Volume 17 Number 22 1989 Nucleic Acids Research

A group II intron in the *Neurospora* mitochondrial *coI* gene: nucleotide sequence and implications for splicing and molecular evolution

Deborah J.Field+, Adele Sommerfield, Barry J.Saville and Richard A.Collins*

Department of Botany and Centre for Plant Biotechnology, University of Toronto, Toronto, Ontario M5S 3B2, Canada

Received August 14, 1989; Revised and Accepted October 18, 1989 EMBL accession no. X14669

ABSTRACT

The temperature-sensitive *Neurospora* nuclear mutant *cyt18-1* is deficient in splicing many Group I mitochondrial introns when grown at its non-permissive temperature; however, splicing of intron 1 in the *coI* gene of the Adiopodoume (formerly called North Africa) strain is unaffected (R.A. Collins and A.M. Lambowitz, J. Mol. Biol. 184: 413–428, 1985). Here we show that *coI* intron 1 is a typical Group II intron, the only one identified to date in *Neurospora*. The differential effect of the *cyt18-1* mutation suggests that splicing of certain introns could be regulated independently of others by nuclear-encoded proteins. The intron contains a long open reading frame (ORF) resembling that of the *Neurospora* Mauriceville mitochondrial plasmid. The intron and plasmid ORFs share unusual features of codon usage that suggest both evolved outside of the *Neurospora* mitochondrial genetic system.

INTRODUCTION

Introns in mitochondrial (mt) genes can be classified as Group I or Group II based on the presence of short conserved sequences that allow the formation of group-specific secondary and tertiary structures. Each group has its own RNA-catalyzed splicing mechanism for which at least some of these structural features have been shown to be important (reviewed in 1).

Even though splicing of both groups of introns is RNA-catalyzed, proteins are often required for splicing. These proteins can be required for splicing of a specific intron or multiple introns (reviewed in 2). In *Neurospora*, the *cyt18* nuclear gene encodes a protein required *in vivo* for the splicing of several mt introns in at least three genes (3,4); all of these are Group I introns (5–7; Collins, unpublished). For some introns (e.g. in the large rRNA gene) the cyt18 protein is even required for splicing *in vitro* (8), while for others (e.g., *cob* intron 1) the need for the cyt18 protein can be overcome by appropriate protein-free reactions conditions (9). Such trans-acting proteins are thought to be involved in formation of the higher order structure required for splicing (10).

Intron 1 in the *coI* gene of the Adiopodoume strain of *Neurospora* is the first intron discovered whose splicing *in vivo* appears to be unaffected by the *cyt18-1* mutation (4). This intron is different from most other *Neurospora* mt introns in that it is very large (2.6 kb) and it accumulates to a high concentration as a covalently-closed, single-stranded RNA after excision from pre-mRNA (4,11). To investigate a possible explanation for the *cyt18*-independent splicing of intron 1, we have determined and analyzed its nucleotide sequence.

MATERIALS AND METHODS.

Strains and Growth Conditions. All strains were obtained from the Fungal Genetics Stock Center. The Adiopodoume strain (FGSC # 430) was formerly named North Africa I. Other strains (and their stock center numbers in parentheses) used were: *N. intermedia* Fiji N6-6 (435); Kurubara Shettihally-1f (1808); Kalastwadi-1c (1810); Mysore-1 (1820); Mysore-1e (1821); Varkud-1b (1822); Varkud-1c (1823); Labelle-1b (1940); Chickkadana-1 (2495); Chickkadana-1j (2496); Nandi Hill 1-4a (2653); Fred-6 (3213); *N. crassa* Lahore-1 (1824); Lahore-1b (1825); Groveland-1c (1945); Mauriceville-1c (2225); Mauriceville-1d (2226); Aarey-1e (2499); Aarey-1 (2500); Spurger-3 (3201); Saratoga-11 (3226). Cells were grown in Vogel's medium according to standard procedures (12).

Cloning and sequencing. Mitochondrial DNA and RNA were isolated from flotation-gradient-purified mitochondria (13) by the UNSET-phenol procedure (14). Restriction fragments containing parts of *coI* intron I and flanking regions were cloned into pUC18/19 or M13mp18/19 (15). DNA sequence was obtained from single- or double-stranded templates using the dideoxy method (16,17). Overlapping clones obtained by unidirectional deletions (18,19) and/or subcloning were used to determine the complete sequence of both strands. dITP was substituted for dGTP in some sequencing reactions (20). The IBI/Pustell programs (International Biotechnologies, Inc., New Haven, Ct., USA) were used for analysis of the DNA sequence.

Northern hybridization. Approximately 10μg of total mt nucleic acids were denatured by glyoxalation and electrophoresed on 1.4% agarose gels (21), transferred to a nitrocellulose membrane and hybridized with cloned mt DNA restriction fragment (pNAA8) labeled with α-^{32}P-dATP by the random primer procedure (22).

RESULTS AND DISCUSSION.

Intron 1 in the coI gene is a Group II intron.

Previous restriction enzyme mapping and heteroduplex electron microscopy showed that the Adiopodoume natural isolate contains four introns in its *coI* gene (Figure 1 and ref. 11), in contrast to the standard laboratory strain 74A which contains no introns (23,24). Intron 1 is also present in a small number of other *Neurospora* natural isolates: Southern hybridizations (not shown) of EcoRI-digested mtDNAs with ^{32}P-labeled intron specific probe pNAA8 (see Figure 1) identified two *Neurospora intermedia* isolates, Varkud-1c (FGSC # 1823) and Chickkadana-1j (FGSC # 2496), both collected from Karnataka, India,

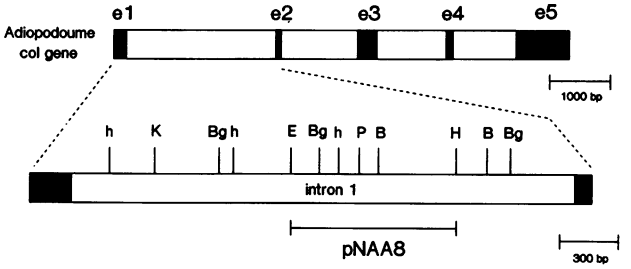


Figure 1. Organization of the *coI* gene in natural isolate Adiopodoume. Exons are represented by filled boxes, introns by open boxes. Some of the restriction sites used for subcloning and sequencing are indicated in the enlarged diagram. B=BamHI; Bg=BglII; E=EcoRI; H=HindIII; h=HincII; K=KpnI; P=PstI. Subclone pNAA8 was used as a probe for Northern hybridizations (see Figure 2).

that contained the intron. mtDNA from other isolates (collected from India, as well as North America and Fiji) of *N. intermedia* (FGSC # 435, 1808, 1810, 1820, 1821, 1822, 1940, 2495, 2653, 3213) and *N. crassa* (1824, 1825, 1945, 2225, 2226, 2499, 2500, 3201, 3226) did not hybridize.

