J Mol Evol
DOI 10.1007/s00239-008-9090-8

Dropout Alignment Allows Homology Recognition and Evolutionary Analysis of rDNA Intergenic Spacers

Seongho Ryu · Yoonkyung Do · David H. A. Fitch · Won Kim · Bud Mishra

Received: 27 August 2007 / Accepted: 21 February 2008

Abstract Subrepeats within the ribosomal gene (rDNA) intergenic spacer (IGS) play an important role in enhancing RNA polymerase I transcription. Despite this functional role and presumed selective constraint, there is surprisingly little sequence similarity among IGS subrepeats of different species. This sequence dissimilarity corresponds with the fast insertion-deletion (indel) rates observed in short mononucleotide microsatellites (here referred to as poly[N] runs, where N is any nucleotide), which are relatively abundant in rDNA IGS subrepeats. Some species have different types of IGS subrepeats that share species-specific poly(N) run patterns. This finding indicates that many IGS subrepeats within species share a common evolutionary history. Furthermore, by aligning sequences after modifying them by the dropout method, i.e., by disregarding poly(N) runs during the sequence aligning step, we sought to uncover evolutionarily shared similarities that fail to be recognized by current alignment programs. To ensure that the improved similarities in the computed alignments are not a chance artifact, we calibrated and corrected the IGS subrepeat sequences for the influence of repeat length and estimated the statistical significance of the alignments (in terms of a stringent *p*-value) obtained by the dropout method by comparing them to null models constructed using random sequence sets from the same genomes. We found that most diverse kinds of rDNA IGS subrepeats in one species must have been derived from a common ancestral subrepeat, and that it is possible to infer the evolutionary relationships among the IGS subrepeats of different species by comparative genomics methods based on dropout alignments.

Keywords Comparative genomics · Ribosomal DNA · Intergenic spacer · Dropout alignment method · Subrepeat · Homopolymeric runs · Mononucleotide microsatellites · Poly(N)

Electronic supplementary material The online version of this article (doi:10.1007/s00239-008-9090-8) contains supplementary material, which is available to authorized users.

S. Ryu (✉) · D. H. A. Fitch
Department of Biology, New York University, Main Building, Room 1009, 100 Washington Square East, New York, NY 10003, USA
e-mail: Seongho@nyu.edu

S. Ryu · B. Mishra (✉)
NYU/Courant Bioinformatics Group, Courant Institute, New York University, 251 Mercer Street, New York, NY 10012, USA
e-mail: mishra@nyu.edu

Y. Do
Laboratory of Cellular Physiology and Immunology, The Rockefeller University, New York, NY 10065, USA

W. Kim
Department of Biological Sciences, Seoul National University, Seoul, Korea

B. Mishra
Department of Cell Biology, NYU School of Medicine, New York, NY 10016, USA

Introduction

For the last several decades, molecular and evolutionary biologists have been intensely studying the intergenic spacer (IGS) region of the ribosomal DNA (rDNA) genes, which separates the 28S and the 18S rDNA coding regions. Not only does the IGS of higher eukaryotes play an important part in RNA polymerase I transcription, but also it contains broadly conserved structural features such as

several kinds of repeating elements (or subrepeats), repetitive enhancer elements, duplicated promoters, and conserved secondary structures, which are useful to study in the context of molecular evolution (Baldridge et al. 1992; Kahl 1988; Reeder 1989; Ruiz Linares et al. 1991; Sollner-Webb and Tower 1986). The rDNA IGS, which is composed of the nontranscribed spacer (NTS) and the external transcribed spacer (ETS) regions, contains typically many reiterated subrepeats (Baldridge et al. 1992; Kahl 1988; Mandal 1984), with one known exception occurring in *Caenorhabditis elegans*, which has a simple, short structure and no subrepeats in the IGS region (Ellis et al. 1986). There are length variations of the IGS in most species. However, the IGS has not lent itself as a useful tool for phylogenies of species that are not very closely related, not only because the IGS has a large number of reiterated subrepeats but also because the subrepeats' lengths and primary sequences are too dissimilar to be aligned properly (Black et al. 1989; MacIntyre 1985; Murtif and Rae 1985; Rogers et al. 1993). Consequently, IGS sequences could only be usefully employed in phylogenetic studies of very closely related species (Bhatia et al. 1996; Borisjuk and Hemleben 1993; Cordesse et al. 1993; Da Rocha and Bertrand 1995; King et al. 1993; Tautz et al. 1987).

