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

Laura E. Bean,*1 William H. Dvorachek, Jr.,* Edward L. Braun,*1,2 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
1National Center for Genome Resources, Santa Fe, New Mexico 87505

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 (Ehrenshافت et al. 1999a,b) and Aspergillus nidulans (Osmanî 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 hypersensitivity to singlet oxygen-generating agents (Ehrenshhaft et al. 1999b).

In many organisms, genes encoding Snz homologs are closely linked to genes encoding Snc 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 Snc proteins...
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 pds-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 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 GG68 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 32P-labeled DNA from a cDNA clone carrying the N. crassa SNZ homolog.

Subcloning of cosmid GG68: Escherichia coli cells containing the GG68 cosmid were grown at 37°C 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, Sau3AI 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 BamHI-digested pUC-18 using a standard ligation protocol (Sambrook et al. 1989). Ligated DNA was transformed into INVoF’ cells (Invitrogen, San Diego). In addition to using Sau3AI, fragments were produced for subcloning by complete digestion of the GG68 cosmid DNA using four different restriction enzymes with 6-bp recognition sequences followed by dephosphorylation. One procedure used cosmid DNA digested with HindIII and EcoRI, while another used KpnI and PstI. 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°C. 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, detyderminator 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 Sequercher 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 GG68 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 [expressed 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 pds-1 mutants: Strains carrying various pds-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 de Serres 1970). Genomic DNA was prepared using the Puregene D-5000A plant DNA isolation kit (Gentra Systems, Research Triangle Park, NC). The N. crassa SNZ homolog was amplified from genomic DNA preparations by polymerase chain reaction (PCR) using forward primer 5’-AAGACCCCTAAGCTCTCAATCGTGGT-3’ and reverse primer 5’-TCCAAGGCCCTTTTTTAGTCTTCG-3’. Sequences were obtained using forward and reverse PCR primers along with internal primers 5’-GGCTGCACTACATGCGAGAAG-3’ and 5’-TCCCTGAGAACCTCAACTCCG-3’. The N. crassa SNO homolog was amplified by PCR using forward primer 5’-CCCTGGTGTAACCCGAGACGTATCG-3’ and reverse primer 5’-AACCGTGACCCCTAATGCTGC-3’. Sequences were obtained using forward and reverse PCR primers along with
internal primers 5'-AGTCTTTTTTTTTTCTTTTTCCTAAACCGG-3' and 5'-ACTTTCCAGCTGGTGCCTTGTA-3'. Primers were tested and wild-type SNZ1 and SNO1 homolog sequences were confirmed with \textit{N. crassa} strain 74-OR23-1A.

\textbf{RESULTS}

\textbf{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 \textit{N. crassa} codon preference (e.g., CHARY \textit{et al.} 1990), and searches for consensus sequences associated with translational start sites (BRUCHEZ \textit{et al.} 1993a) and intron splicing (BRUCHEZ \textit{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 \textit{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 \textit{N. crassa} codon preference and lacks similarity to known genes from other organisms. This ORF is contained within certain \textit{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 \textit{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 \textit{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 \textit{N. crassa}. A different 3-hydroxyisobutyrate dehydrogenase (3HD) homolog was found earlier by the \textit{Neurospora} Genome Project in a cDNA clone (NELSON \textit{et al.} 1997). There is only moderate sequence similarity between the two predicted \textit{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 \textit{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 \textit{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 \textit{Emeriella} (= \textit{Aspergillus}) \textit{nidulans}]

There are four genes in G6G8 that appear to lack \textit{S. cerevisiae} orthologs, despite evidence suggesting they were present in the common ancestor of \textit{N. crassa} and \textit{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 \textit{S. cerevisiae} homolog, the criterion used to establish gene loss by BRAUN \textit{et al.} (2000). Two of the proteins are probable structural enzymes [3HD and \textit{d}-amino acid oxidase (DAO)]. The third appears distantly related to translation initiation factor IA, a protein essential for transfer of the initiator tRNA to 40 S ribosomal subunits to form the 40 S preinitiation complex (CHAUDHURI \textit{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 \textit{S. cerevisiae} lineage (BRAUN \textit{et al.} 2000), and an ortholog of elf1A is present in \textit{S. cerevisiae} (TIF11p, see WEl \textit{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 \textit{Schizosaccharomyces pombe} protein that shows weak identity to the helix-turn-helix structure of \textit{S. cerevisiae} Mbp1p (see TAYLOR \textit{et al.} 1997) in profile searches. However, a protein much more closely related to G6G8.4 is present in \textit{S. pombe}, an apparent outgroup to a clade containing \textit{N. crassa} and \textit{S. cerevisiae} (BRUNS \textit{et al.} 1992), suggesting the absence of an \textit{S. cerevisiae} G6G8.4 ortholog through gene loss.