In both the Adiopodoume and Varkud-1c strains (Chickkadana-1j has not been investigated) the excised intron is present at very high concentration, sufficient to be visible by ethidium bromide staining after agarose gel electrophoresis of total mitochondrial RNA (Figure 2). Introns of a similar large size in some yeast (25,26) and *Podospora* (27,28) mt genes that accumulate after excision as lariats (29−32), have been found to be Group II introns.

Analysis of the primary sequence (Figure 3) and possible secondary and tertiary structure (Figure 4) shows that Adiopodoume *coI* intron 1 is indeed a typical Group II intron, the

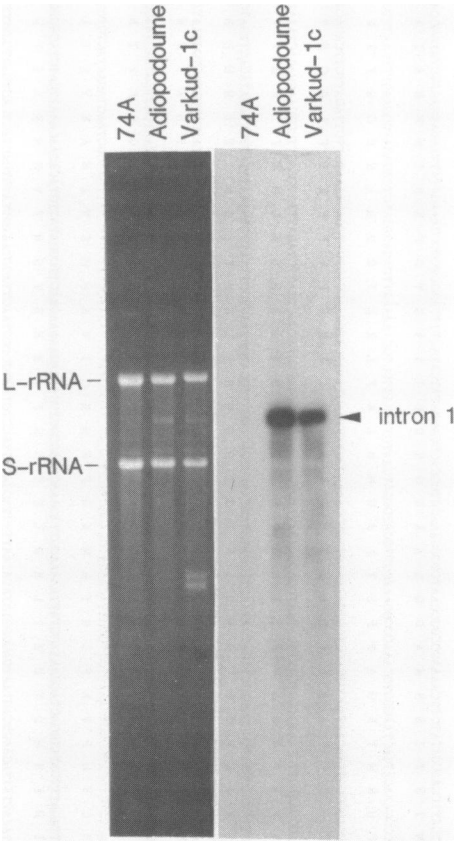
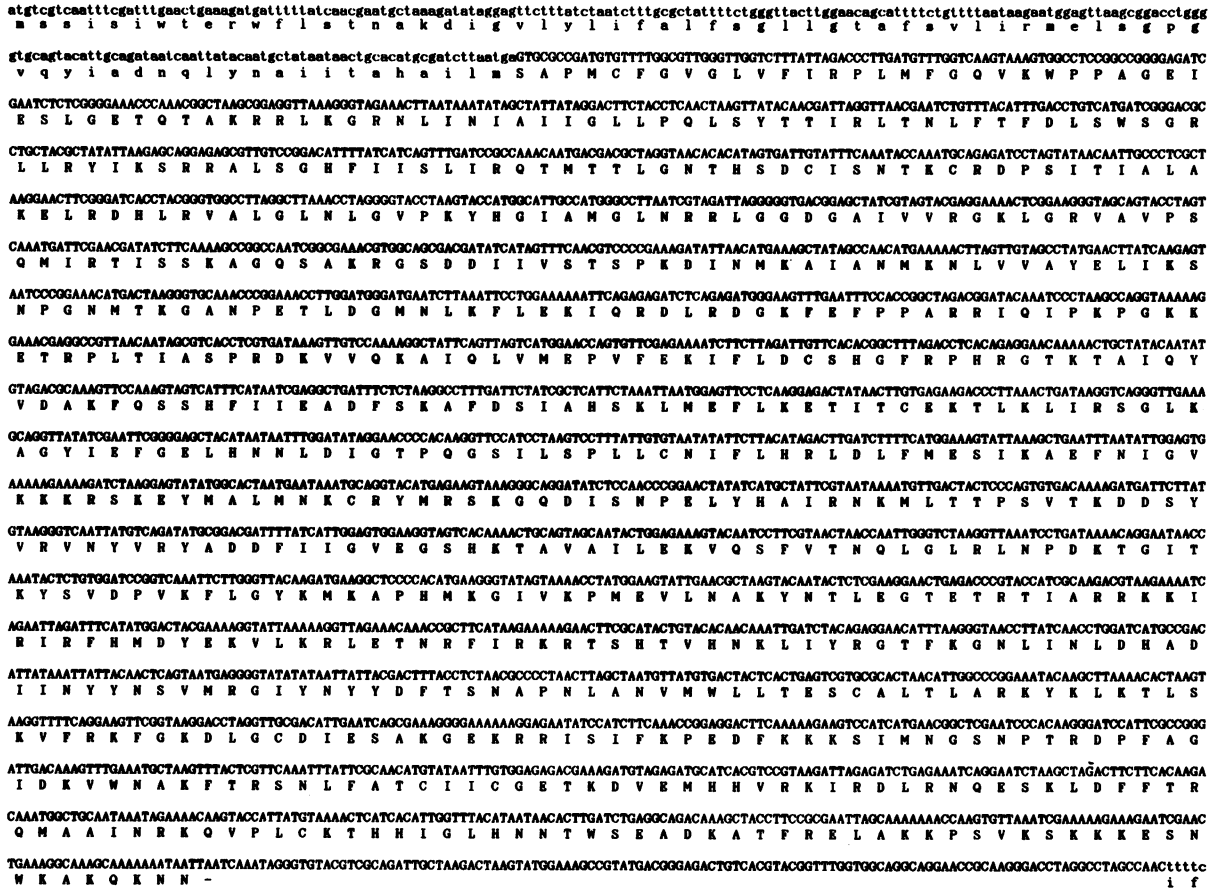


Figure 2. Gel electrophoresis and Northern hybridization demonstrating high concentration of intron 1 RNA. Left: ethidium bromide stained agarose gel; right: autoradiogram of a Northern transfer of a duplicate gel hybridized with intron-specific probe pNAA8 (see Figure 1). Positions of the large (3.3 kb) and small (2.0 kb) mt rRNAs are indicated on the left. The two smaller RNAs visible by ethidium bromide staining in the Varkud-1c strain are related to a mt plasmid (Saville and Collins, manuscript in preparation).


