosen for treatment were first identified by screening immature siliques for aborted seeds and then trimmed to remove excess branches and stems before supplementation began. Embryo rescue experiments were performed under aseptic conditions (Schneider et al., 1989) on plates containing Murashige and Skoog salts, 3% (w/v) Glc, 0.8% (w/v) agar, 0.1 mg/L 1-naphthylacetic acid, and 1.0 mg/L 6-benzylaminopurine. Two different sources of DAPA were used. The first (adjusted to 2 μM final concentration in the culture medium) was derived from an ethanol stock prepared 15 years ago (Shellhammer, 1991) using a powdered sample of DAPA provided by Nicholas Shaw and stored since that time at −20°C. The second sample (1 μM final concentration) also originated from the Shaw laboratory, but was provided in 2005 by Peter Roach (University of Southampton) in powdered form and then dissolved (8.0 mg in 10 mL 50% ethanol) to form a concentrated stock. Both sources of DAPA gave similar responses in culture. D,L-desthiobiotin (2 μM final concentration) and d-biotin (1 μM final concentration) were both obtained from Sigma Chemical Company.

Sequencing of the *bio1-1* Mutant Allele

Genomic DNA was isolated from leaf tissue of plants homozygous for the *bio1-1* allele and grown in the presence of 1 mM biotin. Using a set of primers that spanned the At5g57590 gene, PCR was used to generate a series of overlapping amplicons that were directly sequenced and compared to the wild-type Arabidopsis (*Arabidopsis thaliana*) genome (TAIR 7.0). Any polymorphisms identified between the sequences derived from *bio1-1* plants and the published genomic sequences were confirmed by PCR amplifying the homologous DNA fragment from wild-type Columbia plants.

RT-PCR Analysis of Transcript Diversity

Cauline leaves, young flowers, and siliques with embryos up to the transition stage were harvested from plants grown in soil, flash frozen in liquid nitrogen, and stored at −80°C without thawing until RNA extraction. Frozen tissue (0.1 g) was homogenized in liquid nitrogen. Total RNA was prepared from powdered tissues using the RNeasy plant mini kit (Qiagen), treated with RNase-free DNase I (TaKaRa Bio), quantified with a Shimadzu UV-160 spectrophotometer, and visualized on a 1.0% formaldehyde agarose gel. For the two-step RT-PCR reaction, 5 μg total RNA was reverse transcribed using the SuperScript first-strand synthesis system (Invitrogen). A 1-μL aliquot of reverse transcribed reaction was then used as a template for the second-step PCR with REDTaq DNA polymerase (Sigma-Aldrich). Flower cDNA was used as a template to amplify separate *BIO3* and *BIO1* transcripts. Leaf, flower, and silique cDNAs were used to amplify (−10) and (+10) chimeric transcripts. Reactions were performed with a Biometra Uno II thermocycler. Primers are listed in Supplemental Table S1. PCR parameters were: 94°C for 1 min, followed by 30 cycles of 94°C for 1 min, 55°C for 1 min, 72°C for 2 min, and a final elongation step at 72°C for 10 min. Amplified products from (−10) transcripts (112 bp) and (+10) transcripts (122 bp) were separated using a high-resolution 4% MetaPhor (Cambrex Bio Science) agarose gel.

5′- and 3′-RACE Experiments

Initial 5′-RACE experiments designed to identify the full-length mRNA sequence that corresponded to EST clone RZ128g09R were conducted with the Invitrogen 5′-RACE system. The 5′ and 3′ ends of the *BIO3-BIO1* mRNA were authenticated with an RNA ligase-mediated rapid amplification method (Maruyama and Sugano, 1994) using the GeneRacer kit (Invitrogen).