In a previous study, we described the rDNA IGS region in the swimming crab, *Charybdis japonicaa*, and reported that the swimming crab IGS also shows a typical IGS structural pattern, which has repetitive subrepeats (Ryu et al. 1999). Especially, three size classes of the swimming crab subrepeats, 60, 142, and 391 bp, showed high similarity values, signifying that they shared a common ancestor. We suggested one type of subrepeat (60-bp subrepeats; type c) as a prototype for other types. Nevertheless, the primary structures of subrepeats in the swimming crab are quite divergent. One reason for this divergence may be frequent unequal crossing-over and mutation. It has been a well-accepted model that repeated DNA sequences evolve through successive cycles of tandem duplication and divergence of an ancestral sequence (Dover and Tautz 1986; Grellet et al. 1986; Stark et al. 1989). Similarly, the evolution of the rDNA IGS is thought to include duplication and deletion or divergence processes, resulting in a dynamic change in the subrepeat composition of the IGS (Barker et al. 1988; Cordesse et al. 1993; Ryu et al. 1999). On the other hand, gene conversion and other processes of "concerted evolution" are predicted to maintain similarity among sequences within a subfamily of repeats (Dover and Tautz 1986). Tandem duplication of single nucleotides, perhaps by polymerase "stuttering," producing homopolymeric runs, is thought to be another factor resulting in divergence between new types of subrepeats and the original subrepeats (Cunningham et al. 1991; Jacques et al. 1994). Thus, it is not surprising that DNA sequences have many homopolymeric runs, which are defined as two or more identical consecutive nucleotides. Previous studies showed that genome sequences from many species have long stretches of homopolymeric runs. These homopolymeric runs are also frequently referred to as mononucleotide microsatellites (Denver et al. 2004) but are abbreviated here to the term poly(N). In general, poly(A) or poly(T) runs are more abundant in each taxon than poly(C) and poly(G) runs (Toth et al. 2000). Intergenic spacer regions contain more poly(A/T) than poly(C/G) in each taxon except *C. elegans* (Toth et al. 2000). However, these distributions differ when constrained to relatively short poly(N) runs, such as 2–10 bp. Accounting for these differences is likely to aid reconstruction of IGS evolution, since IGSs contain more short poly(N) runs than expected.

In this study, we observed relatively fast insertion rates for the poly(N) runs in the rDNA IGS and even more rapid rates in the subrepeats within IGS sequences. These rapid rates led us to hypothesize that changes in poly(N) fraction could better explain the nature of subrepeat divergence. We characterized patterns of poly(N) runs in various IGS subrepeats and studied how these poly(N) runs can affect the primary structure of the subrepeats. Furthermore, with the goal of extending this analysis further, we developed a dropout alignment algorithm, which can mask the differences induced by the poly(N) runs and recover the obscured underlying phylogenetic signals from IGS subrepeat comparisons.

Materials and Methods

Sequence Data

The data for various subrepeats of the rDNA IGS region in higher eukaryotes were obtained from published rDNA IGS sequences and GenBank. We used a total of 28 types of available rDNA IGS sequences and 44 types of subrepeats from those species: 3 types of subrepeats from 3 mammalian species, 6 types of subrepeats from 3 amphibian species, 9 types of subrepeats from 5 insect species, 3 types of subrepeats from 2 crustacean species, 22 types of subrepeats from 14 plant species, and 1 type of subrepeats from 1 protozoan species (Table 1). We also used genomic data from five species, *Xenopus laevis* (National Bioresource Project, version 1.6.5; http://www.shigen.lab.nig.ac.jp/xenopus/), *Drosophila melanogaster* (UCSC genome database, version 2, April 2004), human (UCSC genome database, version 18, March 2006), mouse (UCSC genome database, version 8, March 2006), and rat (UCSC genome database, version 4, November 2004).

Table 1 List of species and rDNA IGS subrepeats used in this study

	Species	IGS size (GC%)	Subrepeats (GC%)	GenBank ID	Ref.
Mammalian	Human (*Homo sapiens*)	29,685 (52.1)*	798 (59.8),* 282 (57.8)	U13369	Gonzalez et al. (1992)
	Rat (*Rattus norvegicus*)	NTS: 2987 (47)*	150, 572†*	X03838	Tower et al. (1989); Yavachev et al. (1986)
	Mouse (*Mus musculus*)	31,905 (42.1)*	132*	BK000964	Kuehn and Arnheim (1983)
	Hamster (*Cricetulus longicaudatus*)	NA	356†	M26164†	Tower et al. (1989)
Amphibian	Tailed frog (*Ascaphus truei*)	1,897 (57.2)*	51 (47.1)S*	X12607	Morgan and Middleton (1988)
	Salamander (*Triturus vulgaris*)	8,516 (61.6)*	120 (57.5:59.2)*	X98876	De Lucchini et al. (1997)
	African clawed frog (*Xenopus laevis*)	NA*	100 (82),* 60 (74.6),*† 81 (76.5),* 35 (75.8)*	M23393	Moss et al. (1980); Pikaard and Reeder (1988)
	Kenyan clawed frog (*Xenopus borealis*)	Partial	138	X00184	Bach et al. (1981)
	Xenopus clivii	Partial	130	V01427	
Lower Cordata	Sea squirt (*Herdmania momus*)	1,880 (54.2)*	NA	X53538	Degnan et al. (1990)
Insect	*Drosophila melanogaster*	3,632 (29.0)*	100, 240†*	AF191295	Kohorn and Rae (1982); Ohnishi and Yamamoto (2004); Simeone et al. (1985)
	Drosophila oreana	2,385 (NA)	241	NA	Murtif and Rae (1985); Tautz et al. (1987)
	Drosophila hydeii	4,487 (NA)	226	NA	
	Drosophila virilis	5,276 (NA)	100 (40.0),* 226 (32.0)*	NA	
	Drosophila funebris	4,031 (30.0)*	NA	L17048	UP
	Bulldog ant (*Myrmecia croslandi*)	1,752 (34.3)*	144 (31.3)*	AB121789	Ohnishi and Yamamoto (2004)
	Asian tiger mosquito (*Aedes albopicus*)	4,707 (57.3)*	201 (58.2),* 64 (57.8),* 34 (41.2),* 48 (56.3)*	M65063	Baldridge and Fallon (1992)
	Yellow fever mosquito (*A. aegypti*)	1,797 (57.9)*	49 (63.3:35.0)	AF004986	Wu and Fallon (1998)
	Bombyx mori	Partial	3, 4, 5, 8	X05086	Fujiwara and Ishikawa (1987)
	Tsetse fly (*Glossina morisitans*)	Partial	420 (29.3)*	X05007	Cross and Dover (1987)
	Aphid (*Acyrthosiphon lactucae*)	Partial	247	NA	Kwon and Ishikawa (1992)
Crustacean	Swimming crab (*Charybdis japonica*)	5,376 (47.6)*	60 (45.0),* 142 (48.6), 391 (43.3)	NA	Ryu et al. (1999)
	Water flea (*Daphnia pulex*)	4,819 (45.4)*	330 (43.3), 200*	U34871	Crease (1993)
	Brine shrimp (*Artemia cysts*)	Partial	618 (42.0)*	NA	Koller et al. (1987)
Fungi	*Armillaria jezoensis*	611 (43.4)	NA	D89921	DS
	Armillaria ostoyae	590 (43.1)	NA	D89924	Peyretaillade et al. (1998)
	Armillaria sinapina	611 (42.4)	NA	D89925	
	Encephalitozoon cuniculi	1,748 (50.8)	NA	AJ005581	
Protozoan	*Trypanosoma cruzi*	1,754	172*	Y00055	Schnare and Gray (1982)