1 atgtcgtcaatttcgattgaactgaaagatgatttttatcaacgaatgctaaagatataggagttctttatctaatctttgcgctatttttctgggttacttggaacagcattttctgttttaataaggatggagttaagcggacctggg
m s s i s i w t e r w f l s t n a k d i g v l y l i f a l f s g l l g t a f s v l i r m e l s g p g
151 gtgcagtacattgcagataaatcaattatacaatgctataaataactgcacatgcgatcttaatgaGTGCGCCGATGTGTTTTGGCGTTGGGTTGGTCTTTATTAGACCCTTGATGTTTGGTCAAGTAAAGTGGCCTCCGGCCGGGGAGATC
v q y i a d n q l y n a i i t a h a i l m S A P M C F G V G L V F I R P L M F G Q V K W P P A G E I
301 GAATCTCTCGGGGAAACCCAAACGGCTAAGCGGAGGTTAAAGGGTAGAAACTTAATAAATATAGCTATTATAGGACTTCTACCTCAACTAAGTTATACAACGATTAGGTTAACGAATCTGTTTACATTTGACCTGTCATGATCGGGACGC
E S L G E T Q T A K R R L K G R N L I N I A I I G L L P Q L S Y T T I R L T N L F T F D L S W S G R
451 CTGCTACGCTATATTAAGAGCAGGAGAGCGTTGTCCGGACATTTTATCATCAGTTTGATCCGCCAAACAATGACGACGCTAGGTAACACACATAGTGATTGTATTTCAAATACCAAATGCAGAGATCCTAGTATAACAATTGCCCTCGCT
L L R Y I K S R R A L S G H F I I S L I R Q T M T T L G N T H S D C I S N T K C R D P S I T I A L A
601 AAGGAACTTCGGGATCACCTACGGGTGGCCTTAGGCTTAAACCTAGGGGTACCTAAGTACCATGGCATTGCCATGGGCCTTAATCGTAGATTAGGGGGTGACGGAGCTATCGTAGTACGAGGAAAACTCGGAAGGGTAGCAGTACCTAGT
K E L R D H L R V A L G L N L G V P K Y H G I A M G L N R R L G G D G A I V V R G K L G R V A V P S
751 CAAATGATTCGAACGATATCTTCAAAAGCCGGCCAATCGGCGAAACGTGGCAGCGACGATATCATAGTTTCAACGTCCCCGAAAGATATTAACATGAAAGCTATAGCCAACATGAAAAACTTAGTTGTAGCCTATGAACTTATCAAGAGT
Q M I R T I S S K A G Q S A K R G S D D I I V S T S P K D I N M K A I A N M K N L V V A Y E L I K S
901 AATCCCGGAAACATGACTAAGGGTGCAAACCCGGAAACCTTGGATGGGATGAATCTTAAATTCCTGGAAAAAATTCAGAGAGATCTCAGAGATGGGAAGTTTGAATTTCCACCGGCTAGACGGATACAAATCCCTAAGCCAGGTAAAAAG
N P G N M T K G A N P E T L D G M N L K F L E K I Q R D L R D G K F E F P P A R R I Q I P K P G K K
1151 GAAACGAGGCCGTTAACAATAGCGTCACCTCGTGATAAAGTTGTCCAAAAGGCTATTCAGTTAGTCATGGAACCAGTGTTCGAGAAAATCTTCTTAGATTGTTCACACGGCTTTAGACCTCACAGAGGAACAAAAACTGCTATACAATAT
E T R P L T I A S P R D K V V Q K A I Q L V M E P V F E K I F L D C S H G F R P H R G T K T A I Q Y
1201 GTAGACGCAAAGTTCCAAAGTAGTCATTTCATAATCGAGGCTGATTTCTCTAAGGCCTTTGATTCTATCGCTCATTCTAAATTAATGGAGTTCCTCAAGGAGACTATAACTTGTGAGAAGACCCTTAAACTGATAAGGTCAGGGTTGAAA
V D A K F Q S S H F I I E A D F S K A F D S I A H S K L M E F L K E T I T C E K T L K L I R S G L K
1351 GCAGGTTATATCGAATTCGGGGAGCTACATAATAATTTGGATATAGGAACCCCACAAGGTTCCATCCTAAGTCCTTTATTGTGTAATATATTCTTACATAGACTTGATCTTTTCATGGAAAGTATTAAAGCTGAATTTAATATTGGAGTG
A G Y I E F G E L H N N L D I G T P Q G S I L S P L L C N I F L H R L D L F M E S I K A E F N I G V
1501 AAAAAGAAAAGATCTAAGGAGTATATGGCACTAATGAATAAATGCAGGTACATGAGAAGTAAAGGGCAGGATATCTCCAACCCGGAACTATATCATGCTATTCGTAATAAAATGTTGACTACTCCCAGTGTGACAAAAGATGATTCTTAT
K K K R S K E Y M A L M N K C R Y M R S K G Q D I S N P E L Y H A I R N K M L T T P S V T K D D S Y
1651 GTAAGGGTCAATTATGTCAGATATGCGGACGATTTTATCATTGGAGTGGAAGGTAGTCACAAAACTGCAGTAGCAATACTGGAGAAAGTACAATCCTTCGTAACTAACCAATTGGGTCTAAGGTTAAATCCTGATAAAACAGGAATAACC
V R V N Y V R Y A D D F I I G V E G S H K T A V A I L E K V Q S F V T N Q L G L R L N P D K T G I T
1801 AAATACTCTGTGGATCCGGTCAAATTCTTGGGTTACAAGATGAAGGCTCCCCACATGAAGGGTATAGTAAAACCTATGGAAGTATTGAACGCTAAGTACAATACTCTCGAAGGAACTGAGACCCGTACCATCGCAAGACGTAAGAAAATC
K Y S V D P V K F L G Y K M K A P H M K G I V K P M E V L N A K Y N T L E G T E T R T I A R R K K I
1951 AGAATTAGATTTCATATGGACTACGAAAAGGTATTAAAAAGGTTAGAAACAAACCGCTTCATAAGAAAAAGAACTTCGCATACTGTACACAACAAATTGATCTACAGAGGAACATTTAAGGGTAACCTTATCAACCTGGATCATGCCGAC
R I R F H M D Y E K V L K R L E T N R F I R K R T S H T V H N K L I Y R G T F K G N L I N L D H A D
2101 ATTATAAATTATTACAACTCAGTAATGAGGGGTATATATAATTATTACGACTTTACCTCTAACGCCCCTAACTTAGCTAATGTTATGTGACTACTCACTGAGTCGTGCGCACTAACATTGGCCCGGAAATACAAGCTTAAAACACTAAGT
I I N Y Y N S V M R G I Y N Y Y D F T S N A P N L A N V M W L L T E S C A L T L A R K Y K L K T L S
2251 AAGGTTTTCAGGAAGTTCGGTAAGGACCTAGGTTGCGACATTGAATCAGCGAAAGGGGAAAAAAGGAGAATATCCATCTTCAAACCGGAGGACTTCAAAAAGAAGTCCATCATGAACGGCTCGAATCCCACAAGGGATCCATTCGCCGGG
K V F R K F G K D L G C D I E S A K G E K R R I S I F K P E D F K K K S I M N G S N P T R D P F A G
2401 ATTGACAAAGTTTGAAATGCTAAGTTTACTCGTTCAAATTTATTCGCAACATGTATAATTTGTGGAGAGACGAAAGATGTAGAGATGCATCACGTCCGTAAGATTAGAGATCTGAGAAATCAGGAATCTAAGCTAGACTTCTTCACAAGA
I D K V W N A K F T R S N L F A T C I I C G E T K D V E M H H V R K I R D L R N Q E S K L D F F T R
2551 CAAATGGCTGCAATAAATAGAAAACAAGTACCATTATGTAAAACTCATCACATTGGTTTACATAATAACACTTGATCTGAGGCAGACAAAGCTACCTTCCGCGAATTAGCAAAAAAACCAAGTGTTAAATCGAAAAAGAAAGAATCGAAC
Q M A A I N R K Q V P L C K T H H I G L H N N T W S E A D K A T F R E L A K K P S V K S K K K E S N
2701 TGAAAGGCAAAGCAAAAAAATAATTAATCAAATAGGGTGTACGTCGCAGATTGCTAAGACTAAGTATGGAAAGCCGTATGACGGGAGACTGTCACGTACGGTTTGGTGGCAGGCAGGAACCGCAAGGGACCTAGGCCTAGCCAACTtttc
W K A K Q K N N -
i f

