Analysis of the *pdx-1* (*snz-1/sno-1*) Region of the *Neurospora crassa* Genome: Correlation of Pyridoxine-Requiring Phenotypes With Mutations in Two Structural Genes

Laura E. Bean,*,[1] William H. Dvorachek, Jr.,* Edward L. Braun,*,†,[2] Allison Errett,* Gregory S. Saenz,*,[3] Mara D. Giles,* Margaret Werner-Washburne,* Mary Anne Nelson* and Donald O. Natvig*

**Department of Biology, University of New Mexico, Albuquerque, New Mexico 87131 and* †*National Center for Genome Resources, Santa Fe, New Mexico 87505*

Manuscript received October 3, 2000
Accepted for publication December 15, 2000

ABSTRACT

We report the analysis of a 36-kbp region of the *Neurospora crassa* genome, which contains homologs of two closely linked stationary phase genes, *SNZ1* and *SNO1*, from *Saccharomyces cerevisiae*. Homologs of *SNZ1* encode extremely highly conserved proteins that have been implicated in pyridoxine (vitamin B6) metabolism in the filamentous fungi *Cercospora nicotianae* and in *Aspergillus nidulans*. In *N. crassa*, *SNZ* and *SNO* homologs map to the region occupied by *pdx-1* (pyridoxine requiring), a gene that has been known for several decades, but which was not sequenced previously. In this study, pyridoxine-requiring mutants of *N. crassa* were found to possess mutations that disrupt conserved regions in either the *SNZ* or *SNO* homolog. Previously, nearly all of these mutants were classified as *pdx-1*. However, one mutant with a disrupted *SNO* homolog was at one time designated *pdx-2*. It now appears appropriate to reserve the *pdx-1* designation for the *N. crassa SNZ* homolog and *pdx-2* for the *SNO* homolog. We further report annotation of the entire 36,030-bp region, which contains at least 12 protein coding genes, supporting a previous conclusion of high gene densities (12,000–13,000 total genes) for *N. crassa*. Among genes in this region other than *SNZ* and *SNO* homologs, there was no evidence of shared function. Four of the genes in this region appear to have been lost from the *S. cerevisiae* lineage.

ALTHOUGH efforts are underway to sequence and annotate the genomes of *Neurospora crassa* and other filamentous fungi, there remain few carefully annotated large regions of genomic DNA. Such analyses are required for accurate estimates of gene numbers, and they are extremely valuable for investigations in comparative genomics as well as in gene structure and function. We have sequenced and annotated a cosmid insert carrying *N. crassa* genes homologous to the *SNZ* and *SNO* genes (Braun *et al.* 1996; Padilla *et al.* 1998) from *Saccharomyces cerevisiae*, which encode conserved proteins distantly related to proteins involved in amino acid and nucleotide biosynthesis (Galperin and Koonin 1997). Recent evidence suggests that homologs of *SNZ* participate in pyridoxine (vitamin B6) metabolism in *Cercospora nicotianae* (Ehrenshaft *et al.* 1999a,b) and *Aspergillus nidulans* (Osmani *et al.* 1999). Results presented here indicate a role in pyridoxine metabolism for both *SNZ* and *SNO* homologs in *N. crassa*.

Initial interest in eukaryotic *SNZ* and *SNO* homologs on the part of several researchers stemmed from patterns of expression as well as from a possible role for *SNZ* homologs in avoidance of oxidative damage. The synthesis of the *S. cerevisiae* Snz1 protein increases dramatically when cells enter stationary phase (Fuge *et al.* 1994; Braun *et al.* 1996), and homologs of *SNZ* have been identified as ethylene-inducible mRNAs from the rubber tree plant, *Hevea brasiliensis* (Sivasubramaniam *et al.* 1995) and the marine sponge, *Suberites domuncula* (Krasko *et al.* 1999). The *SNZ* homolog in the filamentous fungus *C. nicotianae* was discovered because mutations in this gene, designated *SOR1*, result in hypersenstitivity to singlet oxygen-generating agents (Ehrenshaft *et al.* 1999b).

In many organisms, genes encoding Snz homologs are closely linked to genes encoding Sno homologs, which are related to amidotransferases involved in amino acid and nucleotide biosynthesis (Galperin and Koonin 1997). The *SNZ* and *SNO* homologs form an apparent operon in some prokaryotes (Galperin and Koonin 1997) and are coregulated in *S. cerevisiae* (Padilla *et al.* 1998). In addition to close genomic linkage of their respective genes, Snz and Sno proteins

Corresponding author: Donald O. Natvig, Department of Biology, University of New Mexico, Albuquerque, NM 87131.
E-mail: dnatvig@unm.edu

[1] *Present address:* Cell and Molecular Biology Program, Michigan State University, East Lansing, MI.

[2] *Present address:* Department of Plant Biology, The Ohio State University, Columbus, OH.

[3] *Present address:* Department of Plant Pathology, Cornell University, Ithaca, NY.

exhibit physical and genetic interactions that suggest they function as components of an oligomeric complex (Padilla *et al.* 1998). It can be inferred, therefore, that *SNZ* and *SNO* cooperate in function, a conclusion strongly supported by results presented here.

This study afforded the opportunity to explore the relationship between the *N. crassa SNZ* and *SNO* homologs and mutations that result in a requirement for pyridoxine, and it allowed a detailed examination of a portion of the genome in which these genes reside. The *N. crassa SNZ* and *SNO* homologs were found to be closely linked, as is observed in other microorganisms, and they map to the *pdx-1* (pyridoxine requiring) region of linkage group IVR (see Nelson *et al.* 1998). Results further indicate that pyridoxine auxotrophy in *N. crassa* can be caused by mutations in either structural gene.