Complementation of Bacterial Biotin Auxotrophs

Escherichia coli strains carrying mutations in biotin biosynthetic genes were obtained from the Keio collection of single gene knockouts (http://ecoli.aist-nara.ac.jp/gb6/Resources/deletion/deletion.html), which replaced each coding region with a kanamycin resistance gene (Baba et al., 2006). Four different strains were used in these studies: BW25113 (wild-type parental strain), JW0761 (*bioD* knockout), JW0757 (*bioA* knockout), and JW5264 (*ynfK* knockout). Mutant strains were confirmed by their ability to grow on kanamycin and by sequencing of PCR products that amplified the mutant allele. Strains were first lysogenized with λ(DE3) to introduce the required T7 RNA polymerase and then transformed with pDEST17-derivative plasmids that carried either the *BIO3-BIO1* (+10) or (−10) cDNA versions. These plasmids were constructed using PCR products corresponding to full-length Arabidopsis transcripts that were amplified using forward (5′-CACCATGATACCCG-TAACCGC-3′) and reverse (5′-AGCTGGAGAGAGAGTTTTGGGT-3′) primers spanning the entire *BIO1-BIO3* locus and then cloned in pENTR vector (Invitrogen). The (+10) and (−10) splice variants were identified by sequencing. Constructs were then moved from pENTR to pDEST17 using Gateway Technology as recommended by the manufacturer. *E. coli* strains were grown in isopropylthio-β-galactoside (0.1 mM) and kanamycin containing solid or liquid (M9 Glc) medium that were either depleted of biotin by the addition of 50 μg mL^{-1} of avidin (basal medium) or supplemented with 1 mM biotin.

Sequence Alignments and Genome Analyses

BLASTP and TBLASTN searches of GenBank datasets were performed using default settings at the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov). Sequences were aligned using the MultAlin program (http://bioinfo.genopole-toulouse.prd.fr/multalin/multalin.html) as described by Corpet (1988). Potential coding regions were identified with ORF finder at the National Center for Biotechnology Information. Putative orthologs of selected Arabidopsis genes were identified using the Kyoto Encyclopedia of Genes and Genomes (www.genome.ad.jp/en/gn_kegg.html) database. The AraCyc dataset (Zhang et al., 2005) of metabolic pathways in Arabidopsis was used to search for adjacent genes with related functions. Two different AraCyc datasets were downloaded in February and July, 2007, from TAIR 7.0 (www.arabidopsis.org): aracyc_dump_20070213 and aracyc_dump_20070703. These datasets were then analyzed using a computer program that we developed to first read into memory the AraCyc pathway designation for each gene and then scan down the ordered list of Arabidopsis Genome Initiative locus identifiers to look for instances where sequential genes shared the same pathway designation. The output was compared with the original list of complex loci obtained from Thimmapuram et al. (2005), with an updated list of such loci provided by David Swarbeck at TAIR, and with detailed AraCyc assignments of Arabidopsis genes to specific metabolic reactions.

Sequence data from this article can be found in the GenBank/EMBL data libraries under accession numbers EU089963 and EU090805.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure S1. Region of the Arabidopsis genome that includes the entire *BIO3-BIO1* locus (At5g57600/At5g57590) and part of the downstream gene (At5g57580).

Supplemental Figure S2. Sequence alignments between four different full-length cDNAs from the *BIO3-BIO1* locus: Nikolau (−10; GenBank EU089963, monocistronic); Nikolau (+10; GenBank EU090805, bicistronic); RIKEN (+10; RAFL22-07-J07, bicistronic); and French (+10; BX842298, bicistronic).

Supplemental Table S1. Primers for RT-PCR and genotype analyses.

ACKNOWLEDGMENTS

We thank Nicholas Shaw (Biotechnology, Lonza AG, Switzerland) and Peter Roach (University of Southampton, UK) for providing samples of DAPA, and Johnny Lloyd (Oklahoma State University) for assistance with plant maintenance.

Received August 16, 2007; accepted November 2, 2007; published November 9, 2007.

LITERATURE CITED

Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, Shinn P, Stevenson DK, Zimmerman J, Barajas P, Cheuk R, et al (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana.* Science **301:** 653–657

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:** 3389–3402

Arnal N, Alban C, Quadrado M, Grandjean O, Mireau H (2006) The *Arabidopsis* Bio2 protein requires mitochondrial targeting for activity. Plant Mol Biol **62:** 471–479

Ashkenazi T, Pinkert D, Nudelman A, Widberg A, Wexler B, Wittenbach V, Flint D, Nudelman A (2007) Aryl chain analogues of the biotin vitamers as potential herbicides. Part 3. Pest Manag Sci **63:** 974–1001

Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol **2:** 2006.0008

Baldet P, Alban C, Douce R (1997) Biotin synthesis in higher plants: purification and characterization of *bioB* gene product equivalent from *Arabidopsis thaliana* overexpressed in *Escherichia coli* and its subcellular localization in pea leaf cells. FEBS Lett **419:** 206–210

