Published in final edited form as:

Science. 2010 November 5; 330(6005): 838–841. doi:10.1126/science.1194554.

Evolution of Yeast Noncoding RNAs Reveals an Alternative Mechanism for Widespread Intron Loss

Quinn M. Mitrovich^{1,2,*}, Brian B. Tuch^{1,3,*†}, Francisco M. De La Vega³, Christine Guthrie^{2,‡}, and Alexander D. Johnson^{1,2,‡}

¹Department of Microbiology and Immunology, University of California, San Francisco, San Francisco, CA 94143-2200, USA

²Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA 94143-2200, USA

³Genetic Systems Division, Research and Development, Life Technologies, Foster City, CA 94404, USA

Abstract

The evolutionary forces responsible for intron loss are unresolved. Whereas research has focused on protein-coding genes, here we analyze noncoding small nucleolar RNA (snoRNA) genes in which introns, rather than exons, are typically the functional elements. Within the yeast lineage exemplified by the human pathogen *Candida albicans*, we find—through deep RNA sequencing and genome-wide annotation of splice junctions—extreme compaction and loss of associated exons, but retention of snoRNAs within introns. In the *Saccharomyces* yeast lineage, however, we find it is the introns that have been lost through widespread degeneration of splicing signals. This intron loss, perhaps facilitated by innovations in snoRNA processing, is distinct from that observed in protein-coding genes with respect to both mechanism and evolutionary timing.

In eukaryotes, protein-coding genes are frequently interrupted by introns, which must be precisely removed from RNA transcripts by the nuclear spliceosome (1). Over evolutionary time scales, the presence of introns is dynamic, with intron gain and loss rates varying substantially across eukaryotic lineages (2-4). The mechanisms of intron gain and loss speak to questions about both the origins of introns and the markedly different intron-exon patterns observed across eukaryotes (5)—for example, whether spliceosomal introns arose within eukaryotes (“introns late”), within an ancestor of both prokaryotes and eukaryotes (“introns early”), or even before the emergence of protein-coding genes (“introns first”) (6). The last two hypotheses depend on the feasibility of comprehensive intron loss within both the prokaryotic and archaeal lineages, whose modern representatives lack spliceosomal introns. Within eukaryotes, the hemiascomycetous yeasts show substantial intron loss, with modern species like *Saccharomyces cerevisiae* and *Candida albicans* devoid of introns in >90% of their genes (7). A postulated mechanism for this loss is reverse transcription of spliced RNA, followed by homologous DNA recombination that replaces the intron-containing genomic sequence with the intronless copy (8). Previous studies of intron loss have focused on protein-coding genes, however, and are therefore likely to be biased toward mechanisms

[‡]To whom correspondence should be addressed. christineguthrie@gmail.com (C.G.); ajohnson@cgl.ucsf.edu (A.D.J.).

*These authors contributed equally to this work.

[†]Present address: Genome Analysis Unit, Amgen, South San Francisco, CA 94080, USA.

Supporting Online Material

www.sciencemag.org/cgi/content/full/330/6005/838/DC1

that lead to precise intron removal. Here, we examine instead the evolution of splicing patterns in yeast noncoding genes.

We performed massively parallel ligation-based sequencing of RNA libraries (RNA-seq) and mapped the resulting ~160 million 50-nucleotide (nt) strand-specific sequence reads to the *C. albicans* genome (9). Our data confirm 89% of 421 previously annotated spliceosomal introns (7, 10) (table S1), while correcting or rejecting seven of these annotations (table S2). We also find 68 previously unannotated splice junctions, identifying 15 new introns in protein-coding genes (table S3), 30 examples of alternative splicing [table S4 and supporting online (SOM) text], and 23 new introns in previously unannotated transcripts that lack substantial protein-coding potential (table S5). Analysis of 11 of these spliced, noncoding RNAs revealed that their exons have no apparent function, but that their introns contain C/D box snoRNAs—noncoding RNAs that target modifications to ribosomal RNA (rRNA) (11). In the nonhemiascomycetous fungus *Neurospora crassa*, snoRNAs are also generally processed from the introns of non-protein-coding precursors (12). This is different, however, from the more closely related hemiascomycete *S. cerevisiae*, where nearly all snoRNAs arise from unspliced primary transcripts and, therefore, require a splicing-independent processing pathway (13). This difference between *C. albicans* and *S. cerevisiae* suggests that the transition of snoRNAs from intron sequences to unspliced, dedicated transcripts occurred within the *Saccharomyces* lineage, well after the onset of rapid intron loss from protein-coding genes in the hemiascomycete ancestor.

