Research article

The exon context and distribution of Euascomycetes rRNA spliceosomal introns

Debashish Bhattacharya*[1], Dawn Simon[1], Jian Huang[2], Jamie J Cannone[3] and Robin R Gutell[3]

Address: [1]Department of Biological Sciences and Center for Comparative Genomics, University of Iowa, Iowa City IA USA, [2]Department of Statistics and Actuarial Science, University of Iowa, Iowa City IA USA and [3]Institute for Cellular and Molecular Biology, University of Texas, Austin TX USA

Email: Debashish Bhattacharya* - dbhattac@blue.weeg.uiowa.edu; Dawn Simon - dsimon@blue.weeg.uiowa.edu; Jian Huang - jian@stat.uiowa.edu; Jamie J Cannone - cannone@mail.utexas.edu; Robin R Gutell - robin.gutell@mail.utexas.edu

* Corresponding author

Abstract

Background: We have studied spliceosomal introns in the ribosomal (r)RNA of fungi to discover the forces that guide their insertion and fixation.

Results: Comparative analyses of flanking sequences at 49 different spliceosomal intron sites showed that the $G - intron - G$ motif is the conserved flanking sequence at sites of intron insertion. Information analysis showed that these rRNA introns contain significant information in the flanking exons. Analysis of all rDNA introns in the three phylogenetic domains and two organelles showed that group I introns are usually located after the most conserved sites in rRNA, whereas spliceosomal introns occur at less conserved positions. The distribution of spliceosomal and group I introns in the primary structure of small and large subunit rRNAs was tested with simulations using the broken-stick model as the null hypothesis. This analysis suggested that the spliceosomal and group I intron distributions were not produced by a random process. Sequence upstream of rRNA spliceosomal introns was significantly enriched in G nucleotides. We speculate that these G-rich regions may function as exonic splicing enhancers that guide the spliceosome and facilitate splicing.

Conclusions: Our results begin to define some of the rules that guide the distribution of rRNA spliceosomal introns and suggest that the exon context is of fundamental importance in intron fixation.

Background

Many eukaryotic genes are interrupted by stretches of non-coding DNA called introns or intervening sequences. Transcription of these genes is followed by RNA-splicing that results in intron removal (for review, see [1]). The majority of eukaryotic spliceosomal introns interrupt pre-mRNA in the nucleus and are removed by a ribonucleo-

protein complex, termed the spliceosome. Two theories have been proposed to explain the present spliceosomal intron distribution; i.e., their presence in eukaryotes and their absence in Bacteria and Archaea. The first, "introns-early", posits that introns were present in most, if not all, protein-coding genes in the last universal common ancestor (LUCA) and have subsequently been lost in the

archaeal and bacterial domains due to strong selection for compact genomes. Eukaryotes have maintained their introns because they confer the capacity to create evolutionary novelty through exon shuffling [2]. The introns-early theory predicts that at least some of the extant eukaryotic introns are direct descendants of the primordial sequences in the LUCA [2–5]. The alternate view, "introns-late", suggests that the last common ancestor was intron-free and that spliceosomal introns have originated in eukaryotes from recent invasions by autocatalytic RNAs (e.g., group II introns) or transposable elements [6–9]. The introns-late view is compatible with the now-established role of exon shuffling in creating eukaryotic genes [10]. It is the ancient origin of introns that is primarily called into question.

In this study, we analyzed the putative spliceosomal introns in Euascomycetes (Ascomycota) small subunit (SSU) and large subunit (LSU) ribosomal (r)RNA genes [11,12] to understand how spliceosomal introns of a recent origin (i.e., introns-late) spread to novel genic sites. Statistical methods were used to study the exon sequences flanking 49 different spliceosomal intron insertion sites in Euascomycetes rRNA and show that the introns interrupt the G – intron – G (hereafter, the intron position is shown with –) proto-splice site that pre-existed in the coding region. A proto-splice site is a short sequence motif that has a high affinity for splicing factors and is a preferred site of intron insertion. The proto-splice site (e.g., MAG – R in pre-mRNA genes [13]) need not be perfectly conserved in organisms but is rather a set of nucleotides that, with some statistical uncertainty, shows a non-random sequence pattern at sites flanking introns. It is also conceivable that proto-splice sites may differ between lineages reflecting, for example, differences in how the spliceosome recognizes introns (e.g., exon definition hypothesis [14,15]).

Our analysis using information theory [16] shows that the significant information is found in exons flanking rRNA spliceosomal introns. We also confirm that introns are not randomly distributed in the primary and secondary structure of the SSU and LSU rRNA and that the group I introns are generally found in the highly conserved (i.e., functionally important) regions of these genes, whereas the spliceosomal introns tend to occur in regions of the rRNA that are not as well conserved or are not directly involved in protein synthesis.

Results
Analysis of Euascomycetes rRNA Spliceosomal Introns
With our data set of 49 (two diatom-specific introns were excluded from this analysis) different spliceosomal intron sites in the SSU and LSU rRNAs of Euascomycetes (alignment available at http://www.rna.icmb.utexas.edu/ANAL

YSIS/FUNGINT/ (for registration details please see http://www.rna.icmb.utexas.edu/cgi-access/access/locked.cgi), we first tested for the presence of a proto-splice site flanking the introns [12]. In this chi-square analysis, the null hypothesis specified that nucleotide usage in 50 nt of exon sequence upstream and downstream of the different intron insertion sites was random and dependent on the nucleotide composition of Euascomycetes SSU and LSU rRNA sequences in general. Previously, we found evidence for the proto-splice site, AG – G, in Euascomycetes rRNA with the greatest support for the G nucleotides (p < 0.001 [12]). The addition of 18 new Euascomycetes SSU and LSU rRNA insertion sites in the new analysis supports this finding (see Fig. 1) but shows strongest evidence for the proto-splice site to encode G – G (p < 0.01 [three degrees of freedom]), with the Gs occurring at frequencies of 65% and 61% in the Euascomycetes rRNAs.