Table 1 continued

	Species	IGS size (GC%)	Subrepeats (GC%)	GenBank ID	Ref.
Plant	Rice (*Oryza sativa*)	2,140 (71.5)*	254 (70.4)*	X54194	Takaiwa et al. (1990)
	Wheat (*Triticum aestivum*)	3,589 (57.0)*	135 (61.0)*	X07841	Barker et al. (1988)
	Oat (*Avena sativa*)	3,982 (54.0)*	92 (46.0),* 148 (55.0)*	X74820	Polanco and Perez de la Vega (1994)
	Fava bean (*Vicia faba*)	2,014 (50.0)*	325 (50.0)*	X16615	Kato et al. (1984)
	Hirsuta bean (*Vicia hirsute*)	2,264 (48.0)*	379 (42.0)*	X62122	
	Adzuki bean (*Vigna angularis*)	Partial	174 (40.0)*	X17210	Unfried et al. (1991)
	Mung bean (*Vigna radiata*)	4,243 (50.0)*	174 (43.0),* 340 (57.0)*	X17209	Unfried et al. (1991)
	Carrot (*Daucus carota*)	5,775 (53.0)*	456 (56.0)*	D16103	Suzuki et al. (1996)
	Potato (*Solanum tuberosum*)	3,232 (56.5)*	54 (63.0),* 74 (57.1)*	X65489	Borisjuk and Hemleben (1993)
	Tomato (*Lycopersicon esculentum*)	3,253 (53.7)*	63 (64.8),* 141 (65.7)*	X14639	Schmidt-Puchta et al. (1989)
	Cucumber (*Cucumis sativus*)	3,451 (61.3)	30 (90.0), 90 (55.6)*	X07991	Ganal et al. (1988)
	Squash (*Cucurbita maxima*)	5,508 (58.3)*	104 (72.4),* 118 (46.1)* 420 (47.7), 41 (90.1)	X13059	UP
	Strawberry (*Fragaria ananassa*)	4,274 (52.0)*	140 (54.0),* 170 (52.0)	X58119	UP
	Tobacco (*Nicotiana tabacum*)	4,996 (56.0)*	216 (68.0), 121 (58.0)*	D76443	UP
	Arabidopsis thaliana	4,726 (49.0)*	495 (58.0)*	X15550	Gruendler et al. (1989)
	Kidney bean (*Phaseolus vulgaris*)	Partial	166 (40.0)*	Z48777	DS
	Garden rocket (*Eruca sativa*)	4,003 (45.0)*	113 (56.0), 120 (26.0)*	X74829	Lakshmikumaran and Negi (1994)

Note. UP, unpublished data; DS, directly submitted to GenBank; NA, not available. *rDNA IGSs and subrepeats used in this study. †Experimentally tested as an enhancer. SSelected subrepeat with the most similarity value compared to others

Calculating the Frequency and the Fraction of Poly(N) Runs

The probability of finding a poly(N) run of a certain length l depends on the length of the run and the probabilities of having a different nucleotide at positions adjacent to either ends of the run.

$$\Pr[\text{poly(N)}_{i=begin,l,x}] = \Pr[\text{N}_{\text{Prev}}] * \Pr[\text{N}_{\text{Given}}]^{l} * \Pr[\text{N}_{\text{Next}}] \quad (1)$$

where the terms $\Pr[\text{N}_{\text{Prev}}]$, $\Pr[\text{N}_{\text{Given}}]$, $\Pr[\text{N}_{\text{Next}}]$, and l denote, respectively, the probability of a different nucleotide at the 5′ adjacent site, the probability of a given nucleotide at the site of interest, the probability of a different nucleotide at the 3′ adjacent site, and the length of run. Based on Eq. 1, we derived equations to calculate the expected frequency and fraction of poly(N) in a certain length of sequence (supplementary). To obtain a standard value for the frequency and the fraction of poly(N) runs, we assumed that all four nucleotides are equiprobable. If $P_{AT} = P_{CG} = 1/4$, then the expected frequency of poly(N) runs is equal to $4\ (1/4)^2\ (1 - 1/4) = 4\ (1/16)\ (3/4) = 3/16$. Also, the expected fraction of poly(N) runs is equal to $4\ (1/4)^2\ (2 - 1/4) = 4\ (1/16)\ (7/4) = 7/16$ (43.75%).