The observation of 4 of 13 genes showing possible loss in \textit{S. cerevisiae} is surprising, since a previous survey based on EST data suggested that \textit{\sim}12% of \textit{N. crassa} genes with detectable homologs were lost in the \textit{S. cerevisiae} lineage (BRAUN \textit{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 \textit{N. crassa} genome are substantially larger than comparable regions in the \textit{S. cerevisiae} genome, as expected (KUPPER \textit{et al.} 1997). However, the intergenic regions separating convergently transcribed genes are only slightly larger than comparable regions in the \textit{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 \textit{S. cerevisiae} genome. Although there is substantial evidence linking the SNZ and SNO genes functionally (GALPERIN and KOOIN 1997; PADILLA \textit{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 \textit{E. nidulans} EST and a hypothetical \textit{S. pombe} elf1A-like ORF (G6G8.9). Surprisingly, a NOT-56 cDNA (SM1G12) shows substantial overlap (at least 130 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

<table>
<thead>
<tr>
<th>ORF</th>
<th>Protein identification</th>
<th>Length (aa)</th>
<th>Best BLAST hit (organism)</th>
<th>Best S. cerevisiae BLAST hit</th>
<th>Exon locations (inferred or deduced)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Serine/threonine protein phosphatase</td>
<td>281</td>
<td>C2H10.04 (S. pombe)</td>
<td>Ppg1p (YNR032w)</td>
<td>2167–2299</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>74% id</td>
<td>64% id</td>
<td>2415–2669</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1 × 10^{-18}</td>
<td>1 × 10^{-104}</td>
<td>2736–3193</td>
</tr>
<tr>
<td>2</td>
<td>Hypothetical 12.6-kD protein, predicted from cDNA</td>
<td>110</td>
<td>None</td>
<td>None</td>
<td>3844–3512</td>
</tr>
<tr>
<td>3</td>
<td>rho GDI (GDP dissociation inhibitor)</td>
<td>161</td>
<td>SPAC6F12.06 (S. pombe)</td>
<td>Rdi1p (YDL135c)</td>
<td>4893–4885</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>43% id</td>
<td>36% id</td>
<td>4656–4311</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6 × 10^{-34}</td>
<td>4 × 10^{-30}</td>
<td>4188–4058</td>
</tr>
<tr>
<td>4</td>
<td>52.4-kD protein possibly distantly related to Mbp1p</td>
<td>483</td>
<td>SPBC19C7.10 (S. pombe)</td>
<td>None</td>
<td>8139–8223</td>
</tr>
<tr>
<td></td>
<td>transcription factor</td>
<td></td>
<td>30% id</td>
<td>(probable loss)</td>
<td>8265–9631</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9 × 10^{-16}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3-Hydroxy isobutyrate dehydrogenase</td>
<td>338</td>
<td>CG15095 (D. melanogaster)</td>
<td>None</td>
<td>11156–11263</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>34% id</td>
<td>(probable loss)</td>
<td>11322–11495</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9 × 10^{-12}</td>
<td></td>
<td>11723–12457</td>
</tr>
<tr>
<td>6</td>
<td>ω-Amino acid oxidase</td>
<td>362</td>
<td>DAO (F. solani)</td>
<td>None</td>
<td>14004–13982</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>56% id</td>
<td>(probable loss)</td>
<td>13852–12787</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1 × 10^{-110}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Thioredoxin</td>
<td>107</td>
<td>Thioredoxin (E. nidulans)</td>
<td>Trx1p (YLR043c)</td>
<td>19146–19124</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>53% id</td>
<td>50% id</td>
<td>18989–18885</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3 × 10^{-71}</td>
<td>8 × 10^{-22}</td>
<td>18688–18493</td>
</tr>
<tr>
<td>8</td>
<td>27.5-kD protein distantly related to 3-phosphoserine</td>
<td>258</td>
<td>SPAC823.14 (S. pombe)</td>
<td>YNL010w</td>
<td>23396–22903</td>
</tr>
<tr>
<td></td>
<td>phosphatase</td>
<td></td>
<td>54% id</td>
<td>53% id</td>
<td>22830–22682</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6 × 10^{-67}</td>
<td>4 × 10^{-64}</td>
<td>22597–22464</td>
</tr>
<tr>
<td>9</td>
<td>18.7-kD protein distantly related to translation initiation</td>
<td>162</td>
<td>SPBC146.08c (S. pombe)</td>
<td>None</td>
<td>24247–24735</td>
</tr>
<tr>
<td></td>
<td>factor</td>
<td></td>
<td>29% id</td>
<td>(probable loss)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3 × 10^{-12}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>NOT-56 mannosyltransferase</td>
<td>442</td>
<td>SPAC7D4.06c (S. pombe)</td>
<td>Rhk1p (YBL082C)</td>
<td>26290–26122</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>43% id</td>
<td>38% id</td>
<td>26058–25814</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9 × 10^{-77}</td>
<td>4 × 10^{-64}</td>
<td>25732–24818</td>
</tr>
<tr>
<td>11</td>
<td>48.2-kD Gln/Pro-rich protein</td>
<td>426</td>
<td>None</td>
<td>None</td>
<td>29016–30296</td>
</tr>
<tr>
<td>12</td>
<td>pdx-1 (SNZ homolog)</td>
<td>308</td>
<td>PYROA (E. nidulans)</td>
<td>Snz3p (YFL059W)</td>
<td>32022–31096</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>67% id</td>
<td>58% id</td>
<td>33178–32845</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1 × 10^{-106}</td>
<td>2 × 10^{-82}</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>pdx-2 (SNO homolog)</td>
<td>252</td>
<td>PDX2 (C. nicotianae)</td>
<td>Snp2p (YNL334c)</td>
<td>34015–34289</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>48% id</td>
<td>38% id</td>
<td>34433–34916</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9 × 10^{-59}</td>
<td>4 × 10^{-32}</td>
<td></td>
</tr>
</tbody>
</table>