To address the possibility that we were counting as independent events cases where introns may have had a single origin but then spread into neighboring sites through intron sliding [e.g., [11]], we reran the chi-square analysis after removal of all introns that were within 5 nt of each other. This substantially reduced our data set to 30 introns at the following sites; SSU – 265, 297, 330, 390, 400, 514, 674, 882, 939, 1057, 1071, 1083, 1226, 1514; LSU – 678, 711, 775, 824, 830, 858, 978, 1024, 1054, 1091, 1098, 1849, 1903, 1929, 2076, 2445, but addressed independence of intron insertion events. This data set showed significant support for the AG – G proto-splice site with the A, G, and G, occurring at frequencies of 50% (chi-square = 12.56, p = 0.0055), 67% (chi-square = 24.48, p < 0.0000), and 67% (chi-square = 25.35, p < 0.0000), respectively. The AG – G and G – G proto-splice sites occurred in 9 and 15 of these sequences, respectively. The increase in signal of the AG – G proto-splice site with removal of neighboring (potentially slid) introns is consistent with the idea that intron sliding may over time obscure the targets originally used for insertion. It should be noted, however, that this procedure was done by retaining the most 5' intron in each set of neighboring insertions and this may not represent the original intron. Determining the role of intron sliding in creating new lineages of insertions will require a fully resolved Euascomycetes phylogeny (not yet available) that can be used to map intron gains, losses, and potential slides. The present data for the 300 – 337 spliceosomal introns, for example, when mapped on the Euascomycetes tree published in Bhattacharya et al. [11] shows these introns to be distributed in at least 4 divergent clades within the Lecanoromycetes. These introns may be related through the sliding of an ancestral intron but without the presence of one of these insertions in a non-Euascomycetes fungus or a robust phylogeny of this lineage, it will not be possible to unambiguously identify the original site of insertion.
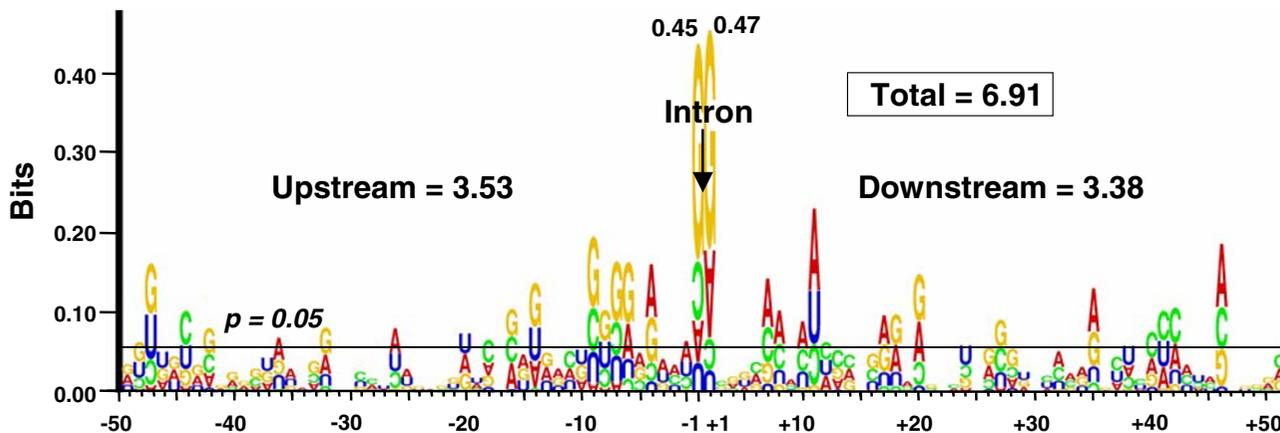
Figure 1
Logo analysis of 50 nt upstream and downstream of insertion sites of 43 different spliceosomal rRNA introns. The information content of the 2 Gs of the intron proto-splice site is shown as is a line at *p* = 0.05 (95% quantile) that is based on simulations using random sequence data. This exon region contains a total of 6.91 bits of information.

Next, we used the "Sequence Logo" method developed by Stephens and Schneider [16] and the expression of Hertz and Stormo [17] to determine the information content in the Euascomycetes rRNA introns and exon flanking sequence. The logo of a subset of 43 of the original 49 spliceosomal introns for which we had complete 50 nt of upstream and 50 nt of downstream exon sequence is shown in Fig. 1. This analysis shows that many of the informative sites encode purines (in particular Gs) and that the region contains a total of 6.91 bits. In general, the information content is highest at the site of intron insertion and the regions within a close proximity (about 10 nt), and decreases as one moves away from this site, with the exception of a significant U+G peak at -48 and C-richness around +40 (Fig. 1). In comparison, the mean value (100,000 iterations) for the total bits of information in a 100 nt random sequence data set was 5.68 bits. The 95% quantile for this distribution was 6.47 bits indicating that the Euascomycetes rRNA exons encode significant information ($p < 0.001$). Logo analysis of the reduced set of 30 non-neighboring spliceosomal introns was consistent with this analysis but showed a stronger signal at the proto-splice site (A = 0.31 bits, G = 0.52 bits, G = 0.59 bits). The finding of significant information in the flanking exons suggests that some regulatory regions (i.e., exonic splicing enhancers, ESEs [18,19] may exist in these sequences.

Sliding Window Analysis of Euascomycetes Spliceosomal Intron Insertion Sites

Intrigued by the finding of G-richness in the upstream exon region flanking introns (see -7 to -17 in Fig. 1), we

determined the association of G-rich regions in 1434 fungal SSU rRNAs and 880 fungal LSU rRNAs with all reported spliceosomal introns in these genes. The G-frequencies were calculated at each rRNA site and are plotted as the green circles in Fig. 2. The SSU (1800 nt [GenBank U53879]) and LSU (3554 nt [U53879]) rRNAs from S. cerevisiae were used as the reference sequence for these alignments. The raw G-frequencies were smoothed (blue curve in Fig. 2), using the loess local regression method [20], and smoothing windows of size 50 nt or 100 nt, prior to analyzing the intron-G-frequency association. The positions of rRNA spliceosomal intron positions are shown as red lines in Fig. 2. From this analysis we can observe that regions of intron insertion strongly associate with high G-frequencies in both the SSU and LSU rRNA. The association is stronger in the 50 nt (i.e., 25 nt exon sequence – intron insertion site – 25 nt exon sequence) window of weighted averages, suggesting that this window size includes most of the exon signal. However, the association is still apparent in the 100 nt window, in particular for the SSU rRNA.

Our analyses show that the average G-frequency at the 25 intron sites using the fitted curve in the SSU rRNA is 0.34, whereas the average G-frequency at the 24 intron sites using the fitted curve in the LSU rRNA is 0.32. To test the significance of this result with the 25 intron sites and the G-contents in the LSU rRNA, we randomly selected 25 sites from the 3554 nt of rRNA and computed the average of their G-frequencies. We repeated this process 10,000 times and plotted the distribution of these average G-frequencies (results not shown). The observed average G-frequency at the LSU intron sites was significantly greater than that in the simulated data (p = 0.0268). Similarly, we carried out the simulation-based test for the SSU rRNA intron sites. In these 10,000 replications, no average from the randomly generated sites was greater than 0.34. Thus, the p-value is less than 0.0001, reinforcing the remarkable association of SSU rRNA introns and G-rich regions apparent in Fig. 2. Taken together, our results suggest that Euascomycetes rRNA spliceosomal introns are fixed at the G – G or AG – G proto-splice site that is found in G-rich regions.

Intron Positions on rRNA Conservation Diagrams
To understand the association of introns with highly conserved regions in the rRNAs, we mapped the intron positions on SSU and LSU rRNA conservation diagrams of the three phylogenetic domains of life and the two eukaryotic organelles (3Dom2O) and the nuclear-encoded rRNA genes in the three phylogenetic domains (3Dom). This analysis shows a significant association of group I intron sites with rRNA sites that are 98–100% conserved within both 3Dom2O and 3Dom LSU rRNA analyses (see Table 1). Only in the 3Dom analysis for SSU rRNA was the asso-

ciation weakly non-significant (p = 0.0577). The observed association of highly conserved rRNA and group I intron sites is, therefore, unlikely to have occurred by chance alone. For rRNA spliceosomal introns, however, the association of conserved rRNA and introns sites is less clear. Within the 3Dom2O analysis of SSU rRNA, spliceosomal intron positions vary significantly from the null model but in the direction of fewer than expected introns at the most highly conserved sites, whereas within the 3Dom analysis of LSU rRNA no significant difference is found (p = 0.0969). The 3Dom2O LSU rRNA and 3Dom SSU rRNA analyses both show an enrichment of spliceosomal introns at the highly conserved genic sites (primarily in sites conserved between 90–97%). Taken together, our analyses suggest that group I introns are fixed primarily in the most highly conserved rRNA sites when analyzed in the 3Dom2O or 3Dom data sets, whereas spliceosomal introns are not strongly associated with highly conserved rRNA sites.

To address more directly the relationship between Euascomycetes spliceosomal introns and rRNA conservation patterns, we positioned these introns on a conservation diagram generated from 1042 fungal SSU rRNA sequences (see Fig. 3). This analysis showed that 19 of 24 fungal SSU rRNA spliceosomal introns follow sites that are conserved in more than 95% of the fungal sequences (1114 nt in this class), one intron follows a site that is 90–95% conserved (149 nt in this class), two introns follow sites that 80–89% conserved (134 nt in this class), and two introns follow sites that <80% conserved (402 nt in this class). More importantly, inspection of the 1800 nt alignment of SSU rRNAs and 3554 nt of LSU rRNAs of all fungi, of fungi containing spliceosomal introns, and of fungi lacking spliceosomal introns shows that most of the introns are inserted between nucleotides that are 99–100% conserved (whether they encode G – G or not) in taxa containing introns and sister groups lacking introns (Table 2). This result provides strong support for the hypothesis that Euascomycetes spliceosomal introns are fixed in a proto-splice site that pre-dates intron insertion. Beyond this pattern of conservation, the G-rich regions in the neighborhood of introns are also often highly conserved among all fungi (see Fig. 3). Most of these Gs are in sites that are >95% conserved in all fungal SSU rRNAs, suggesting that their existence also pre-dates intron insertion.