Berg M, Rogers R, Muralla R, Meinke D (2005) Requirement of aminoacyl-tRNA synthetases for gametogenesis and embryo development in Arabidopsis. Plant J **44:** 866–878

Blumenthal T (2004) Operons in eukaryotes. Brief Funct Genomics Proteomics **3:** 199–211

Castelli V, Aury JM, Jaillon O, Wincker P, Clepet C, Menard M, Cruaud C, Quétier F, Scarpelli C, Schächter V, et al (2004) Whole genome sequence comparisons and ''full-length'' cDNA sequences: a combined approach to evaluate and improve *Arabidopsis* genome annotation. Genome Res **14:** 406–413

Corpet F (1988) Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res **16:** 10881–10890

DeMoll E (1996) Biosynthesis of biotin and lipoic acid. *In* FC Neidhardt, ed, *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology. ASM Press, Washington, DC, pp 704–709

Derelle E, Ferraz C, Rombauts S, Rouzé P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynié S, Cooke R, et al (2006) Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. Proc Natl Acad Sci USA **103:** 11647–11652

Duncan K, Coggins JR (1986) The *serC-aroA* operon of *Escherichia coli.* Biochem J **234:** 49–57

Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protocols **2:** 953–971

Hall C, Dietrich FS (2007) The reacquisition of biotin prototrophy in *Saccharomyces cerevisiae* involved horizontal gene transfer, gene duplication, and gene clustering. Genetics **177:** 2293–2307

Heazlewood JL, Millar AH (2005) AMPDB: the Arabidopsis mitochondrial protein database. Nucleic Acids Res **33:** D605–D610

Henikoff S, Till BJ, Comai L (2004) TILLING: traditional mutagenesis meets functional genomics. Plant Physiol **135:** 630–636

Ilag LL, Kumar AM, Soll D (1994) Light regulation of chlorophyll biosynthesis at the level of 5-aminolevulinate formation in *Arabidopsis.* Plant Cell **6:** 265–275

Lukowitz W, Gillmor CS, Scheible WR (2000) Positional cloning in Arabidopsis: why it feels good to have a genome initiative working for you. Plant Physiol **123:** 795–805

Maruyama K, Sugano S (1994) Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. Gene **138:** 171–174

Matsuzaki M, Misumi O, Shin-I T, Maruyama S, Takahara M, Miyagishima SY, Mori T, Nishida K, Yagisawa F, Nishida K, et al (2004) Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. Nature **428:** 653–657

McElver J, Tzafrir I, Aux G, Rogers R, Ashby C, Smith K, Thomas C, Schetter A, Zhou Q, Cushman MA, et al (2001) Insertional mutagenesis of genes required for seed development in *Arabidopsis thaliana.* Genetics **159:** 1751–1763

Miroux B, Walker JE (1996) Over-production of proteins in *Escherichia coli*: mutant hosts that allow synthesis of some membrane proteins and globular proteins at high levels. J Mol Biol **260:** 289–298

Misumi O, Matsuzaki M, Nozaki H, Miyagishima SY, Mori T, Nishida K, Yagisawa F, Yoshida Y, Kuroiwa H, Kuroiwa T (2005) *Cyanidioschyzon merolae* genome: a tool for facilitating comparable studies on organelle biogenesis in photosynthetic eukaryotes. Plant Physiol **137:** 567–585

Moore B (2004) Bifunctional and moonlighting enzymes: lighting the way to regulatory control. Trends Plant Sci **9:** 221–228

Muralla R, Sweeney C, Stepansky A, Leustek T, Meinke D (2007) Genetic dissection of histidine biosynthesis in Arabidopsis. Plant Physiol **144:** 890–903

Nikolau BJ, Ohlrogge JB, Wurtele ES (2003) Plant biotin-containing carboxylases. Arch Biochem Biophys **414:** 211–222

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1999) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res **27:** 29–34

Patton DA, Franzmann LH, Meinke DW (1991) Mapping genes essential for embryo development in *Arabidopsis thaliana.* Mol Gen Genet **227:** 337–347

Patton DA, Schetter AL, Franzmann LH, Nelson K, Ward ER, Meinke DW (1998) An embryo-defective mutant of Arabidopsis disrupted in the final step of biotin synthesis. Plant Physiol **116:** 935–946