only Group II intron identified to date in *Neurospora*. All of the characteristic secondary structural features described for other Group II introns by Michel and co-workers (33–36) can be identified. In the standard drawing, six helices (I through VI) radiate from a 'central wheel', the nucleotides forming the wheel are quite conserved among Group II introns. The structures and sequences forming helices V and VI are highly conserved. Additional stems and loops subtended by helix I in the Adiopodoume intron (Figure 4A) are also very similar to those in the *S. cerevisiae* Group II introns (33,34). Sequences capable of forming tertiary interactions between loops within the intron (α and α'; Figure 4B) and between intron loops and the upstream exon (exon/intron binding sites; 35) are also found (Figure 4B). Some Group II introns have been shown to be capable of self-splicing *in vitro* in the absence of proteins, while others have not (reviewed in 36). Attempts to detect splicing *in vitro* of *Neurospora coI* intron 1 have thusfar been unsuccessful (Saville and Collins, unpublished).

Previous studies (4) showed that splicing of *coI* intron 1 occurs efficiently in the nuclear mutant *cyt18-1*, which is deficient in splicing several Group I introns. The finding that this intron is a Group II intron suggests that the difference between Group I and Group II intron structure may be one of the reasons for the *cyt18*-independent splicing. The cyt18 protein, or at least the region of the protein affected by the *cyt18-1* mutation, may not be involved in Group II intron splicing. We have recently found that splicing of some Group I introns is also not affected by the *cyt18-1* mutation (Mota and Collins, unpublished). These observations suggest a biological significance for the differences among intron structures: splicing of some introns could be regulated by the availability or activity of the *cyt18* gene product without affecting splicing of other introns.

Molecular evolution of Group II intron structure and location.

DNA sequence analysis of the intron/upstream exon junction of the Varkud-1c strain revealed that the intron is located in the same position as in the Adiopodoume strain. The sequenced region of the Varkud-1c gene (all of exon 1 and the first 410 nt. of the intron) contains nine base substitutions compared to the Adiopodoume gene (Figure 4A). Three of these occur in the last 11 bases of the upstream exon, a region thought to be involved in base-pairing with intron sequences (35). Two of these three substitutions (at positions 211 and 209 (numbered as in Figure 3 where the last base of the exon is 214)) would yield a more stable intron/exon pairing, while that at 205 would disrupt base pairing (Figure 4B). Compensating substitutions have not occurred in the complementary intron sequences (EBS1 and EBS2). The lack of compensating substitutions is not totally unexpected, since intron/exon interactions in other Group II introns contain some unpaired bases (36,37). The rest of the stem I region contains six base substitutions. Three of these occur in regions thought to be involved in secondary structure: only the substitution at position 308 (stem C1) increases base-pairing; those at 352 (stem C2) and 375 (stem D) would de-stabilize helices (Figure 4A).

The intron is 2631 nucleotides long and interrupts the coding sequence after base 214. Splice sites can be assigned unambiguously by comparison with the sequence of the

Figure 3. DNA sequence and deduced amino acid sequence of *coI* intron I and flanking regions in the Adiopodoume strain. The sequence is shown from the position equivalent to the beginning of the *coI* gene in lab strain 74A (23,24). The sequence has been submitted to the EMBL database (accession number X14669). Exon sequences are in lower case, intron in upper case.