However, several exceptions to this general pattern merit closer inspection. The upstream nucleotide at the SSU rRNA 297 site (369 in the S. cerevisiae gene), for example, occurs at a frequency of 63.9% U in taxa lacking introns but at a frequency of 97.8% U in taxa containing introns. On the surface, this suggests that the site may have undergone selective pressure, post-intron insertion, towards a high frequency of Us. Analysis of the SSU rRNA alignment
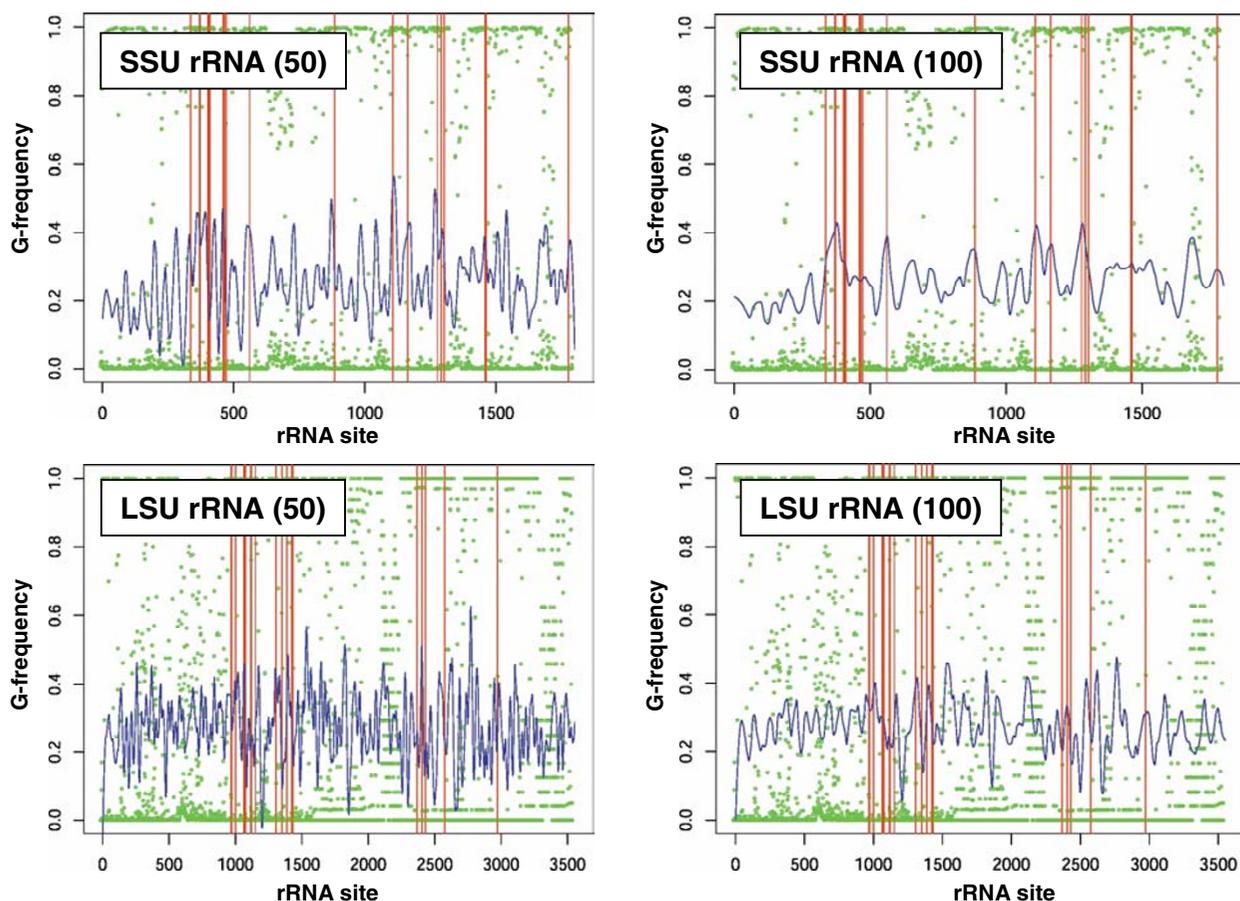
Figure 2
The distribution of SSU and LSU rRNA spliceosomal introns relative to the G-frequency in these genes. The raw G-frequencies are shown in the green circles, the smoothed loess curves for 50 nt and 100 nt smoothing windows are shown with the blue lines, and the positions of introns are shown with the vertical red lines.

shows, however, that the 5 taxa containing the 297 intron share a U at this site with virtually all other intron-containing fungi that lack this particular insertion. This suggests that the high U frequency in the intron-containing fungi is a synapomorphy for the monophyletic intron-containing Euascomycetes and is not an outcome of the 297 intron insertion. A similar result is found when the proto-splice site is checked in all taxa containing introns with those lacking any particular intron.

Intron Positions on the rRNA Primary Structure
The positions of spliceosomal, group I, group II, and archaeal introns were included on a line representing the primary structures of E. coli SSU and LSU rRNA (Fig. 4A). The intron distributions were then studied to determine if they differ significantly from the null hypothesis of a

"broken-stick" distribution [21,22]. This resource division model, which has been used extensively to test hypotheses about patterns of species abundance [e.g., [23]], specifies a distribution that arises when a "stick" of unit length is divided into n number of events with these events scattered with a uniform probability distribution. The events break the stick into n + 1 intervals which can then be studied to determine if they depart from uniformity in the probability density along the stick. Departure will tend to make the longest intervals longer and the shortest intervals shorter [24]. In our analyses, the rRNA genes were the sticks and the intron insertion sites were the events. The metric used to compare the null (i.e., broken-stick) and observed distribution was the standard deviation (SD) from the mean interval length; i.e., lower SDs mean the more uniform are the lengths of the intervals [e.g., [25]].

Table 1: Chi-Square Test of Association of Spliceosomal and Group I Introns with Conserved rRNA Sites

	98–100%	90–97%	80–89%	<80%	Total	P-value
3Dom2O: SSU rRNA						
sites	178	175	116	1073	1542	-
group I	11 [4.85]	5 [4.77]	5 [3.16]	21 [29.23]	42	0.0106*
spliceosomal	0 [3.00]	3 [2.95]	8 [1.96]	15 [18.09]	26	<0.0000*
3Dom2O: LSU rRNA						
sites	150	203	168	2383	2904	-
group I	10 [2.12]	4 [2.82]	4 [2.37]	23 [33.64]	41	<0.0000*
spliceosomal	3 [1.29]	8 [1.75]	2 [1.45]	12 [20.51]	25	<0.0000*
3Dom: SSU rRNA						
sites	355	156	80	951	1542	-
group I	17 [9.67]	3 [4.25]	1 [2.18]	21 [25.90]	42	0.0577
spliceosomal	4 [5.99]	9 [2.63]	2 [1.35]	11 [16.04]	26	0.0003*
3Dom: LSU rRNA						
sites	595	349	283	1677	2904	-
group I	17 [8.40]	5 [4.93]	1 [4.00]	18 [23.68]	41	0.0059*
spliceosomal	10 [5.12]	3 [3.00]	1 [2.44]	11 [14.44]	25	0.0969

Column headings: Introns are positioned relative to SSU and LSU rRNA sites for positions with a nucleotide in more than 95% of the sequences that are 1) 98–100%, 2) 90–97%, 3) 80–89%, and 4) either <80% conserved or positions that are present in <95% of the sequences in genes from; *3Dom2O*, the three phylogenetic domains and two organelles; *3O*, the three phylogenetic domains. Sites are the number of rRNA positions followed by group I and spliceosomal introns in each conservation class and the number of observed and expected introns (in brackets [under a null model of random insertion]) is shown for each gene. The P-values for each analysis are also shown. Significant probability values are marked with an asterisk.