The 36-kbp region examined contains at least 12 genes, including the homologs of *SNZ* and *SNO*. This reflects a gene density consistent with recent estimates (Nelson *et al.* 1997; Kelkar *et al.* 2001) suggesting high gene numbers (12,000–13,000) for *N. crassa*. With the exception of the *SNZ* and *SNO* homologs, there is no evidence for clustering of genes of shared function in this region. In fungi and prokaryotes, the clustering of genes of shared function can reflect dispensable function and a potential for horizontal transfer ("selfish" operons; Lawrence and Roth 1996; Keller and Hohn 1997). However, there is no evidence that the clustering of *SNZ* and *SNO* homologs of *N. crassa* and other organisms reflects either dispensable function or horizontal transfer.

MATERIALS AND METHODS

Library: Cosmid clone G6G8 from the Orbach/Sachs cosmid library (Orbach 1994; Kelkar *et al.* 2001) was obtained from the Fungal Genetics Stock Center (FGSC), University of Kansas Medical Center, Kansas City. This clone has the alternative designation X137G08 (Kelkar *et al.* 2001). In preliminary experiments (not presented) it was found to contain *N. crassa* homologs of the *SNZ* and *SNO* genes of *S. cerevisiae* (Braun *et al.* 1996; Padilla *et al.* 1998). Initial identification was made by colony blot hybridization employing ^{32}P-labeled DNA from a cDNA clone carrying the *N. crassa SNZ* homolog.

Subcloning of cosmid G6G8: *Escherichia coli* cells containing the G6G8 cosmid were grown at 37° for 15 hr in 50 ml Terrific broth (Sambrook *et al.* 1989) with 50 μg/ml ampicillin, and cosmid DNA was isolated using the QIAGEN (Valencia, CA) Plasmid Midi kit.

Cosmid DNA was subcloned for shotgun sequencing using two different methods. First, *Sau*3AI partial digestion was performed (Sambrook *et al.* 1989). To reduce the number of chimeric clones, the partially digested DNA was dephosphorylated using shrimp alkaline phosphatase [United States Biochemical (Cleveland) and Amersham (Buckinghamshire, UK)] according to the manufacturer's recommendations. After purification using QIAquick PCR product clean-up (QIAGEN), the partially digested, dephosphorylated DNA was cloned into *Bam*HI-digested pUC-18 using a standard ligation protocol (Sambrook *et al.* 1989). Ligated DNA was transformed into INVαF′ cells (Invitrogen, San Diego). In addition to using *Sau3AI*, fragments were produced for subcloning by complete digestion of the G6G8 cosmid DNA using four different restriction enzymes with 6-bp recognition sequences followed by dephosphorylation. One procedure used cosmid DNA digested with *Hin*dIII and *Eco*RI, while another used *Kpn*I and *Pst*I. Digestion products were ligated into pUC-18 cut with the corresponding enzymes.

Individual white colonies were transferred to 96-well block plates containing 1.5 ml Terrific broth with ampicillin (50 μg/ml), and cells were grown at 250 rpm for 20 hr at 37°. Template DNA for sequencing was purified using the alkaline lysis protocol of Roe *et al.* (1996) or the QIAprep spin Miniprep kit (QIAGEN) according to the manufacturer's instructions.

DNA sequencing: DNA sequences were obtained with an ABI 377 automated sequencer using cycle-sequencing, dye-terminator procedures with ThermoSequenase (Amersham) and ABI PRISM BigDye chemistries. Sequence gaps left after assembly of random-clone sequences were closed by direct sequencing of cosmid template DNA (prepared as described above) using custom-synthesized oligonucleotide primers.

Sequence assembly: Phred (Green and Ewing 1997) was used to call bases in the raw data files on a SUN workstation. Cloning-vector DNA sequences were deleted from each raw sequencing file using Crossmatch. The insert sequence was assembled into contiguous fragments from ~700 individual sequence reads using Phrap (Green 1996) running on an SGI workstation. We used Sequencher 3.0 (Gene Codes Corporation, 1995) as a quality check of the PHRAP assembly to design primers to fill gaps and improve sequence quality and to confirm sequence across the entire insert using representative chromatograms. The 36,030-nucleotide G6G8 insert sequence has been deposited at GenBank, with annotations, under accession no. AF309689.

Sequence analysis: The nucleotide sequence was searched for homologs of previously identified genes by performing gapped BLAST searches (Altschul *et al.* 1997) using protein and nucleotide databases available from the National Center for Biotechnology Information (NCBI, Bethesda, MD). The databases examined included the nonredundant database (NR) and dbEST [espressed sequence tagged (EST) database] from NCBI, as well as the Saccharomyces genome database (Stanford University, Stanford, CA). The algorithms employed included BLASTX, BLASTP, TBLASTX, and BLASTN, as appropriate for specific databases and queries. MacDNASIS v. 3.2 (Hitachi) was used to find open reading frames (ORFs) using codon bias for *N. crassa*. Identification of open reading frames and determination of codon usage were also aided by services available from the Virtual Genome Center (http://alces.med.umn.edu/webtrans.html).

Analysis of *pdx-1* mutants: Strains carrying various *pdx-1* alleles were obtained from the FGSC in the Department of Microbiology, University of Kansas Medical Center. Mycelium was grown in N medium, supplemented with 1.5 μg/ml pyridoxine (Davis and deSerres 1970). Genomic DNA was prepared using the Puregene D-5000A plant DNA isolation kit (Gentra Systems, Research Triangle Park, NC). The *N. crassa SNZ1* homolog was amplified from genomic DNA preparations by polymerase chain reaction (PCR) using forward primer 5′-ACAAACCTAAGCTCTCAATCGTGGT-3′ and reverse primer 5′-TCCAAGCCCCTTTTTAGTTCGT-3′. Sequences were obtained using forward and reverse PCR primers along with internal primers 5′-GCGTCGACTACATCGACGAGA-3′ and 5′-TTCTTGAGGAGCTCAACATCGG-3′. The *N. crassa SNO1* homolog was amplified by PCR using forward primer 5′-CCTGGTGTAACCAAAAGACCTATCG-3′ and reverse primer 5′-AACCGTGACCCTCATAGTCGC-3′. Sequences were obtained using forward and reverse PCR primers along with

internal primers 5′-AGTCTTTTTTTCTCTTTTCCTAACCCG-3′ and 5′-ACTCTGGAGCTGTGTGCCGTA-3′. Primers were tested and wild-type *SNZ1* and *SNO1* homolog sequences were confirmed with *N. crassa* strain 74-OR23-1*A*.