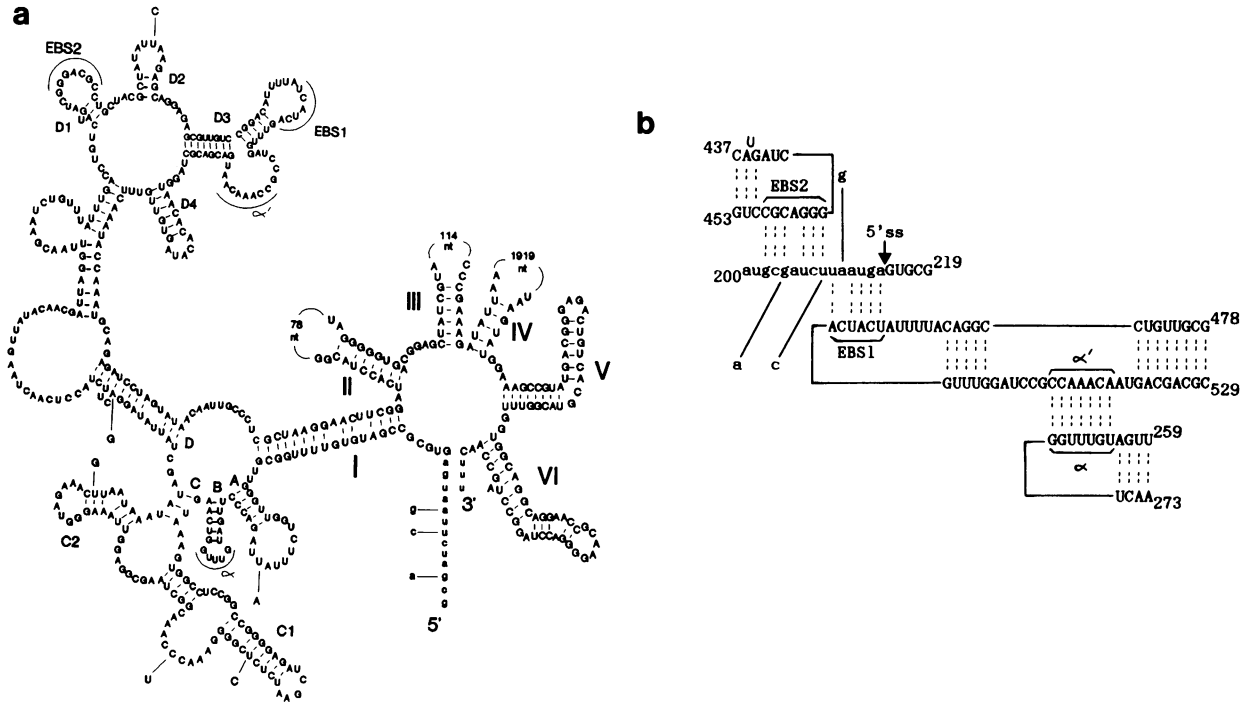


Figure 4. a) Potential secondary structure of the core of *col* intron 1. Short regions of the flanking exons (lowercase letters) are included. b) Potential tertiary interactions between exon binding site (EBS) sequences in loops D1 and D3 with complementary regions in the upstream exon (lower case). The arrow labelled 5′ss marks the 5′ splice site. Conserved structures are drawn and labeled according to the model of Michel and co-workers (33−36). Bases are numbered as in Figure 3, beginning at the start of exon 1; base 215 is the first base of the intron. Single base substitutions observed in the Varkud-1c strain are indicated by connecting them to the corresponding base in the secondary and tertiary structure drawings.

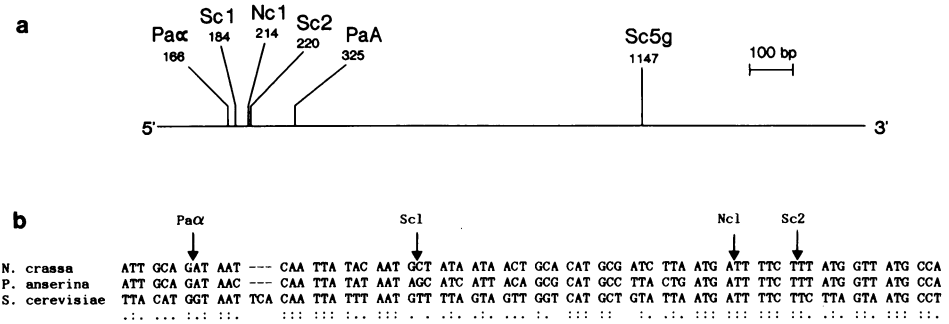


Figure 5. a) Locations of Group II introns in the *coI* genes of fungi. Exon sequences from all genes were aligned with the intronless *coI* gene (1671 bp) from *Neurospora* lab strain 74A. Numbers refer to the nucleotide position in the 74A gene corresponding to positions after which introns in the other genes are located. Sources of the exon sequences are: *N. crassa* (Nc) (23,24); *S. cerevisiae* (Sc) (26); *P. anserina* (Pa) (27,28); b) conservation of DNA sequence in the *coI* exon region surrounding the locations of Group II introns. Nucleotides present in all three sequences are indicated by (:); those in two of the sequences by (.).

intronless *coI* gene of the laboratory strain 74A (23,24). Figure 5 provides a comparison of the locations of Group II introns in the *coI* genes of other fungi. Of the five known large introns (two each in *Saccharomyces* and *Podospora*, one in *Neurospora*) no two are found at the same position, although they are clustered in a small region near the 5′ end of the gene. A small Group II intron lacking a long open reading frame is also found in *S. cerevisiae* near the other end of the *coI* coding sequence.

It seems unlikely that the ancestral *coI* gene contained a large number (at least the present six) of closely-spaced Group II introns, different ones of which have been retained in each current species. In such a scenario, some ancestral *coI* exons would have been extremely short: only a six base exon would separate the the location of the *Neurospora* intron from that of intron 2 in the yeast gene. It also seems unlikely that all of the introns are descendants of a single ancestral intron whose splice sites have changed ('intron sliding'). The *coI* exon sequences in the region where most of the Group II introns are located are very similar among all of these organisms (Figure 5B) and do not contain sequences resembling the conserved first few bases of fungal Group II introns, as might be expected if the apparent position of an intron had changed due to a change in splice sites.

The occurrence of different Group II introns at different positions in different organisms is most consistent with the independent insertion of these introns into each species. However, comparison of the exon sequences surrounding the Group II introns in the *coI* genes of *Neurospora*, *Podospora*, and *Saccharomyces* does not reveal any extensive base sequence similarities that would suggest a similarity in the target sites for insertion of introns into the gene. It may be that each intron recognizes a different target sequence or that the insertion event or subsequent changes have altered the original target site such that it is no longer recognizable. Indeed, many of the sequence differences between the *N. crassa* and *N. intermedia* gene are concentrated near the end of the upstream exon. Exon sequence polymorphisms immediately upstream of Group I introns (possibly related to intron insertion) have also been noticed in several fungal mt genes since they were first pointed out by Hensgens et al. (38).

Figure 6. Alignment of the reading frames of the *Neurospora* Adiopodoume *coI* intron 1 with Group II introns in *coI* genes of two other fungi, and with the *Neurospora* Mauriceville plasmid. Amino acids are numbered at the left, beginning with the first amino acid of the intron ORF or with the presumed initiation codon of the Mauriceville plasmid. Asterisks at the 3′ end indicate the stop codons. Conserved regions I through VII identified in reverse transcriptase genes of mobile elements (40) are underlined. Larger regions shared among Group II introns are boxed and labeled *a* through *g*. Similar alignments including other Group II introns have been presented previously (39). Sources of the sequences are: *S. cerevisiae* i1 (26); *P. anserina* i1 (also called intron α; 27,28); Mauriceville plasmid (42).