Patton DA, Volrath S, Ward ER (1996) Complementation of an *Arabidopsis thaliana* biotin auxotroph with an *Escherichia coli* biotin biosynthetic gene. Mol Gen Genet **251:** 261–266

Picciocchi A, Douce R, Alban C (2003) The plant biotin synthase reaction: identification and characterization of essential mitochondrial accessory protein components. J Biol Chem **278:** 24966–24975

Pinon V, Ravanel S, Douce R, Alban C (2005) Biotin synthesis in plants: the first committed step of the pathway is catalyzed by a cytosolic 7-keto-8-aminopelargonic acid synthase. Plant Physiol **139:** 1666–1676

Rébeillé F, Alban C, Bourguignon J, Ravannel S, Douce R (2007) The role of plant mitochondria in the biosynthesis of coenzymes. Photosynth Res **92:** 149–162

Rodionov DA, Mironov AA, Gelfand MS (2002) Conservation of the biotin regulon and the BirA regulatory signal in eubacteria and archaea. Genome Res **12:** 1507–1516

Roosens NHCJ, Thu TT, Iskander HM, Jacobs M (1998) Isolation of the ornithine-δ-aminotransferase cDNA and effect of salt stress on its expression in *Arabidopsis thaliana.* Plant Physiol **117:** 263–271

Schneider T, Dinkins R, Robinson K, Shellhammer J, Meinke DW (1989) An embryo-lethal mutant of *Arabidopsis thaliana* is a biotin auxotroph. Dev Biol **131:** 161–167

Schomburg FM, Patton DA, Meinke DW, Amasino RM (2001) *FPA*, a gene involved in floral induction in *Arabidopsis*, encodes a protein containing RNA-recognition motifs. Plant Cell **13:** 1427–1436

Seki M, Narusaka M, Kamiya A, Ishida J, Satou M, Sakurai T, Nakajima M, Enju A, Akiyama K, Oono Y, et al (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. Science **296:** 141–145

Sessions A, Burke E, Presting G, Aux G, McElver J, Patton D, Dietrich B, Ho P, Bacwaden J, Ko C, et al (2002) A high-throughput *Arabidopsis* reverse genetics system. Plant Cell **14:** 2985–2994

Shellhammer AJ (1991) Analysis of a biotin auxotroph of *Arabidopsis thaliana.* PhD thesis. Oklahoma State University, Stillwater, OK

Shellhammer J, Meinke DW (1990) Arrested embryos from the *bio1* auxotroph of *Arabidopsis* contain reduced levels of biotin. Plant Physiol **93:** 1162–1167

Streit WR, Entcheva P (2003) Biotin in microbes, the genes involved in its biosynthesis, its biochemical role and perspectives for biotechnological production. Appl Microbiol Biotechnol **61:** 21–31

Suhre K, Claverie JM (2004) FusionDB: a database for in-depth analysis of prokaryotic gene fusion events. Nucleic Acids Res **32:** D273–D276

Tang G, Zhu X, Gakiere B, Levanony H, Kahana A, Galili G (2002) The bifunctional *LKR/SDH* locus of plants also encodes a highly active monofunctional lysine-ketoglutarate reductase using a polyadenylation signal located within an intron. Plant Physiol **130:** 147–154

Thimmapuram J, Duan H, Liu L, Schuler MA (2005) Bicistronic and fused monocistronic transcripts are derived from adjacent loci in the *Arabidopsis* genome. RNA **11:** 128–138

Tzafrir I, Pena-Muralla R, Dickerman A, Berg M, Rogers R, Hutchens S, Sweeney TC, McElver J, Aux G, Patton D, et al (2004) Identification of genes required for embryo development in Arabidopsis. Plant Physiol **135:** 1206–1220

Weaver LM, Yu F, Wurtele ES, Nikolau BJ (1996) Characterization of the cDNA and gene coding for the biotin synthase of *Arabidopsis thaliana.* Plant Physiol **110:** 1021–1028

Zhang P, Foerster H, Tissier CP, Mueller L, Paley S, Karp PD, Rhee SY (2005) MetaCyc and AraCyc: metabolic pathway databases for plant research. Plant Physiol **138:** 27–37