RESULTS

Genes represented in the cosmid insert: Our annotation of the 36,030-bp insert from cosmid G6G8 includes 13 putative protein-coding genes (Table 1), 12 of which were deduced with a high level of certainty. The identification of coding regions employed a combination of analyses including BLAST searches, examination of ORFs for *N. crassa* codon preference (*e.g.*, Chary *et al.* 1990), and searches for consensus sequences associated with translational start sites (Bruchez *et al.* 1993a) and intron splicing (Bruchez *et al.* 1993b). With two exceptions, the validity of each gene was established by identification of a homologous genomic or cDNA sequence using BLAST searches. One exception, ORF G6G8.11, encodes 426 amino acids without interruption and exhibits strong *N. crassa* codon bias. It appears to represent a true protein-coding sequence, despite the fact that no homolog or cDNA was identified. The other exception, ORF G6G8.2, exhibits rather poor *N. crassa* codon preference and lacks similarity to known genes from other organisms. This ORF is contained within certain *N. crassa* cDNA sequences, but nevertheless there is some question whether this region encodes a protein (discussed below).

Six of the 13 annotated genes in Table 1 are represented by partial cDNA sequences at GenBank that are derived from *N. crassa* EST projects at the University of New Mexico and the University of Oklahoma (see footnote to Table 1). In addition, a partial cDNA sequence from *E. nidulans* encoding the probable ortholog of one gene (G6G8.9) has been identified (Table 1).

Two of the genes in G6G8 have paralogs previously identified in *N. crassa*. A different 3-hydroxyisobutyrate dehydrogenase (3HD) homolog was found earlier by the *Neurospora* Genome Project in a cDNA clone (Nelson *et al.* 1997). There is only moderate sequence similarity between the two predicted *N. crassa* 3HD proteins, and BLAST searches indicated that the sequence reported here is more closely related to 3HDs in other organisms (best BLAST match to *Drosophila melanogaster*). It is therefore possible that the gene identified previously encodes a dehydrogenase with a function different from that of characterized 3HDs. There also was a previously identified *N. crassa* thioredoxin. Again, the results of BLAST searches suggest a closer relationship between the protein reported here and thioredoxins from other organisms [best BLAST match to *Emericella* (=*Aspergillus*) *nidulans*].

There are four genes in G6G8 that appear to lack *S. cerevisiae* orthologs, despite evidence suggesting they were present in the common ancestor of *N. crassa* and *S. cerevisiae*. Three of these proteins (encoded by G6G8.5, G6G8.6, and G6G8.9, see Table 1) have homologs in other eukaryotic kingdoms but lack an *S. cerevisiae* homolog, the criterion used to establish gene loss by Braun *et al.* (2000). Two of the proteins are probable structural enzymes [3HD and D-amino acid oxidase (DAO)]. The third appears distantly related to translation initiation factor 1A, a protein essential for transfer of the initiator tRNA to 40 S ribosomal subunits to form the 40 S preinitiation complex (Chaudhuri *et al.* 1997). Although the loss of a translation factor within the fungi may seem surprising, previous analyses have established the loss of translation factor components in the *S. cerevisiae* lineage (Braun *et al.* 2000), and an ortholog of eIF1A is present in *S. cerevisiae* (Tif11p, see Wei *et al.* 1995), suggesting a distinct function for G6G8.9. The possible transcription factor encoded by G6G8.4 (Table 1) has a region of identity to a *Schizosaccharomyces pombe* protein that shows weak identity to the helix-turn-helix structure of *S. cerevisiae* Mbp1p (see Taylor *et al.* 1997) in profile searches. However, a protein much more closely related to G6G8.4 is present in *S. pombe*, an apparent outgroup to a clade containing *N. crassa* and *S. cerevisiae* (Bruns *et al.* 1992), suggesting the absence of an *S. cerevisiae* G6G8.4 ortholog through gene loss.

The observation of 4 of 13 genes showing possible loss in *S. cerevisiae* is surprising, since a previous survey based on EST data suggested that ~12% of *N. crassa* genes with detectable homologs were lost in the *S. cerevisiae* lineage (Braun *et al.* 2000). Although the higher proportion of genes in this category observed here may simply reflect sampling error, it is also possible that such predictions based upon EST data are biased in some manner.

The intergenic regions in the G6G8 portion of the *N. crassa* genome are substantially larger than comparable regions in the *S. cerevisiae* genome, as expected (Kupfer *et al.* 1997). However, the intergenic regions separating convergently transcribed genes are only slightly larger than comparable regions in the *S. cerevisiae* genome (Table 2), while those separating either divergently transcribed genes or genes transcribed in the same direction are substantially larger than comparable regions in the *S. cerevisiae* genome. Although there is substantial evidence linking the *SNZ* and *SNO* genes functionally (Galperin and Koonin 1997; Padilla *et al.* 1998; this work), their start codons do not appear to be unusually close for divergently transcribed genes (separated by 1992 bp).

The shortest intergenic region separates the NOT-56 homolog (G6G8.10) from a convergently transcribed gene related to an *E. nidulans* EST and a hypothetical *S. pombe* eIF1A-like ORF (G6G8.9). Surprisingly, a NOT-56 cDNA (SM1G12) shows substantial overlap (at least 180 nucleotides) with the adjacent G6G8.9 open reading frame. This overlap raises the possibility that these genes exhibit transcriptional interference similar to the