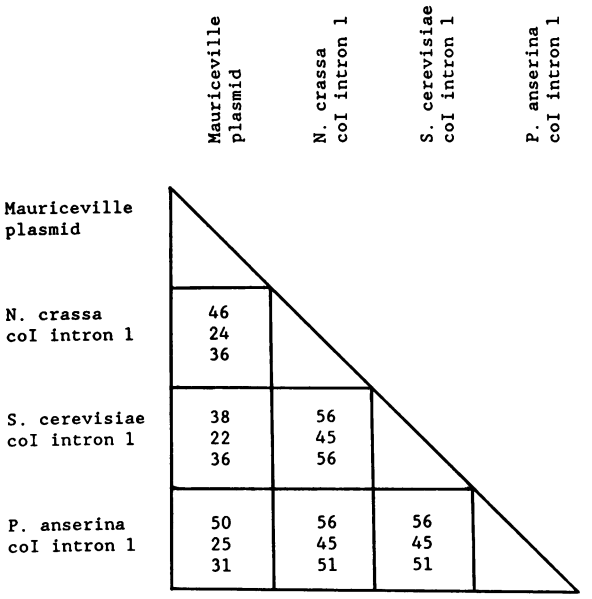


Figure 7. Comparisons of the similar regions of amino acid sequences in the Mauriceville plasmid and Group II intron reading frames. Each box contains three numbers, representing the percentage of positions containing identical amino acids in each of three pairwise comparisons. *Top*: the 'reverse transcriptase' homology regions described by Michel and Lang (ref. 40; 80 positions; Roman numerals I to VII in Figure 6); *middle*: the extended regions of similarity described by Xiong and Eickbush (ref 41; 173 positions); *bottom* regions shared among Group II ORFs (160 positions; blocks a to g in Figure 6; similar to the regions described in ref. 39). Since the sequences cannot be aligned unambiguously outside of these regions, percentage identity values over the entire ORFs are not presented.

The Intron Contains a Long Open Reading Frame Resembling that of a Mitochondrial Plasmid.

The first 2507 nt of the intron comprise an ORF (open reading frame) in frame with the upstream exon. This ORF could encode a large, hydrophilic, basic protein with predicted physical properties very similar to the putative products of Group II ORFs in other fungi. It has been noted previously that other Group II ORFs and the ORF in the mitochondrial plasmid of the Mauriceville strain of *Neurospora* all share regions of high percentage amino acid similarity separated by blocks with virtually no similarity (39); within the conserved blocks are short sequences that are also conserved in reverse transcriptases (40). All of these features are also apparent in the Adiopodoume intron (Figure 6).

Even though the Adiopodoume Group II intron ORF and the Mauriceville plasmid ORF are both *Neurospora* mitochondrial genes, they are not closely related. Thus, the plasmid DNA is not an excised form of a chromosomal Group II intron, as has been found in *Podospora* where a plasmid (called plDNA (27) or α-senDNA (28)) that is identical in sequence to the excised α intron (a Group II intron in the *coI* gene) accumulates during senescence. Pairwise comparisons of amino acid sequences were made among ORFs in Group II introns in the *Neurospora*, *Saccharomyces*, and *Podospora coI* genes and the ORF in the Mauriceville plasmid. The ORFs can not be aligned with certainty over their entire lengths, so these comparisons include only subsets of the sequences. Three separate

comparisons were performed for each pairwise combination using subsets of the sequences: 1) the short regions noticed by Michel and Lang (40) that are shared with reverse transcriptases (seven regions, numbered I to VII in Figure 6; a total of 80 positions); 2) the extended regions described by Xiong and Eickbush that are shared with other non-LTR retrotransposons (see ref. 41; 173 positions); 3) a slightly different set of extended regions shared among Group II intron ORFs (boxed regions a to g in Figure 6; 160 positions). The percentage of positions occupied by identical amino acids in the two sequences being compared is presented for each of the above comparisons, in the top, middle, and bottom row, in each box in Figure 7.

Comparing only the short 'reverse transcriptase' homology regions, the Group II ORFs are quite similar (38 to 50%) to the Mauriceville plasmid ORF. However, comparing the larger regions the Group II ORFs are much less similar to Mauriceville (22 to 36%) than to each other (45 to 56%). Several large insertions and deletions must be introduced into the Mauriceville sequence to allow alignment and even then convincing similarity is confined to reverse transcriptase blocks.

Still, the Adiopodoume *coI* Group II ORF and the Mauriceville plasmid ORF do share several features that distinguish them from *Neurospora* mt chromosomal exons or Group I intron ORFs. Base composition of the Group II intron and plasmid ORFs are similar

Table 1.

		G+C content (%)			
		total	codon position		
	length		1st	2nd	3rd
exons	(bases)				
coI	1671	33	40	42	19
coII	750	31	43	32	19
coIII	807	32	43	39	15
cob	1155	29	36	36	17
ndh1	1113	31	39	34	19
ndh2	1749	26	32	33	16
ndh4L	267	24	34	28	14
ndh5	2145	28	34	34	17
mal	222	36	42	44	19
all exons	**9879**	**30**	**37**	**36**	**17**
introns					
cobi1	951	29	34	34	18
cobi2	939	24	31	27	13
ndh1	912	25	31	31	15
ndh4L	1116	29	35	34	18
ndh5i1	1275	33	40	36	22
ndh5i2	918	26	32	32	12
S5	1278	31	31	28	19
all introns	**7389**	**27**	**34**	**31**	**17**
Group II intron	**2508**	**40**	**42**	**38**	**41**
Mauriceville	**2130**	**42**	**48**	**37**	**41**

Table 2.