Computer simulations were used to determine the level of significance at which the observed distributions could be distinguished from those produced by the broken-stick model.

A cursory analysis of the data suggests that the intron distribution in both SSU and LSU rRNAs is significantly clustered (in particular, the LSU rRNA) and the statistical analysis bears this out. The observed standard deviations for all the analyses (i.e., all the introns together or the spliceosomal and group I introns individually) are significantly different from the expectations of the broken stick model. The departure from the null model is particularly striking for the LSU rRNA, suggesting that the introns in this gene are more strongly clustered than in the SSU rRNA (see Fig. 4A,4B).

Discussion

In this paper, we have focused on spliceosomal introns in the Euascomycetes fungi to address how introns spread in rRNA (and perhaps in all) genes. Potentially, the rRNA spliceosomal introns offer three major advantages over pre-mRNA introns that are relevant to understanding intron spread: 1) the rRNA spliceosomal introns have been inserted recently within the Euascomycetes [11,12]. In contrast, the sporadic distribution of pre-mRNA

introns in different eukaryotes, and the uncertainty about the phylogenetic relationship of these lineages within the eukaryotic radiation often make it difficult to determine unambiguously which spliceosomal introns are of early or late origins [9]. 2) rRNAs have well-characterized secondary and tertiary structures [e.g., [26,27]]; therefore, if the intron distribution reflects in some way RNA-folding patterns, then one can detect this by mapping the intron distribution on rRNA at the primary, secondary, and tertiary structure levels [28]. 3) rRNA genes do not encode proteins; therefore, the Euascomycetes intron distribution will not reflect constraints on sites of intron insertion due to codon structure. In contrast, the role of intron phase (i.e., between codons [phase 0] or within codons [phases 1,2]) and exon symmetry in explaining pre-mRNA intron distribution remains a controversial and unresolved issue in spliceosomal intron evolution [e.g., [29,30]].

The proto-splice site bounding rRNA introns

Our analysis of 100 nt of exon sequence flanking spliceosomal introns in Euascomycetes rRNA shows significant support for a G – G or AG – G proto-splice site (Fig. 1). The proto-splice site pre-dates intron insertion because it is highly conserved in the Euascomycetes rRNAs in both intron-containing and intron-less taxa (see Fig. 3, Table 2). This finding is not anomalous because analysis of exon
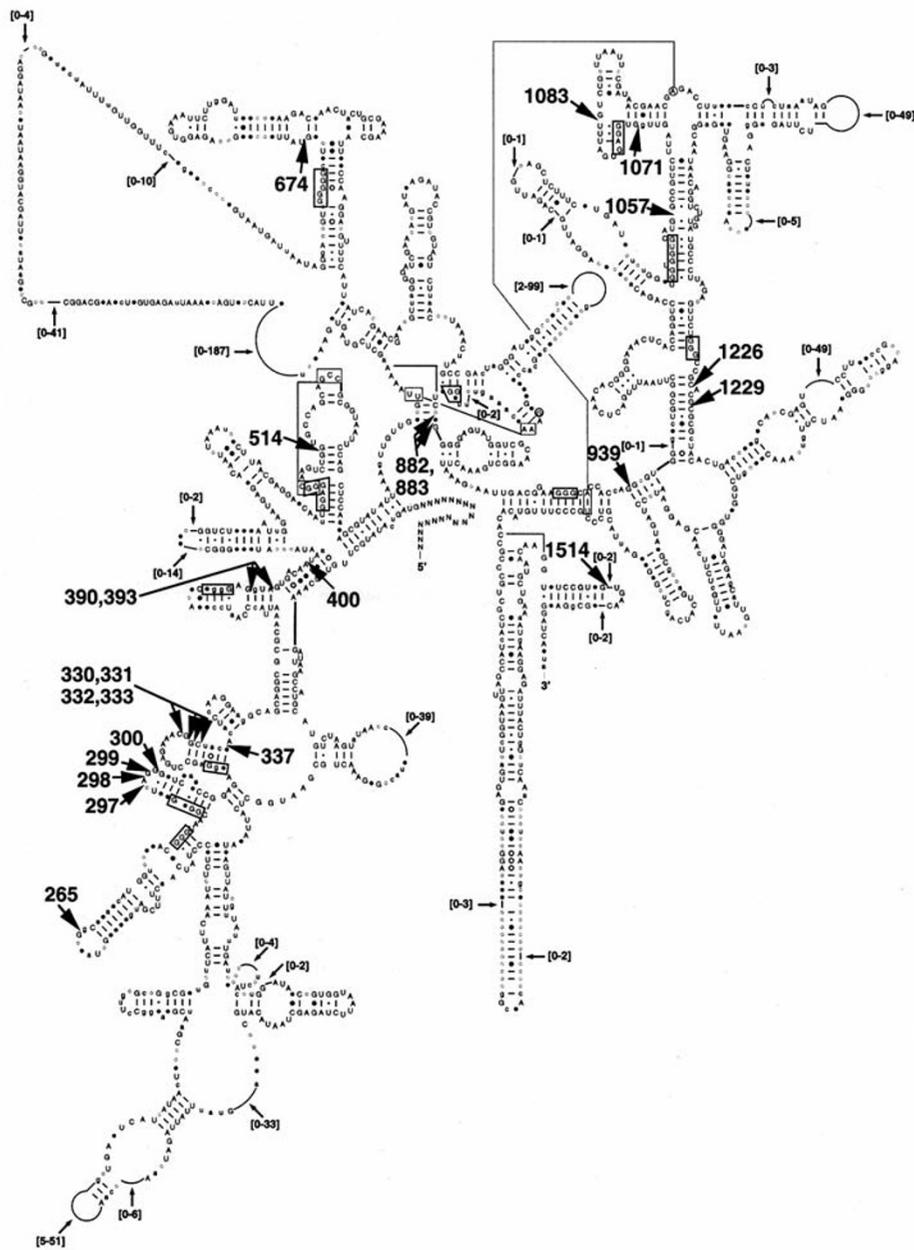
Figure 3
Distribution of Euascomycetes spliceosomal introns on a conservation diagram of fungal SSU rRNA overlaid on a secondary structure model of the *Saccharomyces cerevisiae* SSU rRNA. Spliceosomal introns are shown in large text with arrows denoting their positions. Positions with nucleotides in more than 95% of the 1042 sequences that were studied are shown as following: upper case, conserved at ≥ 95%, lower case, conserved at 90–94%, filled circle, conserved at 80–89%, and open circle, conserved at < 80%. Other regions are denoted as arcs. The numbers at the arcs show the upper and lower number of nucleotides that are found in these variable regions. The boxed regions are G-rich sequences upstream of intron insertion sites. Boxed filled circles indicate that the most frequent nucleotide at this site was a G in our alignment of 1434 fungal rRNAs that included both intron-containing and intron-less taxa.