To trace the evolution of snoRNAs throughout the hemiascomycetes, we first generated 40 high-confidence C/D box snoRNA predictions for *C. albicans* (e.g., Fig. 1). Among these are the 11 identified in our intron analysis (above), as well as the previously identified *snR52* (14). We searched for splicing signals in sequences adjacent to snoRNAs, and predict that 33 of the 40 identified *C. albicans* snoRNAs are intronic. This confirms the difference between *C. albicans* and *S. cerevisiae*, as C/D box snoRNAs in the latter species are rarely intronic [6 out of 47 (13)].

We next identified orthologous snoRNAs from other sequenced yeasts (Fig. 2), beginning with computational predictions (or, for *S. cerevisiae*, existing annotations), identifying candidate orthologs of our *C. albicans* set based on their predicted rRNA target sites, and finally confirming and refining our predictions by searching for limited primary sequence identity among predicted snoRNAs (15). This final refinement identified snoRNAs whose rRNA target sites had changed between species, demonstrating evolutionary plasticity in yeast rRNA methylation sites (fig. S1). We predict that 105 of the 255 analyzed snoRNAs are located within introns (Fig. 2A). Individual species vary considerably: Those in the *Saccharomyces* complex have few intronic snoRNAs (three to five), whereas others have substantially more (23 to 33). The most parsimonious explanation for these data is massive loss of snoRNA-associated introns, most of which took place in the common ancestor of the *Saccharomyces* complex (Fig. 2B).

The intron loss mechanism proposed for protein-coding genes—retrotransposition of spliced mRNAs—cannot explain the pattern we observe, as it would eliminate the snoRNAs along with the introns. Rather, the introns appear to have been lost through degeneration of their splicing signals, effectively converting them into unspliced exon sequence. In *N. crassa* (12) and hemiascomycetes outside the *Saccharomyces* complex, snoRNAs in the *snR72-78* polycistronic cluster are mostly contained within individual introns (Fig. 3A). In *S. cerevisiae*, the genes are arranged identically, but are cotranscribed as a single unspliced precursor, then processed into individual snoRNAs by the RNase III enzyme Rnt1 (16). The conservation of genomic synteny among these species strongly suggests intron loss through splice site degeneration (“de-intronization”) with *Yarrowia lipolytica* and the *Candida* clade

representing an intermediate state of partial intron loss and *S. cerevisiae* representing complete intron loss. The *snR57* cluster similarly supports this idea (below), as do the gene structures of *snR47* and *snR79* (SOM text).

Intron loss through splice-site degeneration would not be expected to occur within protein-coding genes, as this would disrupt the encoded protein. Consistent with this prediction, of the five snoRNAs located in introns within protein-coding regions in *C. albicans*, four remain associated with introns in nearly all the *Saccharomyces* complex species [*snR18*, *snR24*, *snR54*, and *CD39*, (Fig. 2A)]. The fifth, *snR39b*, was likely displaced from its associated coding sequence through a genomic duplication event (fig. S2B). The snoRNA *CD39* is located within the ribosomal protein gene *MRPS35* intron in all but one of the species we analyzed (Fig. 2A). The exception is *S. cerevisiae*, where *MRPS35* has lost its intron, presumably through retrotransposition rather than deintronization. This appears to have eliminated *CD39* from the genome as well: The predicted rRNA methylation site for this snoRNA is unmodified in *S. cerevisiae* (13).