Codon Usage in GROUP II INTRON ORF
Number of identified codons= 836

TTT	F	16	1.9%	TCT	S	11	1.3%	TAT	Y	13	1.6%	TGT	C	8	1.0%	TTT	F	193	5.9%	TCT	S	93	2.8%	TAT	Y	121	3.7%	TGT	C	18	.5%
TTC	F	22	2.6%	TCC	S	7	.8%	TAC	Y	10	1.2%	TGC	C	4	.5%	TTC	F	79	2.4%	TCC	S	13	.4%	TAC	Y	37	1.1%	TGC	C	3	.1%
TTA	L	21	2.5%	TCA	S	10	1.2%	TAA	-	0	--	TGA	W	5	.6%	TTA	L	322	9.8%	TCA	S	64	1.9%	TAA	-	0	--	TGA	W	56	1.6%
TTG	L	14	1.7%	TCG	S	7	.8%	TAG	-	0	--	TGG	W	1	.1%	TTG	L	42	1.3%	TCG	S	7	.2%	TAG	-	0	--	TGG	W	3	.1%
CTT	L	10	1.2%	CCT	P	12	1.4%	CAT	H	14	1.7%	CGT	R	8	1.0%	CTT	L	51	1.5%	CCT	P	84	2.6%	CAT	H	54	1.6%	CGT	R	6	.2%
CTC	L	7	.8%	CCC	P	5	.6%	CAC	H	8	1.0%	CGC	R	5	.6%	CTC	L	5	.2%	CCC	P	5	.2%	CAC	H	15	.5%	CGC	R	0	.0%
CTA	L	16	1.9%	CCA	P	7	.8%	CAA	Q	16	1.9%	CGA	R	2	.2%	CTA	L	52	1.6%	CCA	P	27	.8%	CAA	Q	59	1.8%	CGA	R	2	.1%
CTG	L	8	1.0%	CCG	P	9	1.1%	CAG	Q	4	.5%	CGG	R	5	.6%	CTG	L	5	.2%	CCG	P	5	.2%	CAG	Q	8	.2%	CGG	R	0	.0%
ATT	I	22	2.6%	ACT	T	16	1.9%	AAT	N	26	3.1%	AGT	S	15	1.8%	ATT	I	139	4.2%	ACT	T	69	2.1%	AAT	N	114	3.5%	AGT	S	110	3.3%
ATC	I	21	2.5%	ACC	T	10	1.2%	AAC	N	21	2.5%	AGC	S	2	.2%	ATC	I	41	1.2%	ACC	T	10	.3%	AAC	N	24	.7%	AGC	S	15	.5%
ATA	I	24	2.9%	ACA	T	16	1.9%	AAA	K	50	6.0%	AGA	R	25	3.0%	ATA	I	178	5.4%	ACA	T	83	2.5%	AAA	K	74	2.2%	AGA	R	64	1.9%
ATG	M	25	3.0%	ACG	T	9	1.1%	AAG	K	37	4.4%	AGG	R	14	1.7%	ATG	M	92	2.8%	ACG	T	3	.1%	AAG	K	9	.3%	AGG	R	5	.2%
GTT	V	8	1.0%	GCT	A	18	2.2%	GAT	D	24	2.9%	GGT	G	17	2.0%	GTT	V	93	2.8%	GCT	A	101	3.1%	GAT	D	73	2.2%	GGT	G	115	3.5%
GTC	V	7	.8%	GCC	A	12	1.4%	GAC	D	14	1.7%	GGC	G	8	1.0%	GTC	V	10	.3%	GCC	A	18	.5%	GAC	D	12	.4%	GGC	G	6	.2%
GTA	V	19	2.3%	GCA	A	14	1.7%	GAA	E	23	2.8%	GGA	G	15	1.8%	GTA	V	94	2.9%	GCA	A	68	2.1%	GAA	E	72	2.2%	GGA	G	77	2.3%
GTG	V	6	.7%	GCG	A	6	.7%	GAG	E	15	1.8%	GGG	G	12	1.4%	GTG	V	26	.8%	GCG	A	18	.5%	GAG	E	22	.7%	GGG	G	29	.9%

Codon Usage In total EXONS
Number of codons= 3293

Codon Usage in MAURICEVILLE ORF
Number of identified codons= 710

TTT	F	8	1.1%	TCT	S	14	2.0%	TAT	Y	13	1.8%	TGT	C	4	.6%	TTT	F	114	4.6%	TCT	S	70	2.8%	TAT	Y	106	4.3%	TGT	C	20	.8%
TTC	F	29	4.1%	TCC	S	8	1.1%	TAC	Y	14	2.0%	TGC	C	3	.4%	TTC	F	19	.8%	TCC	S	18	.7%	TAC	Y	16	.6%	TGC	C	1	.0%
TTA	L	16	2.3%	TCA	S	6	.8%	TAA	-	0	--	TGA	W	8	1.1%	TTA	L	170	6.9%	TCA	S	40	1.6%	TAA	-	0	--	TGA	W	25	1.0%
TTG	L	8	1.1%	TCG	S	2	.3%	TAG	-	0	--	TGG	W	6	.8%	TTG	L	24	1.0%	TCG	S	14	.6%	TAG	-	0	--	TGG	W	7	.3%
CTT	L	21	3.0%	CCT	P	14	2.0%	CAT	H	7	1.0%	CGT	R	13	1.8%	CTT	L	37	1.5%	CCT	P	51	2.1%	CAT	H	35	1.4%	CGT	R	14	.6%
CTC	L	13	1.8%	CCC	P	9	1.3%	CAC	H	5	.7%	CGC	R	3	.4%	CTC	L	9	.4%	CCC	P	10	.4%	CAC	H	4	.2%	CGC	R	6	.2%
CTA	L	21	3.0%	CCA	P	13	1.8%	CAA	Q	19	2.7%	CGA	R	1	.1%	CTA	L	30	1.2%	CCA	P	19	.8%	CAA	Q	51	2.1%	CGA	R	9	.4%
CTG	L	7	1.0%	CCG	P	1	.1%	CAG	Q	4	.6%	CGG	R	0	.0%	CTG	L	7	.3%	CCG	P	1	.0%	CAG	Q	8	.3%	CGG	R	3	.1%
ATT	I	16	2.3%	ACT	T	11	1.5%	AAT	N	19	2.7%	AGT	S	11	1.5%	ATT	I	97	3.9%	ACT	T	59	2.4%	AAT	N	187	7.6%	AGT	S	58	2.4%
ATC	I	10	1.4%	ACC	T	4	.6%	AAC	N	16	2.3%	AGC	S	12	1.7%	ATC	I	10	.4%	ACC	T	10	.4%	AAC	N	34	1.4%	AGC	S	9	.4%
ATA	I	10	1.4%	ACA	T	4	.6%	AAA	K	35	4.9%	AGA	R	26	3.7%	ATA	I	113	4.6%	ACA	T	32	1.3%	AAA	K	202	8.2%	AGA	R	65	2.6%
ATG	M	14	2.0%	ACG	T	4	.6%	AAG	K	29	4.1%	AGG	R	4	.6%	ATG	M	35	1.4%	ACG	T	9	.4%	AAG	K	54	2.2%	AGG	R	16	.6%
GTT	V	11	1.5%	GCT	A	7	1.0%	GAT	D	22	3.1%	GGT	G	18	2.5%	GTT	V	56	2.3%	GCT	A	50	2.0%	GAT	D	86	3.5%	GGT	G	68	2.8%
GTC	V	9	1.3%	GCC	A	12	1.7%	GAC	D	14	2.0%	GGC	G	6	.8%	GTC	V	8	.3%	GCC	A	12	.5%	GAC	D	13	.5%	GGC	G	4	.2%
GTA	V	11	1.5%	GCA	A	9	1.3%	GAA	E	32	4.5%	GGA	G	12	1.7%	GTA	V	50	2.0%	GCA	A	17	.7%	GAA	E	83	3.4%	GGA	G	31	1.3%
GTG	V	7	1.0%	GCG	A	7	1.0%	GAG	E	14	2.0%	GGG	G	4	.6%	GTG	V	10	.4%	GCG	A	9	.4%	GAG	E	19	.8%	GGG	G	19	.8%