If we "drop out" the repeated nucleotides except one in each poly(N) run, the estimated decrease in the total length of the sequence with all nucleotides occurring with equal probability of 1/4 is

$$\begin{aligned} &\Pr_{\text{fraction}}\ [[\text{poly(N)}_{i=begin,l\geq 2,x}] \\ &\quad - \Pr_{\text{number}}\ [[\text{poly(N)}_{i=begin,l\geq 2,x}] \\ &= (7/16)\ -\ (3/16) = (1/4) \end{aligned} \quad (2)$$

Thus, theoretically for a given sequence, its total length will decrease by 25% if all but one of the repeated nucleotides in the poly(N) runs are dropped out.

Random Sequence Testing and Statistical Analysis

Since the dropout method compresses runs of nucleotides in two sequences into instances of single nucleotide occurrences before alignment, and since the resulting compression typically reduces the length of the sequences by 25% on average, one may suspect that those length differences may introduce spurious similarities by chance. Thus one must ensure that these observed similarities determined by the dropout alignment are indeed statistically significant. A standard approach involves creating a null model from which random pairs of sequences can be drawn and examined for chance alignment, which can then be used to calibrate the observed alignments. Based on this framework, we estimated the accuracy of the dropout method in subrepeat as follows: we randomly selected sequences (with same length as IGS sequences) with the subrepeats from a full genome database in the same species and compared them with the subrepeats used in this study. For the statistical test, we repeated the comparisons 10,000 times, involving independently and randomly drawn sequences, whose total length adds up to about 1% fraction of their full genomes. All random selections were determined by a random-number-generation function in LISP programming language (Lispworks version 4.2.0, Xanalys Inc). In comparing alignments of the IGS sequences by the dropout method against the distributions observed from the null model, the Student's t test was used; a stringent p-value of $p < 0.01$ was considered to be statistically significant.

We also used a somewhat different null model based on random shuffling of the sequences in addition to the earlier methods based on random sequences from unrelated regions of the genome. Since randomly selected sequences from their genomes might have slightly different base composition than the IGS subrepeats, it could be argued that this provides a relatively uninformative prior. Therefore, we randomly shuffled IGS subrepeat sequences so that the sequences have only random base order, but not a different base composition. We then estimated statistical significance of the alignments based on dropout method before and after shuffling.

Sequence Alignment and the Secondary Structure Prediction

Sequence alignments were carried out using the global alignment algorithm developed by Needleman and Wunsch (1970) with matching score 2, mismatching score –1, and gap penalty −2. For multiple sequence comparisons, we used the Clustal W (Higgins et al. 1992) method in MegAlign program (version 5.07; DNASTAR Inc.) with gap penalty 15 and gap length penalty 6. We also used WinDotter (downloaded from the web site http://www.cgb.ki.se/cgb/groups/sonnhammer/Dotter.html), a dot-matrix program developed by Sonnhammer and Durbin (1995), with the default window width of 25 residues and score threshold 35. Some portions of the alignments were edited by hand to further improve similarity. To predict the secondary structure of the IGS subrepeat sequences, we used MFOLD (Zuker 2003).

Results

Patterns of Poly(N) in Subrepeats of the IGS Region

Poly(N) runs up to several nucleotides long appear frequently in the subrepeats of the rDNA IGS region. We