Table 2: Frequencies of Fungal Nucleotides at Sites of Spliceosomal Intron Insertion

| Intron Position | | | | | Insertion Site | | | | | |
Ec	Sc	All	+ Int	- Int	5'-nt	3'-nt	All	+ Int	- Int	# Int
265	336	99.8	100.0	99.8	G	G	95.0	85.7	95.3	4
297	369	65.3	97.8	63.9	U	A	100.0	100.0	99.5	5
298	370	100.0	100.0	99.5	A	G	100.0	100.0	100.0	2
299	371	100.0	100.0	100.0	G	G	99.8	100.0	99.8	12
300	372	99.8	100.0	99.8	G	G	99.7	100.0	99.6	1
330	402	99.1	100.0	99.1	C	G	99.5	100.0	99.4	15
331	403	99.5	100.0	99.4	G	G	99.7	100.0	99.7	8
332	404	99.7	100.0	99.7	G	C	99.6	100.0	99.6	1
333	405	99.6	100.0	99.6	C	U	91.5	100.0	91.1	1
337	409	76.7	87.0	76.3	C	A	99.8	100.0	99.8	1
390	461	99.3	97.5	99.4	G	G	93.1	100.0	92.9	2
393	464	99.6	100.0	99.6	A	G	99.6	100.0	99.5	10
400	471	99.6	100.0	99.6	A	U	97.6	95.0	97.7	1
514	561	99.5	100.0	99.4	G	G	99.5	100.0	99.4	1
674	885	99.7	100.0	99.7	G	U	99.8	100.0	99.8	4
882	1106	67.3	75.8	67.0	U	G	80.5	84.9	80.3	1
883	1107	80.5	84.9	80.3	G	G	99.4	100.0	99.4	6
939	1164	99.2	97.1	99.3	G	G	98.7	97.1	98.8	8
1057	1277	99.8	100.0	99.8	G	G	98.8	100.0	98.7	1
1071	1291	93.4	100.0	93.3	G	G	99.4	100.0	99.4	1
1083	1303	99.5	100.0	99.5	U	G	99.6	100.0	99.6	1
1226	1459	99.1	100.0	99.1	C	A	99.8	100.0	99.8	2
1229	1462	99.7	100.0	99.7	G	C	99.4	100.0	99.3	8
1514	1777	98.6	100.0	98.5	G	G	91.1	100.0	90.9	2
678	967	99.9	100.0	99.9	G	A	99.9	97.0	100.0	16
681	970	99.7	97.0	99.9	G	G	98.1	93.9	98.3	1
711	1000	99.6	97.0	99.7	G	A	98.2	81.8	99.0	3
775	1065	99.8	100.0	99.8	G	G	100.0	100.0	100.0	1
776	1066	100.0	100.0	100.0	G	G	100.0	100.0	100.0	5
777	1067	100.0	100.0	100.0	G	G	100.0	100.0	100.0	1
780	1070	100.0	100.0	100.0	G	A	100.0	100.0	100.0	1
783	1073	100.0	100.0	100.0	A	G	100.0	100.0	100.0	2
784	1074	100.0	100.0	100.0	G	A	100.0	100.0	100.0	3
786	1076	99.8	100.0	99.8	C	U	95.8	91.2	96.1	1
787	1077	95.8	91.2	96.1	U	A	98.7	91.2	99.1	1
824	1114	100.0	100.0	100.0	U	C	100.0	100.0	100.0	1
830	1120	99.8	100.0	99.8	A	G	99.8	100.0	99.8	1
858	1151	100.0	100.0	100.0	G	G	100.0	100.0	100.0	2
978	1306	99.3	95.7	99.6	G	G	100.0	100.0	100.0	3
1024	1351	98.6	100.0	98.5	A	G	99.3	100.0	99.3	1
1054	1387	97.6	100.0	97.4	G	G	100.0	100.0	100.0	4
1091	1424	100.0	100.0	100.0	G	U	99.3	100.0	99.2	1
1093	1426	100.0	100.0	100.0	G	U	100.0	100.0	100.0	1
1098	1431	100.0	100.0	100.0	A	A	99.3	100.0	99.2	1
1849	2367	100.0	100.0	100.0	U	G	100.0	100.0	100.0	1
1903	2404	97.3	100.0	97.1	G	G	100.0	100.0	100.0	1
1929	2430	97.3	100.0	97.1	G	G	97.3	100.0	97.1	1
2076	2576	100.0	100.0	100.0	G	A	100.0	100.0	100.0	1
2445	2972	100.0	100.0	100.0	G	G	100.0	100.0	100.0	1

Column headings: *Intron Position*, the sites of spliceosomal intron insertion in the SSU and LSU (below the broken line) rRNA genes. The homologous intron sites in the *Escherichia coli* (*Ec*, GenBank #J01695) and *Saccharomyces cerevisiae* (*Sc*, GenBank #U53879) genes are shown. The 5' and 3' nucleotides (*5'-nt*, *3'-nt*) flanking the intron insertion sites (*Insertion Site*), the frequency of these nucleotides in the alignment of all fungal SSU and LSU rRNAs (*All*, 1434 and 880 sequences, respectively), of fungi containing spliceosomal introns (*+ Int*, 73 and 40 sequences, respectively), and of fungi lacking spliceosomal introns (*- Int*, 1361 and 840 sequences, respectively), and the number of taxa containing introns at each site (*# Int*) are shown.
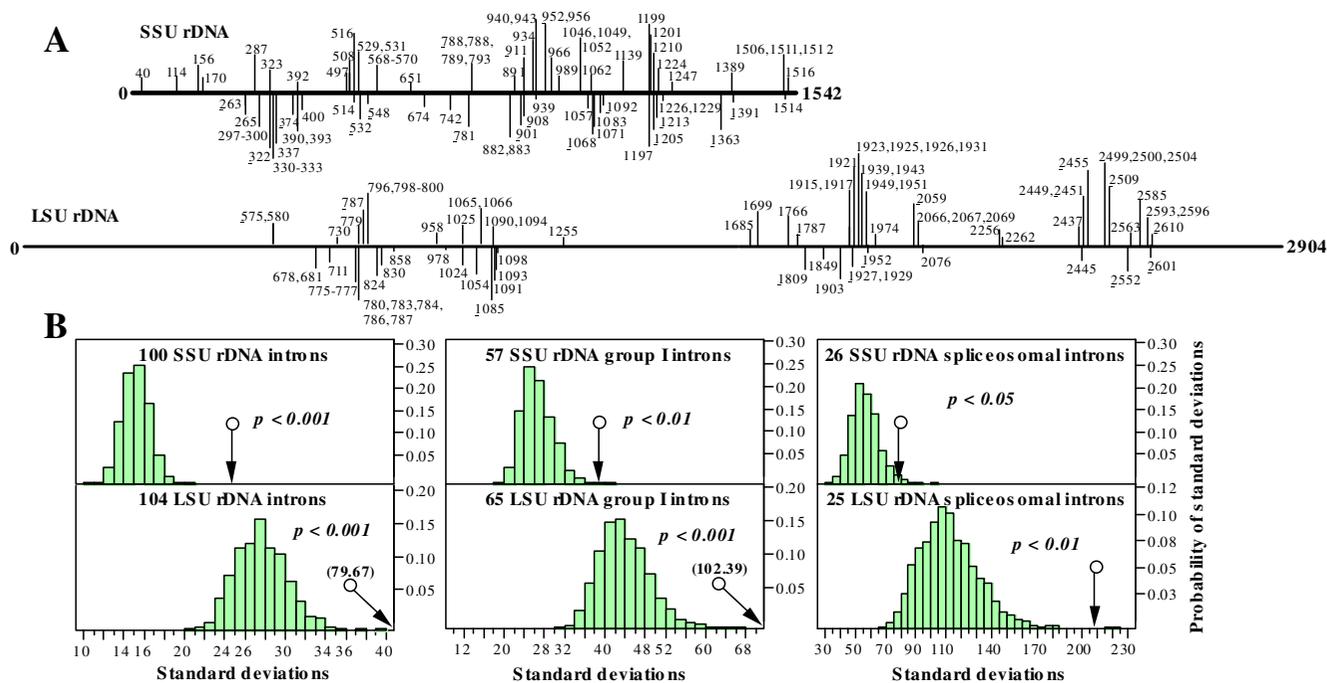
Figure 4
Analysis of rRNA intron distribution. **A**. The positions of introns mapped on the homologous sites in the primary structure of *E. coli* SSU and LSU rRNA. Group I and group II (underlined) introns are shown above the lines, whereas spliceosomal and archaeal (underlined) introns are shown below the lines. **B**. Results of the broken-stick analysis of rRNA intron distribution. The results of the simulations are shown as are the observed standard deviations for all introns or group I and spliceosomal introns individually for both SSU and LSU rRNA genes.