Codon Usage in total GROUP I INTRON ORFs
Number of identified codons= 2463

to each other, 40% and 42% G+C, respectively, noticeably higher than the approximately 30% G+C found in chromosomal ORFs (Table 1). Codon base composition is also very different: there is a marked bias against G or C in the third position of chromosomal codons, less than 20% end in G or C, while the G or C content of third positions in the Group II and plasmid ORFs is over 40%. This bias is particularly evident in certain codon families (boxed in Table 2): chromosomal codons for phenylalanine, tyrosine, histidine, isoleucine, asparagine, lysine and aspartic acid show strong preference for codons ending in A or T; this bias is eliminated or even reversed (in the case of phenylalanine) in the Group II and plasmid ORFs.

The distinctive features of codon usage and base composition shared by the Group II and plasmid ORFs suggest that both ORFs may have evolved in a similar ancestral genetic environment that was not *Neurospora* mitochondria.

ACKNOWLEDGEMENTS

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada and the University of Toronto. We thank Liz Pason, Collette LaRochelle and Catherine Reynolds for technical assistance.

*To whom correspondence should be addressed

+Present address: Department of Genetics, Hospital for Sick Children, 555 University Avenue, Toronto, Ontario M5G 1X8, Canada

REFERENCES

1. Cech,T.R. (1988) Gene 73, 259–271.
2. Burke,J.M. (1988) Gene 73, 273–294.
3. Mannella,C.A., Collins,R.A., Green,M.R. and Lambowitz,A.M. (1979) Proc. Nat. Acad. Sci. USA 76, 2635–2639.
4. Collins,R.A. and Lambowitz,A.M. (1985) J. Mol. Biol. 184, 413–428.
5. Helmer-Citterich,M., Morelli,G. and Macino,G. (1983) EMBO J. 2, 1235–1242.
6. Burke,J.M., Breitenberger,C., Heckman,J.E., Dujon,B. and RajBhandary,U.L. (1984) J. Biol. Chem. 259, 504–511.
7. Burke,J.M. and RajBhandary,U.L. (1982) Cell 31, 509–520.
8. Garriga,G. and Lambowitz,A.M. (1986) Cell 46, 669–680.
9. Garriga,G. and Lambowitz,A.M. (1984) Cell 39, 631–641.
10. Wollenzein,P., Cantor,C.R., Grant,D.M. and Lambowitz,A.M. (1983) Cell 32, 397–407.
11. Collins,R.A. and Lambowitz,A.M. (1983) Plasmid 9, 53–70.
12. Davis,R.H. and deSerres,F.J. (1970) Meth. Enzymol. 17A, 79–143.
13. Lambowitz,A.M. (1979) Meth. Enzymol. 59, 421–433.
14. Garriga,G., Bertrand,H., and Lambowitz,A.M. (1984) Cell 36, 623–634.
15. Yanisch-Perron,C., Vieira,J. and Messing,J. (1985) Gene 33, 103–119.
16. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) Proc. Nat. Acad. Sci. USA 74, 5463–5467.
17. Korneluk,R., Quan,F. and Gravel,R.A. (1985) Gene 40, 317–323.
18. Henikoff,S. (1984) Gene 28, 351–359.
19. Dale,R.M.K., McClure,B.A. and Houchins,J.P. (1985) Plasmid 13, 31–40.
20. Mizusawa,S., Nishimura,S. and Seela,F. (1986) Nucl. Acids Res. 14, 1319–1324.
21. McMaster,G.K. and Carmichael,G.G. (1977) Proc. Nat. Acad. Sci. USA 74, 4835–4838.
22. Feinberg,A.P. and Vogelstein,B. (1983) Anal. Biochem. 132, 6–13.
23. Burger,G., Scriven,C., Machleidt,W. and Werner,S. (1982) EMBO J. 1, 1385–1391.
24. de Jonge,J.C. and de Vries,H. (1983) Curr. Genet. 7, 21–28.
25. Bonitz,S., Coruzzi,G., Thalenfeld,B.E., Tzagoloff,A. and Macino,G. (1980) J. Biol. Chem. 255, 11927–11941.
26. Arnberg,A.C., van Ommen,G.-J.B., Grivell,L.A., van Bruggen,E.F.J. and Borst,P. (1980) Cell 19, 313–319.

27. Oziewacz,H.D. and Esser,K. (1984) Curr. Genet 8, 299–305.
28. Cummings,D.J., MacNeil,I.A., Domenico,J. and Matsuura,E.T. (1985) J. Mol. Biol. 185, 659–680.
29. Peebles,C.L., Perlman,P.S., Mecklenburg,K.L., Petrillo,J.L., Tabor,J.H., Jarrell,K.A., and Cheng,H-L. (1986) Cell 44, 213–223.
30. van der Veen,R., Arnberg,A.C., van der Horst,G., Bonen,L., Tabak,H.F. and Grivell,L.A. (1986) Cell 44, 225–234.
31. Schmelzer,C. and Schweyen,R.J. (1986) Cell 46, 557–565.
32. Schmidt,U., Kosack,M. and Stahl,U. (1987) Curr. Genet. 12, 291–295.
33. Michel,F., Jacquier,A. and Dujon,B. (1982) Biochimie 64, 867–881.
34. Michel,F., and Dujon,B. (1983) EMBO J. 2, 33–38.
35. Jacquier,A. and Michel,F. (1987) Cell 50, 17–29.
36. Michel,F. (1989) Gene, in press.
37. Wesler,F. and Wolf,K. (1988) Nucl. Acids Res. 16, 1629.
38. Hensgens,L.A.M., Bonen,L., de Haan,M., van der Horst,G. and Grivell,L.A. (1983) Cell 32, 379–389.
39. Lang,B.F., Anhe,F. and Bonen,L. (1985) J. Mol. Biol. 184, 353–366.
40. Michel,F. and Lang,B.F. (1985) Nature 316, 641–643.
41. Xiong,Y. and Eickbush,T. (1988) Mol. Biol. Evol. 5, 675–690.
42. Nargang,F.E., Bell,J.B., Stohl,L.L. and Lambowitz,A.M. (1984) Cell 38,

This article, submitted on disc, has been automatically converted into this typeset format by the publisher.