In *C. albicans*, one unusual case of splicing may reflect the particular processing requirements imposed by intron-hosted snoRNAs. The snoRNAs *snR57*, *snR55*, and *snR61* are processed from three introns of a single precursor (Fig. 3B). We find that two of the introns (introns 2 and 3) lie entirely within the sequence of an enveloping intron (intron 1). The splicing signals for intron 1 are not present in the primary transcript, but are created upon splicing of introns 2 and 3. Thus, splicing of intron 1 can occur only after introns 2 and 3 have been removed. (See SOM text for similar scenarios in protein-coding genes of other eukaryotes.) Analysis of both the *snR57/55/61* and the *snR72-78* clusters in other species indicates a significant reduction in the sizes of internal exons within the hemiascomycetes, driven perhaps by the same pressures that streamlined their genomes (17) and leading ultimately to nested splicing within the *Candida* clade (fig. S3). In animals and fungi, intron-hosted C/D box snoRNAs obey a strict “one-snoRNA-per-intron” rule, a requirement imposed by their exonucleolytic maturation pathways (18). Nested splicing of snoRNA host transcripts fulfills this requirement by allowing sequential removal of individual introns despite the absence of intervening exons.

Selective pressures are proposed to have driven intron loss from hemiascomycete protein-coding genes (19), and these same pressures may have driven the loss we observe here. The dependence of snoRNAs on splicing for their proper maturation, however, would have imposed a constraint against loss of their associated introns (18). This constraint may have been overcome by the innovation of an alternative snoRNA processing pathway in the *Saccharomyces* lineage, where exonic C/D box snoRNAs first undergo endonucleolytic cleavage by the RNase III enzyme Rnt1 (20). None-theless, some capacity to process exonic snoRNAs must be more ancient, because other hemiascomycetes (Fig. 2) and more distantly related fungi (12) do have some exon-hosted snoRNAs. It is unknown, however, whether processing of such snoRNAs involves Rnt1 (SOM text).

Studies of protein-coding genes have revealed a dramatic increase in the rate of intron loss in the hemiascomycete ancestor (21). By focusing instead on noncoding RNAs, we describe unexpected patterns of both exon and intron loss. A drastic reduction in the sizes of internal exons has ultimately led to their complete loss in *Candida* species, resulting in a complex form of splicing that maintains the independent removal of now over-lapping introns. The *Saccharomyces* complex, however, has experienced a second wave of intron loss, perhaps facilitated by innovations in snoRNA processing and mediated by a mechanism—splice site degeneration—distinct from that which has acted on protein-coding genes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank C. Barbacioru, O. Homann, S. Kuersten, S. Ledoux, I. Listerman, T. Lowe, C. Maeder, K. Pande, A. Price, S. Roy, J. Steitz, and members of the Guthrie and Johnson laboratories for helpful discussions and support; and M. Barker and C. Monighetti for providing sequence reads. This work was supported by NIH grants GM021119 (C.G.) and GM37049 (A.D.J.) and by Life Technologies Corp. C.G. is an American Cancer Society Research Professor of Molecular Genetics. Sequence data are available at the Gene Expression Omnibus (accession number GSE21291).

References

1. Wahl MC, Will CL, Lührmann R. Cell. 2009; 136:701. [PubMed: 19239890]
2. Li W, Tucker AE, Sung W, Thomas WK, Lynch M. Science. 2009; 326:1260. [PubMed: 19965475]
3. Carmel L, Wolf YI, Rogozin IB, Koonin EV. Genome Res. 2007; 17:1034. [PubMed: 17495008]
4. Roy SW, Gilbert W. Proc. Natl. Acad. Sci. U.S.A. 2005; 102:5773. [PubMed: 15827119]
5. Roy SW, Gilbert W. Nat. Rev. Genet. 2006; 7:211. [PubMed: 16485020]
6. Penny D, Hoepfner MP, Poole AM, Jeffares DC. J. Mol. Evol. 2009; 69:527. [PubMed: 19777149]
7. Mitrovich QM, Tuch BB, Guthrie C, Johnson AD. Genome Res. 2007; 17:492. [PubMed: 17351132]
8. Fink GR. Cell. 1987; 49:5. [PubMed: 3549000]
9. Tuch BB, et al. PLoS Genet. 2010; 6:e1001070. [PubMed: 20808890]
10. Braun BR, et al. PLoS Genet. 2005; 1:36. [PubMed: 16103911]
11. Weinstein LB, Steitz JA. Curr. Opin. Cell Biol. 1999; 11:378. [PubMed: 10395551]
12. Liu N, et al. BMC Genomics. 2009; 10:515. [PubMed: 19895704]
13. Piekna-Przybylska D, Decatur WA, Fournier MJ. RNA. 2007; 13:305. [PubMed: 17283215]
14. Marck C, et al. Nucleic Acids Res. 2006; 34:1816. [PubMed: 16600899]
15. Materials and methods are available as supporting material on Science Online.
16. Qu LH, et al. Mol. Cell. Biol. 1999; 19:1144. [PubMed: 9891049]
17. Dujon B, et al. Nature. 2004; 430:35. [PubMed: 15229592]
18. Brown JW, Marshall DF, Echeverria M. Trends Plant Sci. 2008; 13:335. [PubMed: 18555733]
19. Lynch M. Proc. Natl. Acad. Sci. U.S.A. 2002; 99:6118. [PubMed: 11983904]
20. Chanfreau G, Legrain P, Jacquier A. J. Mol. Biol. 1998; 284:975. [PubMed: 9837720]
21. Stajich JE, Dietrich FS, Roy SW. Genome Biol. 2007; 8:R223. [PubMed: 17949488]
22. Lowe TM, Eddy SR. Science. 1999; 283:1168. [PubMed: 10024243]