sequences surrounding the total set of introns in *S. cerevisiae* pre-mRNA genes shows a preference for AAAG at the 5' splice site [31]. The final G in this motif has been established as significantly conserved in yeast [32]. The sequence at the proximal 5' exon region is required for interactions with the spliceosomal small nuclear ribonucleoprotein particle U1 [19]. Our data are, therefore, consistent with present understanding of yeast pre-mRNA splicing. Furthermore, taking at least 40% as the minimum for a consensus nucleotide in the proto-splice site, Long et al. [33] have shown that this region in six model eukaryotes often encode the AG – G or G – G motif. In humans, for example, the nucleotides in the AG – G motif are found in abundances of 61%, 81%, and 56%, respectively. The finding of a similar motif in rRNA genes

for which there is neither a requirement to incorporate amino acid phase distribution nor to invoke exon-shuffling provides support for the idea that a proto-splice site for intron insertion not only exists in Euascomycetes rRNA but also may exist in pre-mRNA genes. The introns appear to be inserted into some of the most conserved regions of Euascomycetes SSU rRNA, as evident in the fungal conservation diagram (Fig. 3) and the analysis of fungal nucleotide frequencies at the 5' and 3' nt flanking introns (Table 2). However, the spliceosomal introns do not map to the most conserved positions in the 3Dom or 3Dom2O rRNA datasets (Table 1).

Furthermore, exon sequences, outside of the proto-splice site, may be required for splice site recognition by the spli-

ceosome [34–38]. Our rRNA analyses suggest that G-rich regions in the neighborhood (often upstream) of the intron insertion sites may be potential ESEs. The exon context may, therefore, play a fundamental role in controlling intron splicing and, thus, sites of intron fixation. This idea has growing support in the literature [e.g., [19,38,39]]. Combined with this observation is the finding that rRNA spliceosomal introns map primarily to regions in the interface surface of the SSU and LSU ribosome [28]. These sites presumably facilitate intron splicing during ribosome biogenesis.

We find that in contrast to the spliceosomal introns in rRNA, group I intron insertion sites show a stronger positive association with highly conserved rRNA regions (Fig. 3, Table 2), including those that bind tRNA [28], and are more clustered than are spliceosomal introns in the rRNA primary structure (Fig. 4). This suggests that group I intron fixation may be even more highly constrained by the exon context than are spliceosomal introns. A possible explanation for this observation is that group I introns are more dependent on specific upstream and downstream exon sequences to build the P1 and P10 domains [40] to facilitate proper folding prior to excision [e.g., [41]]). This could limit the number of rRNA sites at which group I introns can be fixed in comparison to spliceosomal introns which have less specific exon sequence requirements for splicing.

Conclusions
Our findings provide concrete insights into rRNA intron fixation and are more compatible with the view that both the spliceosomal and group I intron distributions reflect fundamental features of present-day genes and genomes and that introns may not be relics of an ancient intron-rich period of cells. An intriguing view on intron origin was recently published using the tools of population genetics. In this view, the richness of introns in multicellular organisms may primarily reflect the smaller population sizes of these taxa relative to protists, which generally contain few introns. The large population sizes of unicellular eukaryotes may prevent widespread intron spread due to secondary mutations that lead to their loss from populations [42]. Interestingly, the lichenized Euascomycetes, which are particularly rich in both spliceosomal and group I introns in their nuclear rRNA, are typically extremely slow-growing taxa many of which have small population sizes [e.g., [43]].

Methods
PCR Methods and the Intron Data
The spliceosomal introns described in Bhattacharya et al. [12], plus 12 new positions that have become available in GenBank, were used in this study, as well as 6 new sites that we have found in the LSU rRNA genes of *Buellia capi-*

tis-regum, *Buellia muriformis*, *Ionaspis lacustris*, *Physconia enteroxantha*, and *Rinodina tunicata*. To allow direct comparison between all rRNAs, the numbering of introns reflects their relative positions in the *E. coli* coding regions. DNA samples for *Buellia* spp., *Rinodina*, and *Physconi* were generously provided by T. Friedl (Göttingen). Tissue from *Ionaspis* was a gift from F. Lutzoni (Duke). DNA was extracted from *Ionaspis* as in Bhattacharya et al. (2000). PCR reactions were done with the following primers: 1825-5'GTGATTTCTGCCCAGTGCTC3', 2252-5'TTTAACAGATGTGCCGCC3', 2252-5'GGCGGCACATCTGTTAAA3', and 2746-5'GATTCTGRCTTAGAGGCGTTC3'. The primer names refer to their position relative to the LSU rRNA of *E. coli*. PCR amplification products were cloned in the pGEM-T (Promega) vector and sequenced over both strands. Together, the fungal spliceosomal data set included 49 different introns at the following sites (the species from which they were isolated and GenBank accession numbers, where available, are also shown): SSU rRNA – 265 (*Arthroraphis citrinella*, AF279375), 297 (*Anaptychia runcinata*, AJ421692), 298 (*Physconia perisidiosa*, AJ421689), 299 (*Roccella canariensis*, AF110342), 300 (*Rhynchostoma minutum*, AF242268), 330 (*Stereocaulon paschale*, AF279412), 331 (*Physconia perisidiosa*, AJ421689), 332 (*Pyrenula cruenta*, AF279406), 333 (*Pertusaria amara*, AF274104), 337 (*Graphis scripta*, AF038878), 390 (*Dermatocarpon americanum*, AF279383), 393 (*Hymenelia epulotica*, AF279393), 400 (*Halosarpheia fibrosa*, AF352078), 514 (*Porpidia crustulata*, L37735), 674 (*Physconia detersa*, AJ240495), 882 (*Dimerella lutea*, AF279386), 883 (*Diploschistes scruposus*, AF279388), 939 (*Dimerella lutea*, AF279386), 1057 (*Graphina poitiaei*, AF465459), 1071 (*Rhynchostoma minutum*, AF242268), 1083 (*Rhamphoria delicatula*, AF242267), 1226 (*Rhynchostoma minutum*, AF242269), 1229 (*Physconia perisidiosa*, AJ421689), and 1514 (*Phialophora americana*, X65199); LSU rRNA – 678 (*Gyalecta jenensis*, AF279391), 681 (*Stictis radiata*, AF356663), 711 (*Gyalecta jenensis*, AF279391), 775 (*Dibaeis baeomyces*, AF279385), 776 (*Capronia pilosella*, AF279378), 777 (*Rinodina tunicata*, AF457569), 780 (*Pertusaria tejocotensis*, AF279301), 784 (*Melanochaeta* sp. 8, AF279421), 786 (*Pertusaria kalelae*, AF279298), 787 (*Dibaeis baeomyces*, AF279385), 824 (*Stictis radiata*, AF356663), 830 (*Coenogonium leprieurii*, AF465442), 858 (*Trapeliopsis granulosa*, AF279415), 978 (*Gyalecta jenensis*, AF279391), 1024 (*Ocellularia alborosella*, AF465452), 1054 (*Rinodina tunicata*, AF457569), 1091 (*Dimerella lutea*, AF279387), 1093 (*Ocellularia alborosella*, AF465452), 1098 (*Coenogonium leprieurii*, AF465442), 1849 (*Cordyceps prolifica*, AB044640), 1903 (*Physconia enteroxantha*, AF457573), 1929 (*Buellia capitis-regum*, AF457572), 2076 (*Buellia muriformis*, AF457571), and 2445 (*Ionaspis lacustris*, AF457570). We did not study the 742 and 1197 spliceosomal introns in the SSU rRNA gene

of the distantly related stramenopile, *Cymatosira belgica* (X85387). This diatom is the sole known organism outside of the fungi to contain rDNA spliceosomal introns. In addition, the fungal 674, 1057, 1514 SSU rRNA and 1093, 1098, and 1849 LSU rRNA introns were not included in the information analysis because of missing data or ambiguous sequences (see below). The SSU 674 site (*Physconia detersa*, AJ240495), for example, only included 10 nt of the 5' and 3' region in the GenBank accession [11]. All fungal and diatom intron sites were, however, mapped on the conservation diagrams to understand their distribution (see below).