TABLE 1

Genes identified in cosmid G6G8 insert

ORF	Protein identification	Length (aa)	Best BLAST hit (organism)[a]	Best *S. cerevisiae* BLAST hit[b]	Exon locations (inferred or deduced)
1	Serine/threonine protein phosphatase	281	C22H10.04 (*S. pombe*) 74% id 1×10^{-118}	Ppg1p (YNR032w) 64% id 1×10^{-104}	2167–2299[c] 2415–2669 2736–3193
2	Hypothetical 12.6-kD protein, predicted from cDNA	110	None	None	3844–3512[d,e]
3	*rho* GDI (GDP dissociation inhibitor)	161	SPAC6F12.06 (*S. pombe*) 43% id 6×10^{-34}	Rdi1p (YDL135c) 36% id 4×10^{-30}	4893–4885[c,d,e] 4656–4311 4188–4058
4	52.4-kD protein possibly distantly related to Mbp1p transcription factor	483	SPBC19C7.10 (*S. pombe*) 30% id 9×10^{-16}	None (probable loss)	8139–8223[c] 8265–9631
5	3-Hydroxy isobutyrate dehydrogenase	338	CG15093 (*D. melanogaster*) 34% id 9×10^{-42}	None (probable loss)	11156–11263[c] 11322–11495 11723–12457
6	D-Amino acid oxidase	362	DAO (*F. solani*) 56% id 1×10^{-110}	None (probable loss)	14004–13982[c,e] 13852–12787
7	Thioredoxin	107	Thioredoxin (*E. nidulans*) 53% id 3×10^{-21}	Trx1p (YLR043c) 50% id 8×10^{-22}	19146–19124[c,d,e] 18989–18885 18688–18493
8	27.5-kD protein distantly related to 3-phosphoserine phosphatase	258	SPAC823.14 (*S. pombe*) 54% id 6×10^{-67}	YNL010w 53% id 4×10^{-64}	23396–22903[c,e] 22830–22682 22597–22464
9	18.7-kD protein distantly related to translation initiation factor	162	SPBC146.08c (*S. pombe*) 29% id 3×10^{-12}	None (probable loss)	24247–24735[c,d]
10	NOT-56 mannosyltransferase	442	SPAC7D4.06c (*S. pombe*) 43% id 9×10^{-77}	Rhk1p (YBL082C) 38% id 4×10^{-64}	26290–26122[c,d,e] 26058–25814 25732–24818
11	48.2-kD Gln/Pro-rich protein	426	None	None	29016–30296
12	*pdx-1* (*SNZ* homolog)	308	PYROA (*E. nidulans*) 67% id 1×10^{-106}	Snz3p (YFL059W) 58% id 2×10^{-82}	32022–31096[c,d,e]
13	*pdx-2* (*SNO* homolog)	252	PDX2 (*C. nicotianae*)[f] 48% id 9×10^{-59}	Sno2p (YNL334c) 38% id 4×10^{-32}	34015–34289[c,d,e] 34433–34916

aa, amino acid.

[a] Top hit from BLASTP search of the NCBI NR database conducted in September 2000. Percentage identity in the aligned region and expect (E) values are also presented. Organisms listed in this column are: Fungi, *C. nicotianae, E. nidulans*, *Fusarium solani*, *S. pombe*; animals, *D. melanogaster*. Six of the ORFs (G6G8.2, 3, 7, 10, 12, and 13) are represented by partial *N. crassa* cDNA sequences in the GenBank EST division.

[b] Top BLASTP hit from a search of annotated *S. cerevisiae* open reading frames. Percentage identity in the aligned region is also presented. Cases that correspond to probable gene loss in *S. cerevisiae* according to the criteria outlined by Braun *et al.* (2000) are also indicated.

[c] Exon locations established by BLASTX.

[d] Exon locations established on the basis of ESTs or cDNA clones.

[e] Encoded on complementary strand.

[f] From a recent GenBank submission (accession no. AAG09049), which annotated the *C. nicotianae SNO* homolog as a pyridoxine biosynthetic protein.

TABLE 2

Summary of intergenic regions in cosmid G6G8 insert

Type[a]	No.	Mean length (bp)	Range (bp)	% GC	Mean length in *S. cerevisiae*[b]
Divergent	4	2203	850–3245	50.9	618
Parallel	4	2385.5	213–4488	51.9	517
Convergent	4	400.5	82–873	46.1	326

[a] Intergenic regions are categorized on the basis of flanking genes, which can be divergently transcribed, transcribed in the same direction, or convergently transcribed.
[b] The *S. cerevisiae* data were taken from Dujon (1996).

convergently transcribed *S. cerevisiae POT1* and *YIL161w* genes (Puig *et al.* 1999).

The most closely spaced putative parallel transcription units (G6G8.2 and G6G8.3) may present an even more substantial transcriptional overlap. An mRNA (d4b03ne) that has a 3′ end downstream of G6G8.2 extends into the second exon of G6G8.3 and lacks a putative intron present in G6G8.3 (Table 1). Thus, it is possible that G6G8.2 mRNAs actually correspond to the 3′ untranslated region of G6G8.3, making our annotation of G6G8.2 as a protein coding region more tentative than the other genes in this region. On the basis of an in-frame stop codon in G6G8.3, verified in genomic and cDNA sequences, and similarity between G6G8.3 and known *rho* GDI homologs from other organisms, G6G8.2 apparently does not represent an extension of the G6G8.3 coding region.

Analysis of *SNZ* and *SNO* homologs: The *N. crassa SNZ* and *SNO* homologs were first identified as cDNAs by the *Neurospora* Genome Project at the University of New Mexico (Nelson *et al.* 1997). The genomic sequence reported here reveals that the *SNZ* homolog contains no introns, while the *SNO* homolog contains a single intron. The two genes are divergently transcribed and separated by 2 kbp (Table 1). There are two overlapping ORFs between the two genes that could each encode a polypeptide >100 amino acids (not shown). However, these ORFs lack strong consensus sequences for translational start, they do not exhibit codon preference typical for *N. crassa*, and they lack homologs in other organisms or corresponding EST sequences from *N. crassa*. Therefore, neither ORF was included among the predicted genes for the region.