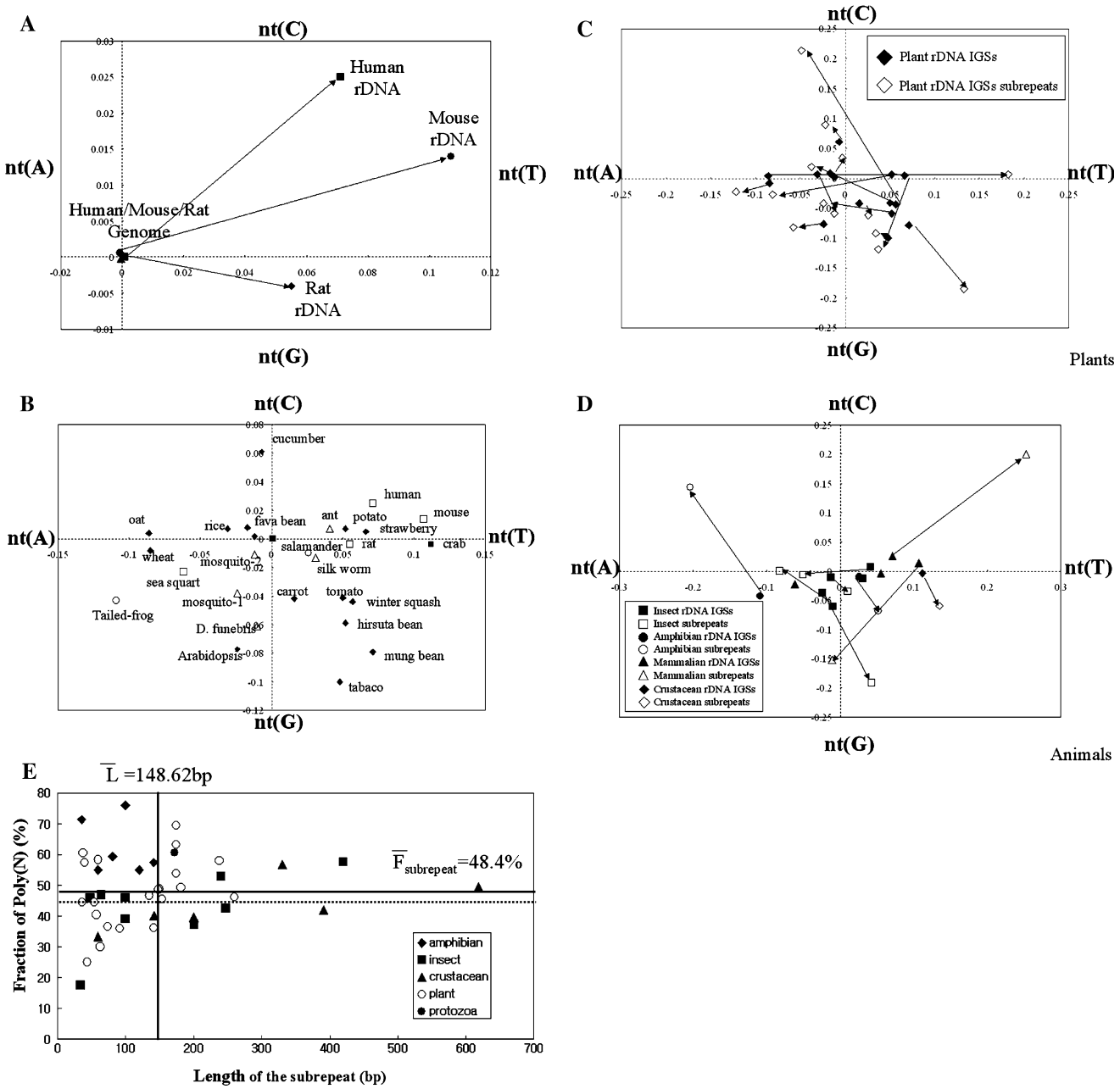

Fig. 1 Biased base composition of the rDNA IGS and subrepeats. (**A**) Base composition of mammalian rDNA IGS compared to mammalian genomic sequences. Arrows indicate shifting of biased base compositions from genomic DNA to rDNA IGS. (**B**) Base composition of full rDNA IGS sequences in many species. Comparison of the base compositions between rDNA IGS and subrepeats in animals (**C**) and plants (**D**). (**E**) Relationship between poly (N) and lengths of subrepeats. Dashed line indicates the expected content of total poly(N), which is computed to be 43.75%. Black line indicates the mean values of subrepeat lengths ($\bar{L}$)

hypothesized that this high frequency of the poly(N) runs was a major factor in the divergence of primary sequence and the length of subrepeats of the rDNA IGS within a species. We collected for analysis 28 rDNA IGS sequences and 44 subrepeats from 28 species including mammals, insects, crustaceans, amphibians, and plants. First, we tested for and detected a significant bias in base composition in almost all rDNA IGSs in comparison to genomic sequences (Figs. 1A and B). The degrees to which they exhibited these biases in base composition were dramatically increased in almost all subrepeats from both animals and plants (Figs. 1C and D). To visualize these biases, we calculated the fractions of the rDNA IGS comprising poly(N) and plotted them.

The percentage of the poly(N) runs in most rDNA IGS subrepeats ranges variously from 17.6% to 76% even

though their average (48.4% ± 12% [SD]) is a little higher than the expected percentage of poly(N) runs, 43.75% (Fig. 1E). We also investigated the possible relationship between sequence (subrepeat) length and percentage of poly(N) runs. The average length of the subrepeats was 148.62 bp. Plotting the relationship between poly(N) percentage and lengths of the subrepeats showed that the different types of subrepeats from the same species have similar poly(N) percentages, for example, 64- and 48-bp subrepeats from mosquito have 46.9% and 45.8% poly(N), respectively.

We also analyzed the base patterns of the poly(N) runs in several species in which rDNA IGS has been well characterized (Fig. 2). In this study, we found three different patterns. The 141-bp (60/81-bp) subrepeat of Xenopus (Moss et al. 1980) showed the first pattern, or a high frequency of specifically poly(G/C) runs (Fig. 2A). The 100-bp subrepeat of Xenopus (Moss et al. 1980), the 150-bp subrepeat of rice (Takaiwa et al. 1990), and the 64-bp and 201-bp subrepeats of mosquito (Baldridge and Fallon 1992) also have a high frequency of poly(G/C). Another pattern, a high frequency of poly(A/T), is apparent in the 420-bp subrepeat of the tsetse fly (Cross and Dover 1987) (Fig. 2B). This pattern is also found in the 240-bp subrepeat of D. melanogaster (Simeone et al. 1985), and the 330-bp subrepeat of water flea (Crease 1993) (data not shown). Third, there is a bias toward a single nucleotide occurring in poly(N) runs. For example, the 142- and 390-bp subrepeats of the swimming crab (Ryu et al. 1999) have numerous poly(T)s (Fig. 2C); 'TT' is found 4 times in the 142-bp subrepeats, 'TTT' is found about 3 times, and 'TTTT' as well as 'TTTTT' is found just once, whereas 'AA' or 'GG' is found just once, 'CC' is found 10 times, and 'CCC' is found just once. Therefore, crab subrepeats show primarily the poly(T) reiterating pattern. Similarly, the 141-bp subrepeat of tomato (Schmidt-Puchta et al. 1989) and the 247-bp subrepeat of pea aphid show predominantly poly(G) reiterating patterns (Kwon and Ishikawa 1992) (data not shown). Finally, we noted that there are also examples of mixed patterns. The 618-bp subrepeat of Artemia and the 238-bp subrepeat of carrot show minor poly(A) reiterating patterns as well as reiterating patterns of all other nucleotides (Koller et al. 1987; Suzuki et al. 1996) (Fig. 2D).