We have made, on the basis of detailed analysis of rRNA flanking regions, a number of corrections in the positions of the introns within the SSU rRNA (e.g., 1129 is now at 1229 and 1510 is now at 1514). Copies of the manuscript figures and tables and additional materials related to this work are available from the Gutell Laboratory's CRW Site at http://www.rna.icmb.utexas.edu/ANALYSIS/FUN GINT/[44]. This page includes detailed rRNA conservation and intron position data (both the version used for the manuscript and current values that are updated daily), fungal nucleotide frequency values, and the SSU and LSU rRNA sequence alignments used in Table 2.

Information Analysis of the rRNA Introns

An information analysis was done of the 50 nt upstream and downstream of the different rRNA spliceosomal intron sites to determine the total amount of exonic information (in "bits") that is available to the spliceosome for splicing. We used the web-based logo program of Gorodkin et al. [45]http://www.cbs.dtu.dk/~gorodkin/ appl/slogo.html to derive the sequence logos and the information content of individual sites was calculated according to the expression of Hertz and Stormo [17]. Type 2 logos were drawn in which the height of the nucleotides in the sequence column represented their frequency in proportion to their expected frequency. The expected nucleotide probabilities were estimated from the observed nucleotide frequencies over all sites for 80 Euascomycetes rRNA sequences (A = 26%, C = 22%, G = 27%, T = 25% [12]). The nucleotides were turned upside-down when the observed frequency was less than expected [45]. A total of 43 spliceosomal intron sites, for which 50 nt of both upstream and downstream exon sequence are available, were included in this analysis.

To put the information content in perspective, we also did simulations in which 43 random sequence data sets of length 100 nt (for flanking exons) and 109 (total number of introns analyzed) random data sets of length 29 nt (for conserved intron regions) were generated at the nucleotide frequencies of Euascomycetes rRNA and the information content of these was calculated. A total of 100,000

iterations were done with each data set to create null distributions of random information content. The observed information values were then compared to the null distributions to infer their probabilities.

Analysis of G-Content in Euascomycetes SSU rRNAs

Because it is difficult to see the pattern of G-content along the sequence based on the raw data, we fit a smooth curve to the frequencies of G using the method of local regression (loess, [20]). This smooth curve captures the G-content pattern along the nucleotide sites. Loess is a nonparametric curve fitting technique that fits the data in a local fashion. That is, for the fit at site x, the fit is made using the G-frequencies at the points in a neighbourhood of x, weighted by their distance from x. A tricubic weighting function (proportional to $[1 - (distance/max\ distance)^3)^3]$) is used for calculating the weights. For both the LSUrRNA and SSUrRNA sequence alignment data sets, we used a neighborhood of 50 nt (and 100 nt) in fitting the loess curve. Thus the value of the curve at each site is computed as a weighted average of the G-frequency at the site itself, the G-frequencies at the 25 up-stream sites, and the G-frequencies at the 25 down-stream sites.

Positions of Introns Relative to Conserved rRNA Regions

To assess the patterns of sequence conservation in exon sequences flanking all rRNA spliceosomal and group I introns, we mapped intron positions on structure conservation diagrams. Group I introns in different subclasses (e.g., IC1, IE [46,47]) which occupied the same rRNA site were counted as separate intron insertions. This accounted for our observation that certain rRNA sites (e.g., SSU 788, 1199, LSU 1949, 2500 [see CRW Site for details]) are "hot" spots for insertion with multiple, evolutionarily divergent introns being fixed at the same site in different species or in different genomes (i.e., nuclear vs. organellar). The actual number of independent hits at rRNA sites is, however, likely to be much greater than our estimate but this can only be proven with rigorous phylogenetic analysis of group I introns at different insertion sites to show that in some cases, introns in the same subclass at the same site in different species have a high probability of independent origin [e.g., [48,49]]. The first set of conservation diagrams used in our analysis was based on the comparison of 6389 and 922 different SSU and LSU rRNA sequences, respectively, from the three phylogenetic domains and the two organelles (3Dom2O) that were superimposed on the secondary structures of the *Escherichia coli* rRNAs. The second set of diagrams was a summary of 5591 and 585 different SSU and LSU rRNA sequences, respectively, from the three phylogenetic domains (3Dom) also mapped on the *E. coli* rRNAs. These diagrams are available at the CRW Site. Multiway contingency table analysis was done to determine whether sites that were 98–100%, 90–97%, 80–89%, and <80% con-

served in the diagrams were independent of intron insertion sites (the null hypothesis). Intron sites were taken as the nucleotide immediately preceding the intron insertion. We also calculated nucleotide frequencies for each SSU and LSU rRNA site using the *S. cerevisiae* genes for numbering. Frequencies were calculated for alignments of all available fungal rRNAs (1434 SSU and 880 LSU sequences) and of only fungi containing spliceosomal introns (73 SSU and 40 LSU sequences), or of fungi lacking spliceosomal introns (1361 sequences for SSU, 840 for LSU). These frequencies were used to determine the level of conservation of nucleotides encoding the proto-splice site in intron-containing and intron-less fungal species.

rRNA Intron Distribution
The positions of all known spliceosomal, group I, group II, and tRNA-like archaeal [50] introns were marked on the primary structures of *E. coli* SSU and LSU rRNA. These data, which also accounted for multiple group I intron hits at the same rRNA site, were then studied to determine whether they differ significantly from the null expectation of a random distribution (i.e., "the broken stick distribution"). We used the program PowerNiche V1.0 (P. Drozd, V. Novotny, unpublished data) to generate sticks of length 1542 nt (SSU rRNA) or 2904 nt (LSU rRNA) which were randomly broken by $n = 101$ events for all introns (including group II and archaeal), or $n = 56$ for only group I, or $n = 26$ for only spliceosomal introns in SSU rRNA. For the LSU rRNA, the stick was broken into $n = 107$ events for all introns, or $n = 68$ for only group I, or $n = 25$ for only spliceosomal introns. The paucity of rRNA group II introns (3 and 8 introns in the SSU and LSU rRNA, respectively) and archaeal introns (14 and 6 introns in the SSU and LSU rRNA, respectively) did not allow their individual analysis. A mean number of intervals and a SD were calculated for each broken-stick. The SDs of 1000 simulations were compared to the SD of the observed data to test whether the observed pattern was likely to have been produced under the assumptions of the broken-stick model.

Authors' Contributions
DS generated the new intron sequences. JH did the statistical analyses of G-frequencies and information content. JJC and RRG established and maintain the CRW database and the rRNA-intron database, generated the rRNA G-frequencies and the yeast conservation diagram, and produced the 3Dom and 3Dom2O rRNA conservation data. DB conceived of the study, did the broken-stick analysis, participated in the design and coordination of the other analyses, and wrote the paper. All authors read, modified, and approved the final manuscript.