Given mapping results that placed this region close to the *pdx-1* locus (linkage group IVR; Nelson *et al.* 1998), together with recent reports that *SNZ* homologs are involved in pyridoxine metabolism, we hypothesized that mutations in the *N. crassa SNZ* homolog were responsible for the *pdx-1* phenotype. Sequences obtained from several known mutants, designated *pdx-1*, strongly suggest that this is the case. Five of nine *pdx-1* mutants examined possessed mutations in the coding region of the *SNZ* homolog that either altered the amino acid sequence in highly conserved regions or caused a frameshift (Table 3). However, analysis of four *pdx-1* mutants revealed no mutations in the *SNZ* homolog but, instead, demonstrated mutations in conserved regions of the *SNO* homolog (Table 3). This represents the first direct evidence that mutations in *SNZ* and *SNO* homologs disrupt a shared metabolic pathway.

The conclusion that the observed mutations in *SNZ* and *SNO* homologs cause the pyridoxine-requiring phenotypes of the mutants examined is supported by complementation studies reported by Radford (1966) for six of the strains—FGSC numbers 1407, 1409, 1411, 1413, 1415, and 4055 (alleles 35405, 39106, 44602, 44204, 39706, and 37803, respectively). Working with alleles that were presumed to represent a single locus, Radford failed to obtain complementation between strains carrying alleles for which we identified corresponding mutations in the *SNZ* homolog (35405, 37803), as well as among strains with alleles with mutations in the *SNO* homolog (39106, 39706, 44602, 44204). In contrast, Radford reported successful complementation in tests where one strain possessed a mutation in the *SNZ* homolog while the other possessed a mutation in the *SNO* homolog. The one exception was a reported failure to obtain complementation between strains with alleles 44204 (*SNO*) and 37803 (*SNZ*), an anomaly for which there is no obvious explanation.

Strains 1409 and 1415, carrying alleles designated 39106 and 39706, possess identical mutations in the *SNO* homolog. It is likely that this reflects confusion in allele labeling in the laboratory history of these strains.

A shared function for *SNZ* and *SNO* homologs is further supported by high-resolution "intragenic" mapping data obtained by Radford (1968). Radford reported evidence for three separate clusters of mutations at the *pdx-1* locus, and he designated these clusters α, β, and γ (Figure 1). Our sequence analysis agrees with the chromosomal order suggested by Radford for alleles in the α group (35405, 37803), which possess mutations in the *SNZ* homolog, relative to alleles in the β (39106) and γ (44602, 44204) groups combined, which possess mutations in the *SNO* homolog (Table 3, Figure 1). Also in agreement with sequence analysis, Radford's results indicated that all β and γ mutations were closer to one another than to any mutations in the α group. However, the Radford study tentatively placed the β allele group proximal to the α group. Sequence results indicate in-

TABLE 3

Sequence analysis of *SNZ* and *SNO* homologs in *N. crassa pdx* mutants

FGSC strain no.	Allele and references	Mutation position and homologous gene	Change	Result[a]
1314	Y31393 (TATUM *et al.* 1950)	664 (*SNZ*)	G → C	FAA → FAP
1407	35405 (BARRATT *et al.* 1954; RADFORD 1965, 1966)	338 (*SNZ*)	T → C	EVL → EVS
1409[c]	39106	734 (*SNO*)	C → T	N → stop
1415[c]	39706 (BARRATT *et al.* 1954; RADFORD 1965, 1966)	734 (*SNO*)	C → T	N → stop
1411	44602 (RADFORD 1966)	208 (*SNO*)	G → A	GGE → GGK
1413	44204 (RADFORD 1965, 1966)	82 (*SNO*)	1-bp deletion (C)	Frameshift
1418	Y2329 (RADFORD 1967)	3′ of 338 (*SNZ*)	4-bp insertion (TCGA)	Frameshift
3261	Y30978 (TATUM *et al.* 1950)	272 (*SNZ*)	G → A	RIG → RID
4055	37803 (KAFER 1982)	401 (*SNZ*)	G → A	VCG → VCE

Genes are designated with respect to the homologous gene in *S. cerevisiae* (see DISCUSSION for nomenclature suggested for *N. crassa*). Nucleotide positions are given with respect to the deduced initiation codon (ATG) for the *SNZ* or *SNO* homolog.

[a] All mutations result in changes in conserved regions (refer to BRAUN *et al.* 1996).

[b] Strains 1409 and 1415 carry the same mutation (see text).

stead that the γ group is proximal to the α group. The positions of α, β, and γ groups approximated by Radford were based in part on recombination frequencies between *pdx* alleles and genetic markers flanking the *pdx* region. However, considering only the frequencies of prototrophs recovered in crosses with alternative *pdx* alleles, one RADFORD study (1968) was inconclusive with respect to the positions of β and γ relative to α, while another study (RADFORD 1967) in fact supported the order indicated by our sequence analysis of mutants.

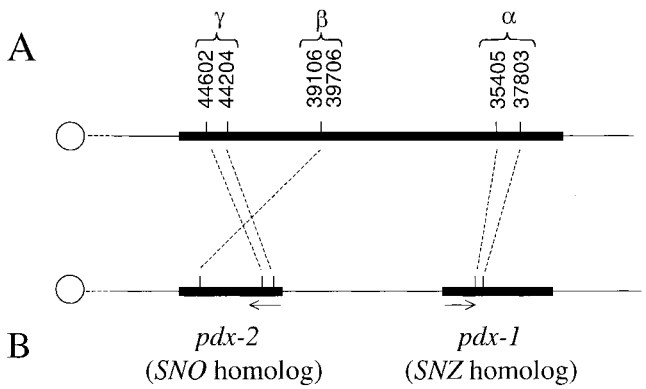

FIGURE 1.—Comparison of high-resolution allele mapping results and sequence analysis of pyridoxine-requiring mutants. (A) Map adapted from RADFORD (1968) showing relative positions of α, β, and γ allele groups. Only alleles examined in both studies are shown. Note: this map was not represented by the author as being to scale. The direction of the centromere is indicated by an open circle. (B) Mutant sequence analysis (Table 3). Positions of mutations observed in *pdx-1* (*SNZ* homolog) and *pdx-2* (*SNO* homolog) coding regions, shown approximately to scale. Arrows indicate direction of transcription.