The Dropout Method and Its Efficacy in Revealing Similarity Among Different Types of Subrepeats from a Species

Because many differences between sequences involve poly(N) expansions, we reasoned that eliminating poly(N) runs would reveal similarities that might have been masked by poly(N) runs. Thus, prior to aligning sequences, we "dropped out" (deleted) all consecutive bases in each poly(N) run except one base.

The majority of the rDNA IGS region of *Xenopus* is composed of four subrepeats, 35, 60, 81, and 100 bp long. The 60- and 81-bp subrepeats are known as enhancer for RNA polymerase I machinery (Pikaard and Reeder 1988). We found that the percentages of poly(N) runs in the 35-bp subrepeat (74.5%) and the 100-bp subrepeat (76%) are higher than the percentages of the 60-bp and the 81-bp subrepeats, 55% and 59.3%, respectively. When we applied the dropout method, we found that the three types of

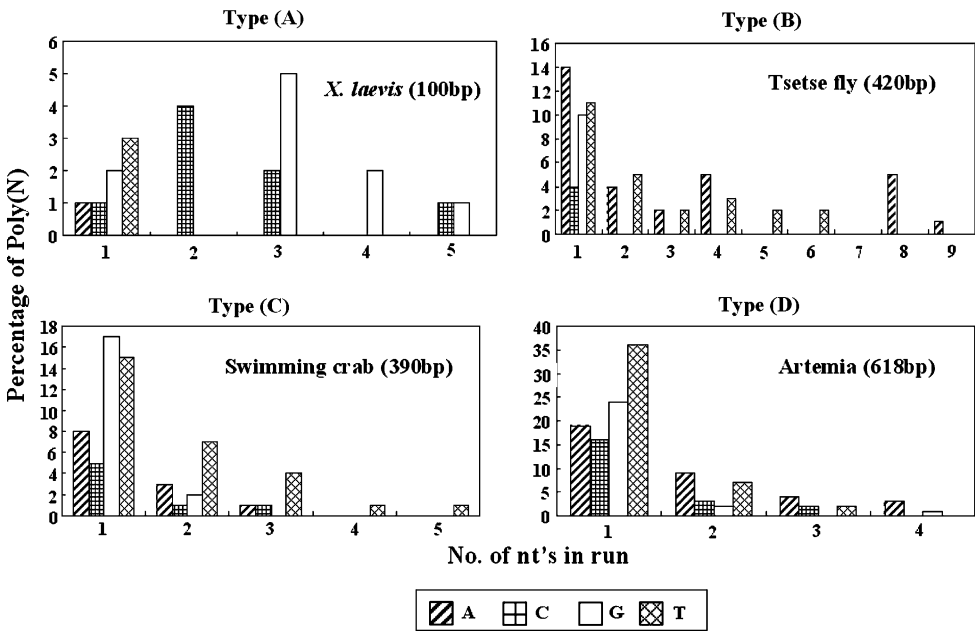

Fig. 2 Comparison between the poly(N) base composition of subrepeats changes according to its frequency. Species name and its subrepeat are marked at the top of each graph. The *x*-axis indicates the number of nucleotides repeated and the *y*-axis indicates the frequency. Type A indicates poly(C and G) reiterating pattern. Type B indicates poly(A and T) reiteration. Type C indicates single poly(N) reiterating pattern. Type D indicates mixed patterns.

subrepeats, 60, 81, and 100 bp, were more easily aligned, thus revealing possible shared ancestry (Fig. 3B). We used the Clustal W (Higgins et al. 1992) method in the MegAlign program (version 5.07; DNASTAR Inc) with gap penalty 15 and gap length penalty 6. After dropping out the poly(N) runs, the similarity value between the 35-bp and the 60-bp subrepeats was 41.9% (85.3% increased compared to 22.6% similarity before dropping out poly[N]'s) and the similarity value between the 81-bp and the 100-bp subrepeats was 58.6% (132.5% increase compared to 25.2% similarity before dropping out poly[N]'s) in the Clustal W method with the same condition. We also discovered three conserved regions in the secondary structures: S1 and S2 for the stem region and L1 for the loop region. Both S1 and S2 regions are perfectly complementary (Fig. 3C). In order to exercise caution, lest increased similarity might be an artifact resulting from a decrease in overall length due to the dropout method (as opposed to revealed similarity), we examined the efficiency of the dropout method using the genomic data. First, we randomly selected the 100-bp sequence from the *Xenopus* genome and measured the similarity value with 81-bp rDNA IGS subrepeat. Next, we dropped out poly(N) runs and remeasured similarity value. After that, we compared two similarity values. We repeated this random selection 10,000 times. The average percentage of the poly(N) before dropping out any poly(N) runs was 48.8% ± 0.08% (average ± standard error), the similarity between a sequence pair increased by about 5.46 ± 0.05 percentage points, from 28.08% ± 0.038% to 33.54% ± 0.042% after the poly(N) runs were dropped out. We found that increased similarity values between the 81-bp and the 100-bp *Xenopus* rDNA subrepeats were statistically significant ($p < 0.001$, Student's t test). We also compared these similarity values with those of the shuffled sequences and found that the increased similarity values were not merely due to decreased length resulting from dropout ($p < 0.001$).