aa, amino acid.

* 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.

* 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.

* Exon locations established by BLASTX.

* Exon locations established on the basis of ESTs or cDNA clones.

* Encoded on complementary strand.

* From a recent GenBank submission (accession no. AAG09049), which annotated the C. nicotianae SNO homolog as a pyridoxine biosynthetic protein.
convergently transcribed \textit{S. cerevisiae} \textit{POT1} and \textit{YIL161w} genes (Puig \textit{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 (dHb03ne) 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 \textit{rho} GDI homologs from other organisms, G6G8.2 apparently does not represent an extension of the G6G8.3 coding region.

\textbf{Analysis of \textit{SNZ} and \textit{SNO} homologs:} The \textit{N. crassa} \textit{SNZ} and \textit{SNO} homologs were first identified as cDNAs by the \textit{Neurospora} Genome Project at the University of New Mexico (Nelson \textit{et al.} 1997). The genomic sequence reported here reveals that the \textit{SNZ} homolog contains no introns, while the \textit{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 \textit{N. crassa}, and they lack homologs in other organisms or corresponding EST sequences from \textit{N. crassa}. Therefore, neither ORF was included among the predicted genes for the region.

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

The conclusion that the observed mutations in \textit{SNZ} and \textit{SNO} homologs cause the pyridoxine-requiring phenotype 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 \textit{SNZ} homolog (35405, 37803), as well as among strains with alleles with mutations in the \textit{SNO} homolog (39106, 39706, 44602, 44204). In contrast, Radford reported successful complementation in tests where one strain possessed a mutation in the \textit{SNZ} homolog while the other possessed a mutation in the \textit{SNO} homolog. The one exception was a reported failure to obtain complementation between strains with alleles 44204 (\textit{SNO}) and 37803 (\textit{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 \textit{SNO} homolog. It is likely that this reflects confusion in allele labeling in the laboratory history of these strains.