DISCUSSION

Comments on annotation: Our attempt to identify genes in the G6G8 insert highlights the difficulties of annotation with filamentous fungi when only genomic sequence data are available, and it underscores the value of supplemental information. The 36-kbp region in question contains 106 ORFs that could encode peptides of at least 100 amino acids each. Thirty-eight of these ORFs begin with a start ATG. In contrast, the actual estimate for this region is 13 genes (Table 1). None of the ORFs excluded from the gene list in Table 1 exhibited strong *N. crassa* codon preference, nor did any produce a BLAST *E*-value $< 10^{-3}$. Several of the excluded ORFs overlapped verified genes, raising additional doubt with respect to possible protein-coding function. Eleven protein-coding genes could be verified by BLAST analyses revealing homology with known genes from other organisms or fungal ESTs (Table 1). An additional gene, not identified by BLAST analysis, was inferred from a long ORF (426 codons) with strong *N. crassa* codon preference. If additional protein-coding genes exist in this region, they were not identified, either because they do not exhibit strong codon preference or because they encode relatively short polypeptides. Further, the presence of an identifiable 5′ start ATG is not a reliable criterion for ORF identification due to the frequent occurrence of introns in the 5′ regions of *N. crassa* genes. This point is well illustrated by the genes identified in Table 1. Nine of the 13 annotated genes possess introns, 5 of which could be deduced by comparison with cDNA (EST) sequences. Among these 9 genes, 6 possess an initial intron within the first 100 codons, exemplifying the poor predictive value of a start ATG for gene finding in this organism.

Significance of observed gene density: *N. crassa* is a multicellular fungus with a complex life cycle that involves both asexual and sexual reproduction. It possesses a genome size of 42.9 Mbp (Orbach *et al.* 1988; Orbach 1992), nearly three times that of its ascomycete relative *S. cerevisiae*. In *N. crassa*, asexual reproduction involves the generation of two different types of conidia, while sexual reproduction involves the development of ascospores within a morphologically complex perithecium (Springer 1993). The developmental complexity and relatively large genome size of *N. crassa* suggest that it might possess a substantially larger number of genes than do unicellular fungi such as *S. cerevisiae* and *S. pombe*. Previous analyses suggested that at least some of these differences in genome complexity reflect gene loss in *S. cerevisiae* (Braun *et al.* 1998, 2000).

Although all recent estimates suggest substantially larger gene numbers for *N. crassa* and other filamentous ascomycetes than for *S. cerevisiae*, specific estimates for *N. crassa* differ. Kupfer *et al.* (1997) estimated that filamentous fungi typically harbor 8000–9000 genes and suggested that *N. crassa* has 9200 genes based on a nonlinear extrapolation of gene number from genome size. In contrast, Nelson *et al.* (1997) estimated a larger number of genes for *N. crassa*, up to 13,000, on the basis of gene densities in the mating-type and *qa*-cluster regions.

There exists a minimum of 12 protein-coding genes in the region represented by the 36,030-bp insert in cosmid G6G8, corresponding to a genetic unit of 3000 bp. Assuming 39×10^6 bp of genomic DNA, after subtracting rDNA repeats and other low complexity sequences (Krumlauf and Marzluf 1980; Nelson *et al.* 1997), the predicted total gene number is 13,000. This estimate is consistent with our previous estimate (Nelson *et al.* 1997) and with another recent estimate based on analysis of a distinct cosmid sequence (Kelkar *et al.* 2001).

Function and evolution of the *pdx-1* region: Our results demonstrate that the *pdx-1* mutant phenotype can derive from mutations in either the *SNZ* or *SNO* homolog of *N. crassa*. The coordinate function for these two genes was inferred for other organisms from previous studies of regulation and gene linkage. Our analysis of *N. crassa pdx-1* mutants provides confirming experimental evidence in support of this inference.

A nomenclature problem exists with respect to mutant alleles currently designated *pdx-1*. The three separate allele clusters identified by Radford (1968) in high-resolution mapping studies—designated α, β, and γ—were interpreted as intragenic on the basis of close physical proximity and shared phenotype. Sequence analyses demonstrate that α alleles possess mutations in the *SNZ* homolog, whereas β and γ alleles possess mutations in the *SNO* homolog. Alleles from α and β groups alike were among those originally described (Houlahan *et al.* 1949; Radford 1968). Given that pyridoxine metabolism was first linked experimentally to *SNZ* homologs (Ehrenshaft *et al.* 1999a; Osmani *et al.* 1999), we suggest that in *N. crassa* the *pdx-1* designation is most appropriate for the *SNZ* homolog.

An allele from the Radford γ group, 44204, was at one time designated *pdx-2* but was considered by Radford (1965) to belong to *pdx-1*. Allele 44204 and another allele from the Radford γ group, 44602, possess mutations in conserved regions of the *SNO* homolog (Table 3). We therefore suggest that the *pdx-2* designation is appropriate for the *SNO* homolog (Figure 1).

SNZ and *SNO* homologs are closely linked in diverse prokaryotes and eukaryotes. It has been proposed that in general such clustering in prokaryotes occurs with "selfish operons," operons whose products provide functions that are under weak or sporadic positive selection (Lawrence and Roth 1996). Because the functions encoded by such clusters are dispensable under certain environmental conditions, the genes encoding such functions may be subject to accumulation of deleterious mutations and loss. Gene clustering is thought to facilitate the horizontal transfer of an intact functional operon, allowing the acquisition or reacquisition of function. Support for a process analogous to the selfish operon theory exists for clustered genes in fungi. Often, clustered genes encode dispensable catabolic functions or components of secondary-metabolite pathways (Keller and Hohn 1997; Prade *et al.* 1997). Clustering may provide a mechanism for horizontal transfer or other forms of transfer not involving sexual reproduction among individuals of the same species.