We also tested the dropout method in other species, *Drosophila melanogaster*, rDNA IGS subrepeats in order to confirm possible relationship among different types of subrepeats from one species. *D. melanogaster* has two subrepeats, 100 and 240 bp, and there exists no reported possible relationship between the two subrepeats, primarily because they fail to align in a statistically significant manner by the existing alignment algorithms. The 240-bp subrepeats are known to be an enhancer for RNA polymerase I machinery (Kohorn and Rae 1982). However, these sequences have many poly(A) and poly(T) runs that obscure the deeper biological signals. By using the dropout method,

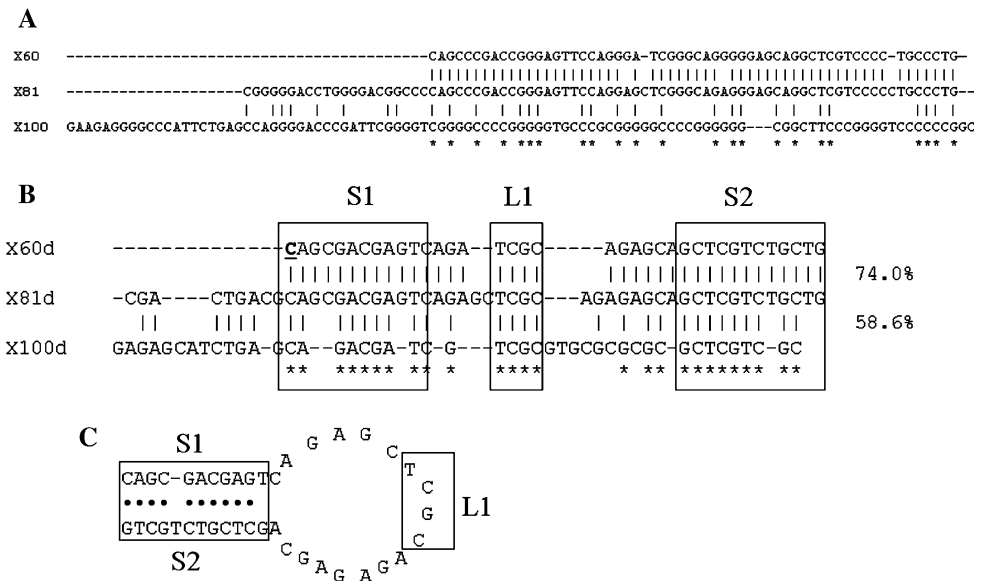

Fig. 3 Sequence alignment among three subrepeats of the *Xenopus* IGS region. At left margin, labels indicate lengths; i.e., X60, X81, and X100 indicate 60-, 81-, and 100-bp subrepeats, respectively. Horizontal lines between bases represent gaps, which have been included for maximum alignment; the vertical bars indicate identical nucleotides in alignment. (**A**) Multiple alignment among unmodified subrepeat sequences. (**B**) Subrepeats are modified by the dropout method, which deletes poly(N) runs except one base in the run. X60d, X81d, and X100d indicate the 60-, 81-, and 100-bp subrepeats, respectively, after dropping out poly(N) runs. Numbers (%) at the right margin indicate similarity values between the 60- and the 81-bp subrepeats and between the 81- and the 100-bp subrepeats. Asterisks indicate the matched nucleotides between 81- and 100-bp subrepeats. Three conserved regions are marked S1 (stem 1), S2 (stem 2), and L1 (loop 1). (**C**) Projected secondary structure of the 60-bp subrepeat after dropping out poly(N). MFOLD was used to predict the secondary structure of the IGS subrepeat. Alignments were generated by Clustal W in MegAlign (version 5.07; DNASTAR Inc.) with gap initiation penalty of 15 and gap extension penalty of 6

which relieved the alignment ambiguities due to poly(N)'s, we discovered regions with true high similarity between *Drosophila* 100-bp and *Drosophila* 240-bp subrepeats (Fig. 4). Conserved regions with 75.9% similarity value are marked by box D in lower panels in Fig. 4. To calibrate the significance of this similarity value, we computed a *p*-value by comparing it with a null model created by randomly selecting unrelated sequences from *D. melanogaster* genomic sequences. We repeated this random selection process 10,000 times by drawing the sequences independently but with replacement. The average poly(N) percentage of such randomly selected sequences before dropping out any poly(N) runs was 48.36% ± 0.086% (average ± standard error); the similarity between a sequence pair increased by about 1.54 ± 0.023 percentage points, from 16.73% ± 0.018% to 20.24% ± 0.024%, after the poly(N) runs were dropped out. The Student *t*-test showed that increased similarity values between two *Drosophila* rDNA subrepeats, 100 and 240 bp, were statistically much more significant ($p < 0.001$). We also compared these values with those obtained by shuffling the sequences and found that increased similarity values were not due merely to length variation resulting from dropout itself ($p < 0.001$).