Acknowledgements

References
1. Burge CB, Tuschl T and Sharp PA: **Splicing precursors to mRNAs by the spliceosomes.** In: *The RNA World* 2nd edition. Edited by: Gesteland RF, Cech TR, Atkins JF. Cold Spring Harbor, Cold Spring Harbor Laboratory, New York; 1999:525-560.
2. Gilbert W: **The exon theory of genes.** *Cold Spring Harb Symp Quant Biol* 1987, **52**:901-905.
3. Doolittle WF: **Genes in pieces: were they ever together?** *Nature* 1978, **272**:581-582.
4. Darnell JE and Doolittle WF: **Speculations on the early course of evolution.** *Proc Natl Acad Sci USA* 1986, **83**:1271-1275.
5. Roy SW, Federov A and Gilbert W: **The signal of ancient introns is obscured by intron density and homolog number.** *Proc Nat Acad Sci USA* 2002, **99**:15513-15517.
6. Cavalier-Smith T: **Intron phylogeny: A new hypothesis.** *Trends Genet* 1991, **7**:145-148.
7. Palmer JD and Logsdon JM Jr: **The recent origins of introns.** *Curr Opin Genet Dev* 1991, **1**:470-477.
8. Stoltzfus A, Spencer DF, Zuker M, Logsdon JM Jr and Doolittle WF: **Testing the exon theory of genes:the evidence from protein structure.** *Science* 1994, **265**:202-207.
9. Logsdon JM Jr: **The recent origins of spliceosomal introns revisited.** *Curr Opin Genet Dev* 1998, **8**:637-648.
10. Patthy L: **Genome evolution and the evolution of exon-shuffling-a review.** *Gene* 1999, **238**:103-114.
11. Cubero OF, Bridge PD and Crespo A: **Terminal-sequence conservation identifies spliceosomal introns in ascomycete 18s RNA genes.** *Mol Biol Evol* 2000, **17**:751-756.
12. Bhattacharya D, Lutzoni F, Reeb V, Simon D and Fernandez F: **Widespread occurrence of spliceosomal introns in the rRNA genes of ascomycetes.** *Mol Biol Evol* 2000, **17**:1971-1984.
13. Dibb NJ and Newman AJ: **Evidence that introns arose at protosplice sites.** *EMBO J* 1989, **8**:2015-2021.
14. Berget SM: **Exon recognition in vertebrate splicing.** *J Biol Chem* 1995, **270**:2411-2414.
15. McCullough AJ and Berget SM: **G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection.** *Mol Cell Biol* 1997, **17**:4562-4571.
16. Stephens RM and Schneider TD: **Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites.** *J Mol Biol* 1992, **228**:1124-1136.
17. Hertz GZ and Stormo GD: **Identification of consensus patterns in unaligned DNA and protein sequences: a large-deviation statistical basis for penalizing gaps.** In: *Proceedings of the Third International Conference on Bioinformatics and Genome Research* 2nd edition. Edited by: Lim HA, Cantor CR. Singapore, Scientific Publishing Co Ltd; 1995:201-216.
18. Blencowe BJ: **Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases.** *Trends Biochem Sci* 2000, **25**:106-110.
19. Graveley BR: **Sorting out the complexity of SR protein functions.** *RNA* 2000, **6**:1197-1211.
20. Cleveland WS: **Robust locally weighted regression and smoothing scatterplots.** *J Amer Statist Assoc* 1979, **74**:829-836.
21. MacArthur RH: **On the relative abundance of bird species.** *Proc Natl Acad Sci USA* 1957, **43**:293-295.
22. Pielou EC: **Mathematical Ecology.** *Wiley* 1977.
23. MacArthur RH: **On the relative abundance of species.** *Am Nat* 1960, **94**:25-36.
24. Goss PJE and Lewontin RC: **Detecting heterogeneity of substitution along DNA and protein sequences.** *Genetics* 1996, **143**:589-602.
25. Baumiller TK and Ausich WI: **The Broken-Stick model as a null hypothesis for crinoid stalk taphonomy and as a guide to the distribution of connective tissue in fossils.** *Paleobiol* 1992, **18**:288-298.
26. RR Gutell: **Comparative sequence analysis and the structure of 16S and 23S rRNA.** In: *Ribosomal RNA: Structure, Evolution, Processing, and Function in Protein Biosynthesis* Edited by: Zimmermann RA, Dahlberg AE. Boca Raton, CRC Press; 1996:111-128.

27. Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JH and Noller HF: **Crystal structure of the ribosome at 55 Å resolution.** *Science* 2001, **292:**883-896.
28. Jackson SA, Cannone JJ, Lee JC, Gutell RR and Woodson SA: **Distribution of rRNA introns in the three-dimensional structure of the ribosome.** *J Mol Biol* 2002, **323:**215-234.
29. Gilbert W, de Souza SJ and Long M: **Origin of genes.** *Proc Natl Acad Sci USA* 1997, **94:**7698-7703.
30. Wolf YI, Kondrashov FA and Koonin EV: **No footprints of primordial introns in a eukaryotic genome.** *Trends Genet* 2000, **16:**333-334.
31. Spingola M, Grate L, Haussler D and Ares M Jr: **Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*.** *RNA* 1999, **5:**221-234.
32. Lopez PJ and Séraphin B: **Genomic-scale quantitative analysis of yeast pre-mRNA splicing: implications for splice-site recognition.** *RNA* 1999, **5:**1135-1137.
33. Long M, de Souza SJ, Rosenberg C and Gilbert W: **Relationship between "proto-splice sites" and intron phases:evidence from dicodon analysis.** *Proc Natl Acad Sci USA* 1998, **95:**219-223.
34. Green MR: **Pre-mRNA splicing.** *Annu Rev Genet* 1986, **20:**671-708.
35. Newman AJ and Norman C: **U5 snRNA interacts with exon sequences at 5' and 3' splice sites.** *Cell* 1992, **68:**743-754.
36. Ramchatesingh J, Zahler AM, Neugebauer KM, Roth MB and Cooper TA: **A subset of SR proteins activates splicing of the cardiac troponin T alternative exon by direct interactions with an exonic enhancer.** *Mol Cell Biol* 1995, **15:**4898-4907.
37. Graveley BR, Hertel KJ and Maniatis T: **A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers.** *EMBO J* 1998, **17:**6747-6756.
38. Lam BJ and Hertel KJ: **A general role for splicing enhancers in exon definition.** *RNA* 2002, **8:**1233-1241.
39. Orban TI and Olah E: **Expression profiles of BRCA1 splice variants in asynchronous and in G1/S synchronized tumor cell lines.** *Biochem Biophys Res Commun* 2001, **280:**32-38.
40. Cech TR, Damberger SH and RR Gutell: **Representation of the secondary and tertiary structure of group I introns.** *Nature Struc Biol* 1994, **1:**273-280.
41. Woodson SA and Emerick VL: **An alternative helix in the 26S rRNA promotes excision and integration of the *Tetrahymena* intervening sequence.** *Mol Cell Biol* 1993, **13:**1137-1145.
42. Lynch M: **Intron evolution as a population-genetic process.** *Proc Natl Acad Sci USA* 2002, **99:**6118-6123.
43. Zoller S, Lutzoni F and Scheidegger C: **Genetic variation within and among populations of the threatened lichen *Lobaria pulmonaria* in Switzerland and implications for its conservation.** *Mol Ecol* 1999, **8:**2049-2059.
44. Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV and Mueller KM *et al.*: **The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs.** *BMC Bioinformatics* 2002, **3:**2.
45. Gorodkin J, Heyer LJ, Brunak S and Stormo GD: **Displaying the information contents of structural RNA alignments: the structure logos.** *Comput Appl Biosci* 1997, **13:**583-586.
46. Michel F and Westhof E: **Modeling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis.** *J Mol Biol* 1990, **216:**585-610.
47. Suh SO, Jones KG and Blackwell M: **A group I intron in the nuclear small subunit rRNA gene of *Cryptendoxyla hypophloia*, an ascomycetous fungus.** *J Mol Evol* 1999, **48:**493-500.
48. Bhattacharya D and Oliveira M: **The SSU rDNA coding region of a filose amoeba contains a group I intron lacking the universally conserved G at the 3'-terminus.** *J Euk Microbiol* 2000, **47:**585-589.
49. Bhattacharya. D, Cannone JJ and Gutell RR: **Group I intron lateral transfer between red and brown algal ribosomal RNA.** *Curr Genet* 2001, **40:**82-90.
50. Kjems J and Garrett RA: **Ribosomal RNA introns in archaea and evidence for RNA conformational changes associated with splicing.** *Proc Natl Acad Sci USA* 1991, **88:**439-443.