The selfish operon model does not appear to provide an adequate explanation for the close linkage of *SNZ* and *SNO* homologs. Quite clearly, within the genera Neurospora and Emericella, *SNZ* homologs do not fit the profile of dispensable genes well. Mutations in the *SNZ* homologs in members of these genera create pyridoxine auxotropy, which to our knowledge has not been observed among thousands of wild-type strains. Furthermore, mutations in *SNZ* homologs increase susceptibility to oxidative stress (Ehrenshaft *et al.* 1999a,b), and *SNZ* homologs have been observed in all ascomycetes for which substantial genome sequencing has been performed. Phylogenetic tree-building analyses using predicted amino-acid sequences for *SNZ* and *SNO* homologs suggest that the evolution of these genes is compatible with organismal phylogeny (results not presented). Together, these observations make it unlikely that the physical linkage of *SNZ* and *SNO* genes in fungi reflects a recent horizontal transfer from prokaryotes. Instead, this linkage likely reflects selection for coordinate regulation.

Conclusion: Our analysis of this 36-kbp region of the *N. crassa* genome demonstrates that efforts in fungal genomics to identify coding regions and determine gene function will be most successful with combined approaches. Results illustrate the difficulties of annotation, given only genomic sequence data, and they reveal

the added value of information from cDNA sequences, biochemistry, bioinformatics, and classical genetics.

This study also underscores the diversity of processes underlying genome evolution. Two genes were identified with *N. crassa* paralogs, despite the relative paucity of duplicated genes in *N. crassa* (Nelson *et al.* 1997; Braun *et al.* 2000). Four of 13 genes appear to have been lost in *S. cerevisiae*, emphasizing the contribution of gene loss to the *S. cerevisiae* lineage (see Braun *et al.* 1998, 2000). Although the contribution of gene loss to the evolution of other small fungal genomes is not known at present, it is likely that gene loss has had a similar impact upon such genomes.

The close linkage of *SNZ* and *SNO* genes in many organisms, including *N. crassa*, signals the functional information present in the genomic context of genes. Although the correlation between location and function has long been appreciated in prokaryotes (reviewed by Aravind 2000), results such as those presented here suggest that this correlation will also be valuable in eukaryotes. The elucidation of evolutionary mechanisms driving correlation between gene location and function should aid in efforts to predict gene function.

We thank Dr. Alan Radford for very helpful comments during the course of *pdx-1* analyses. This work was supported by National Science Foundation grants HRD-9550649 (D.O.N., M.A.N., M.W.-W., and Robert K. Miller), MCB-9603902 (D.O.N.), IBN-9870878 (M.W.-W.) and MCB-9874488 (M.A.N.). A.E. and M.G. were supported in part by the Minority Biomedical Research Support program of the University of New Mexico (National Institutes of Health grant GM-52576). E.L.B. was supported in part by United States Department of Agriculture fellowship 1999-01582. G.S.S. was supported in part by a postdoctoral fellowship from the Ford Foundation. We gratefully acknowledge computer and computational support from the Albuquerque High Performance Computing Center at the University of New Mexico.

LITERATURE CITED

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:** 3389–3492.

Aravind, L., 2000 Guilt by association: contextual information in genome analysis. Genome Res. **10:** 1074–1077.

Barratt, R. W., D. Newmeyer, D. D. Perkins and L. Garnjobst, 1954 Map construction in *Neurospora crassa*. Adv. Genet. **6:** 1–93.

Braun, E. L., E. K. Fuge, P. A. Padilla and M. Werner-Washburne, 1996 A stationary phase gene in *Saccharomyces cerevisiae* is a member of a novel, highly conserved gene family. J. Bacteriol. **178:** 6865–6872.

Braun, E. L., S. Kang, M. A. Nelson and D. O. Natvig, 1998 Identification of the first fungal annexin: analysis of annexin gene duplications and implications for eukaryotic evolution. J. Mol. Evol. **47:** 531–543.

Braun, E. L., A. L. Halpern, M. A. Nelson and D. O. Natvig, 2000 Large scale comparison of fungal sequence information: mechanisms of innovation in *Neurospora crassa* and gene loss in *Saccharomyces cerevisiae*. Genome Res. **10:** 416–430.

Bruchez, J. J. P., J. Eberle and V. E. A. Russo, 1993a Regulatory sequences in the transcription of *Neurospora crassa* genes: CAAT box, TATA box, introns, poly(A) tail formation sequences. Fungal Genet. Newsl. **40:** 89–96.

Bruchez, J. J. P., J. Eberle and V. E. A. Russo, 1993b Regulatory sequences involved in the translation of *Neurospora crassa* mRNA: Kozak sequences and stop codons. Fungal Genet. Newsl. **40:** 85–88.

Bruns, T. D., R. Vilgalys, S. M. Barns, D. Gonzalez, D. S. Hibbett *et al.*, 1992 Evolutionary relationships within the fungi: analyses of nuclear small subunit rRNA sequences. Mol. Phylogenet. Evol. **1:** 531–543.

Chary, P., R. A. Hallewell and D. O. Natvig, 1990 Structure, exon pattern and chromosome mapping of the gene for cytosolic copper-zinc superoxide dismutase (*sod-1*) from *Neurospora crassa*. J. Biol. Chem. **265:** 18961–18967.

Chaudhuri, J., K. Si and U. Maitra, 1997 Function of eukaryotic translation initiation factor 1A (eIF1A) (formerly called eIF-4C) in initiation of protein synthesis. J. Biol. Chem. **272:** 7883–7891.

Davis, R. H., and F. J. deSerres, 1970 Genetic and microbiological research techniques for *Neurospora crassa*. Methods Enzymol. **17:** 79–143.

Dujon, B., 1996 The yeast genome project: what did we learn? Trends Genet. **12:** 263–270.

Ehrenshaft, M., P. Bilski, M. Y. Li, C. F. Chignell and M. E. Daub, 1999a A highly conserved sequence is a novel gene involved in de novo vitamin B6 biosynthesis. Proc. Natl. Acad. Sci. USA **96:** 9374–9378.

Ehrenshaft, M., K.-R. Chung, A. E. Jenns and M. E. Daub, 1999b Functional characterization of *SOR1*, a gene required for resistance to photosensitizing toxins in the fungus *Cercospora nicotianae*. Curr. Genet. **34:** 478–485.