A shared function for \textit{SNZ} and \textit{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 \textit{pdx-I} locus, and he designated these clusters \textalpha, \textbeta, and \textgamma (Figure 1). Our sequence analysis agrees with the chromosomal order suggested by Radford for alleles in the \textalpha group (35405, 37803), which possess mutations in the \textit{SNZ} homolog, relative to alleles in the \textbeta (39106) and \textgamma (44602, 44204) groups combined, which possess mutations in the \textit{SNO} homolog (Table 3, Figure 1). Also in agreement with sequence analysis, Radford's results indicated that all \textbeta and \textgamma mutations were closer to one another than to any mutations in the \textalpha group. However, the Radford study tentatively placed the \textbeta allele group proximal to the \textalpha group. Sequence results indicate in-

\begin{table}
\centering
\caption{Summary of intergenic regions in cosmid G6G8 insert}
\begin{tabular}{llllll}
\hline
Type\textsuperscript{a} & No. & Mean length (bp) & Range (bp) & % GC & Mean length in \textit{S. cerevisiae}\textsuperscript{b} \\
\hline
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 \\
\hline
\end{tabular}
\textsuperscript{a}Intergenic regions are categorized on the basis of flanking genes, which can be divergently transcribed, transcribed in the same direction, or convergently transcribed.
\textsuperscript{b}The \textit{S. cerevisiae} data were taken from Dujon (1996).
\end{table}

\textit{N. crassa pdx-I Region}
TABLE 3
Sequence analysis of SNZ and SNO homologs in N. crassa pdx mutants

<table>
<thead>
<tr>
<th>FGSC strain no.</th>
<th>Allele and references</th>
<th>Mutation position and homologous gene</th>
<th>Change</th>
<th>Result*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1314</td>
<td>Y31393 (Tatum et al. 1950)</td>
<td>664 (SNZ) G → C</td>
<td>FAA → FAP</td>
<td></td>
</tr>
<tr>
<td>1407</td>
<td>35405 (Barratt et al. 1954; Radford 1965, 1966)</td>
<td>338 (SNZ) T → C</td>
<td>EVL → EVS</td>
<td></td>
</tr>
<tr>
<td>1409</td>
<td>39106</td>
<td>734 (SNO) C → T</td>
<td>N → stop</td>
<td></td>
</tr>
<tr>
<td>1415</td>
<td>39706 (Barratt et al. 1954; Radford 1965, 1966)</td>
<td>734 (SNO) C → T</td>
<td>N → stop</td>
<td></td>
</tr>
<tr>
<td>1411</td>
<td>44602 (Radford 1966)</td>
<td>208 (SNO) G → AG</td>
<td>GGE → GGK</td>
<td></td>
</tr>
<tr>
<td>1413</td>
<td>44204 (Radford 1965, 1966)</td>
<td>82 (SNO) 1-bp deletion (C)</td>
<td>Frameshift</td>
<td></td>
</tr>
<tr>
<td>1418</td>
<td>Y2329 (Radford 1967)</td>
<td>3′ of 338 (SNZ) 4-bp insertion (TCGA)</td>
<td>Frameshift</td>
<td></td>
</tr>
<tr>
<td>3261</td>
<td>Y30978 (Tatum et al. 1950)</td>
<td>272 (SNZ) G → A</td>
<td>RIG → RID</td>
<td></td>
</tr>
<tr>
<td>4055</td>
<td>37803 (Kaf er 1982)</td>
<td>401 (SNZ) G → A</td>
<td>VCG → VCE</td>
<td></td>
</tr>
</tbody>
</table>

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.

* All mutations result in changes in conserved regions (refer to Bra u n et al. 1996).

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

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 < 10⁻³¹. 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. Kupper 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 (Houihan 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 auxotrophy, 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 SVO 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 *fotA* analyses. This work was supported by National Science Foundation grants HRD-9556649 (D.O.N., M.A.N., M.W.-W., and Roberts K. Miller), MCB-9602992 (M.A.N.), 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**


Bruchez, J. J. P., J. Eberle and V. E. A. Russo, 1993b Regulatory sequences involved in the transcription of *Neurospora crassa* mRNA.


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.