Fig. 1.

Sequence library comparisons reveal noncoding RNAs. RNA sequence data are shown for *MRPS35*, a representative protein-coding gene that hosts a snoRNA within its intron. Nonadenylated RNAs, such as mature snoRNAs, are enriched in the rRNA-depleted library relative to the poly(A)-selected library. Sequence depth is represented on a $\log_2$ axis. (Bottom track) One of 1706 lower-confidence snoRNA predictions generated using the snoscan algorithm (22).

Fig. 2.

snoRNA-associated introns were lost in the *Saccharomyces* lineage. (A) Intron prediction scores [combined 5' splice site and branch site species-specific PWM scores (15)] for 40 C/D box snoRNA flanking sequences (± 200 nt) in seven representative hemiascomycetes. Each snoRNA (horizontal row) is labeled according to the nomenclature of *S. cerevisiae* (where applicable) or *N. crassa* (CD39) or by the predicted *C. albicans* rRNA modification site. Intron scores greater than 5.0 (false-positive rate $<0.4\%$) are shaded green. Inferred intron loss events for each snoRNA, based on parsimony, are indicated on the left and correspond to labeled branches (not drawn to scale) in the phylogeny shown above and in (B). snoRNAs that were not identified or that lack sufficient flanking sequence to score are

gray and labeled NA or ND, respectively. Locations of *N. crassa* orthologs (green for intronic, white for exonic) are derived from (12). Intron scores for reverse-complements of flanking sequences (right) are provided as a negative control. **(B)** Phylogenetic pattern of snoRNA-associated intron loss. The number of loss events assigned to each branch is indicated in yellow. Species: *N. crassa*, *Y. lipolytica*, *Debaryomyces hansenii*, *C. albicans*, *Kluyveromyces lactis*, *Kluyveromyces waltii*, *Zygosaccharomyces rouxii*, and *S. cerevisiae*.

\$watermark-text

\$watermark-text

\$watermark-text

Fig. 3. Polycistronic snoRNA precursors exhibit unusual splicing patterns. **(A)** Splicing of the *snR72-78* cluster in various fungal species (phylogenetic relations on left). snoRNAs are shown in green, introns as lines, and exons as yellow boxes, with internal exons labeled by size. Cotranscription of entire clusters has been demonstrated only for *S. cerevisiae* (16) and *N. crassa* (12). **(B)** Nested splicing of the *snR57/55/61* snoRNA cluster in *C. albicans*. The 5' splice site, branch site, and 3' splice site sequences are shown for introns 1 (pink), 2 (blue), and 3 (orange). Gray "exons" represent sequences joined with intron 1, by splicing of introns 2 and 3, to create de novo intron 1 splice sites. (Inset) Reverse transcription polymerase chain reaction products of the snoRNA host transcript from wild-type cells and those deficient for nonsense-mediated mRNA decay (*upf1*-Δ) or nuclear exonucleolytic decay (*trp6*-Δ). We infer the order of splicing events from the observable products: Intron 3 is nearly always removed, intron 2 is removed in a subset of transcripts (lower two bands), and intron 1 only when 2 and 3 have also been removed (lower band). Patterns in decay mutants are consistent with degradation of partially spliced host transcripts in the nucleus.