Fuge, E. K., E. L. Braun and M. Werner-Washburne, 1994 Protein synthesis in long-term stationary-phase cultures of *Saccharomyces cerevisiae*. J. Bacteriol. **176:** 5802–5813.

Galperin, M. Y., and E. V. Koonin, 1997 Sequence analysis of an exceptionally conserved operon suggests enzymes for a new link between histidine and purine biosynthesis. Mol. Microbiol. **24:** 443–445.

Green, P., 1996 PHRAP ("phragment assembly program," or "Phil's revised assembly program"). Department of Molecular Biotechnology, University of Washington, Seattle.

Green, P., and B. Ewing, 1997 PHRED. Department of Molecular Biotechnology, University of Washington, Seattle.

Houlahan, M. B., G. W. Beadle and H. G. Calhoun, 1949 Linkage studies with biochemical mutants of *Neurospora crassa*. Genetics **34:** 493–507.

Kelkar, H. S., J. Griffith, M. E. Case, S. F. Covert, R. D. Hall *et al.*, 2001 The *Neurospora crassa* genome: cosmid libraries sorted by chromosome. Genetics **157:** 979-990

Keller, N. P., and T. M. Hohn, 1997 Metabolic pathway gene clusters in filamentous fungi. Fungal Genet. Biol. **21:** 17–29.

Krasko, A., H. C. Schröder, S. Perovic, R. Steffen, M. Kruse *et al.*, 1999 Ethylene modulates gene expression in cells of the marine sponge *Suberites domuncula* and reduces the degree of apoptosis. J. Biol. Chem. **274:** 31524–31530.

Krumlauf, R., and G. A. Marzluf, 1980 Genome organization and characterization of the repetitive and inverted repeat DNA sequences in *Neurospora crassa*. J. Biol. Chem. **255:** 1138–1145.

Kupfer, D. M., C. A. Reece, S. W. Clifton, B. A. Roe and R. A. Prade, 1997 Multicellular ascomycetous fungal genomes contain more than 8000 genes. Fungal Genet. Biol. **21:** 364–372.

Lawrence, J. G., and J. R. Roth, 1996 Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics **143:** 1843–1860.

Nelson, M. A., S. Kang, E. L. Braun, M. E. Crawford, P. Dolan *et al.*, 1997 Expressed sequences from conidial, mycelial, and sexual stages of *Neurospora crassa*. Fungal Genet. Biol. **21:** 348–363.

Nelson, M. A., M. E. Crawford and D. O. Natvig, 1998 Restriction polymorphism maps of *Neurospora crassa*: 1998 update. Fungal Genet. Newsl. **45:** 44–54.

Orbach, M. J., 1994 A cosmid with a Hy^R marker for fungal library construction and screening. Gene **150:** 159–162.

Orbach, M. J., D. Vollrath, R. W. Davis and C. Yanofsky, 1988 An electrophoretic karyotype of *Neurospora crassa*. Mol. Cell. Biol. **8:** 1469–1473.

Osmani, A. H., G. S. May and S. A. Osmani, 1999 The extremely conserved *pyroA* gene of *Aspergillus nidulans* is required for pyridoxine synthesis and is required indirectly for resistance to photosensitizers. J. Biol. Chem. **274:** 23565–23569.

Padilla, P. A., E. K. Fuge, M. E. Crawford, A. Errett and M. Werner-Washburne, 1998 The highly conserved, coregulated

SNO and *SNZ* gene families in *Saccharomyces cerevisiae* respond to nutrient limitation. J. Bacteriol. **180:** 5718–5726.

Prade, R. A., J. Griffith, K. Kochut, J. Arnold and W. E. Timberlake, 1997 In vitro reconstruction of the *Aspergillus (= Emericella) nidulans* genome. Proc. Natl. Acad. Sci. USA **94:** 14564–14569.

Puig, S., J. E. Pérez-Ortín and E. Matallana, 1999 Transcriptional and structural study of a region of two convergent overlapping yeast genes. Curr. Microbiol. **39:** 369–373.

Radford, A., 1965 Heterokaryon complementation among the pyridoxine auxotrophs of *Neurospora crassa.* Can. J. Genet. Cytol. **7:** 472–477.

Radford, A., 1966 Further studies on complementation at the *pdx-1* locus of *Neurospora crassa.* Can. J. Genet. Cytol. **8:** 672–676.

Radford, A., 1967 Prototroph frequencies from crosses between pyridoxine auxotrophs. Neurospora Newsl. **11**: 4.

Radford, A., 1968 High resolution recombination analysis of the pyridoxine-1 locus of *Neurospora.* Can. J. Genet. Cytol. **10:** 893–897.

Roe, B. A., J. S. Crabtree and A. S. Khan, 1996 *DNA Isolation and Sequencing.* John Wiley & Sons, New York.

Sambrook, J., E. F. Fritsch and T. Maniatis, 1989 *Molecular Cloning: A Laboratory Manual,* Ed. 2. Cold Spring Harbor Laboratory Press, Plainview, NY.

Sivasubramaniam, S., V. M. Vanniasingham, C.-T. Tan and N.-H. Chua, 1995 Characterization of HEVER, a novel stress-induced gene from *Hevea brasiliensis.* Plant Mol. Biol. **29:** 173–178.

Springer, M. L., 1993 Genetic control of fungal differentiation: the three sporulation pathways of *Neurospora crassa.* Bioessays **15:** 365–374.

Tatum, E. L., R. W. Barratt, N. Fries and D. Bonner, 1950 Biochemical mutant strains of *Neurospora* produced by physical and chemical treatment. Am. J. Bot. **37:** 38–46.

Taylor, I. A., M. K. Treiber, L. Olivi and S. J. Smerdon, 1997 The X-ray structure of the DNA-binding domain from the *Saccharomyces cerevisiae* cell-cycle transcription factor Mbp1 at 2.1 Å resolution. J. Mol. Biol. **272:** 1–8.

Wei, C. L., M. Kainuma and J. W. Hershey, 1995 Characterization of yeast translation initiation factor 1A and cloning of its essential gene. J. Biol. Chem. **270:** 22788–22794.

Communicating editor: J. Arnold