In order to ascertain the biological universality of the underlying mechanisms, we also applied the dropout method to similar regions in one plant species, namely, the tomato *Lycopersicon esculentum*. The rDNA IGS of tomato is composed of two subrepeat types, 63-bp (RE I)- and 141-bp (RE II)-long sequence (Schmidt-Puchta et al. 1989). The poly(N) percentage of two subrepeat types, the 63-bp and 141-bp subrepeats of the tomato rDNA IGS is 30% and 36.1%, respectively (Fig. 1). In particular, the 141-bp subrepeats have a high percentage of guanine, which occurs five times as triplets (GGG) and eight times as dinucleotides (GG). The alignment of "dropped-out" 63- and 141-bp subrepeats resulted in an increase in similarity by 52.7% to 33%, compared to 17.4% similarity of the original sequences (Fig. 5). Most importantly, the tomato 63-bp subrepeat showed a high similarity value with a 5′-portion of 141-bp subrepeat, which we shall refer to as a Box T. The similarity value in this box was 60.6%, a much higher percentage than would be expected by chance alignment. We also tested the dropout method in another plant species, potato, *Solanum tuberosum*. Potato rDNA IGS contains two types of subrepeats, or 54-bp (type II) and 74-bp (type I) subrepeats (Borisjuk and Hemleben 1993). The similarity value between them using the dropout method is 60.7% (data not shown), and the analysis uncovered many of the same features as in the other unrelated species.

Fig. 4 Sequence alignment between two subrepeats of the *D. melanogaster* IGS region. At the left margin, two subrepeats are indicated by their lengths, 100 and 240 bp, and labels indicate sequence lengths; i.e., D100 and D240 indicate 100- and 240-bp subrepeats, respectively. Upper panel: alignment of unmodified subrepeat sequences. Lower panel: subrepeat sequences modified by the dropout method aligned with gaps. D100d and D240d indicate 100- and 240-bp subrepeats, respectively, after dropping out poly(N) runs. Conserved region is marked by box D in the panel. Numbers (%) at the right margin indicate similarity between 100- and 240-bp subrepeats. Sequence alignments were carried out using the Needleman-Wunsch global alignment algorithm applied with a match score of 2, a mismatch score of −1, and a gap penalty of −2

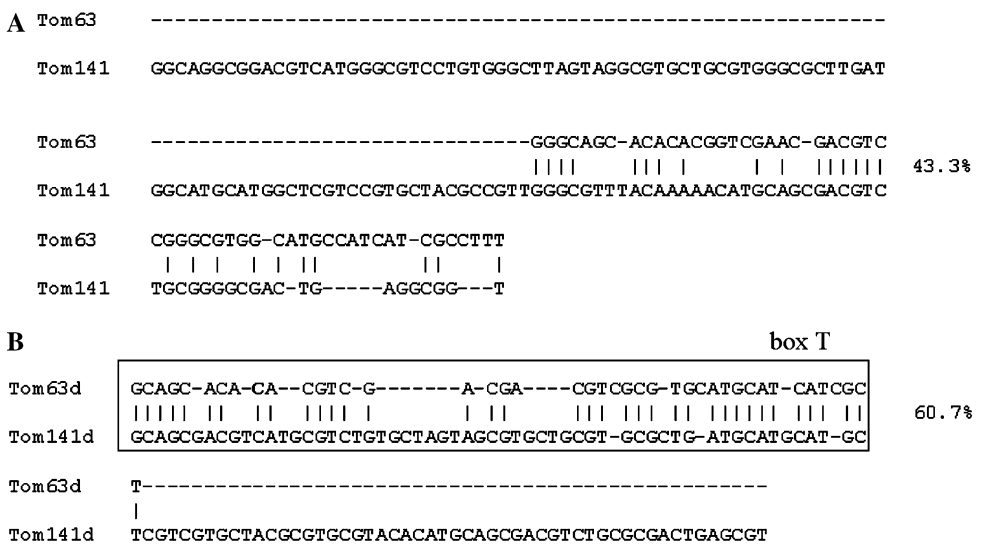

Fig. 5 Sequence alignment between the 63- and the 141-bp subrepeats of tomato. Tom63 and Tom141 indicate 63- and 141-bp subrepeats, respectively. Upper panel: alignment of unmodified subrepeats sequences. Lower panel: subrepeats modified by the dropout method with poly(N) deleted except one base. Tom63d and Tom141d indicate 63- and 141-bp subrepeats, respectively, after dropping out poly(N) runs. The inverted triangle indicates the position of the first nucleotide in the lower panel and the corresponding position in the upper panel. Highly conserved regions after dropping-out are marked by "box T" (62.3% similarity within the box T region). Sequence alignments were carried out using the Needleman-Wunsch global alignment algorithm applied with a match score of 2, a mismatch score of –1, and a gap penalty of –2

Use of the Dropout Alignment Method to Compare Subrepeats Between Different Species

The dropout method can also be used to compare different subrepeats from different species. First, we tested the dropout effect on closely related species. We selected several subrepeats from plants because the sizes of plant subrepeats are relatively long and similar to each other. We observed that similarity values were increased from 33.9 ± 0.83 to 56.0 ± 1.98 (mean ± standard error) by removing poly(N) runs (Table 2). Although mean values are not consistent with evolutionary distance among plant species, we confirmed the usefulness of the dropout method in finding possible relations among subrepeats. For example, the similarity value of subrepeats between two bean species from the same genus, *Vicia*, was extremely low and unnoticeable (35.8%) before dropping out poly(N) runs. However, after dropping out poly(N) runs, the similarity value was found to be relatively high (66.7%) compared with others.

We also tested the usefulness of the dropout method in short rDNA IGS subrepeats. The similarity value of subrepeats between two *Drosophila* species, 100-bp subrepeats of *D. melanogaster* and *D. virilis*, was increased from 36.8% to 58.5% by shrinking poly(N) runs. Moreover, the similarity value in the high-similarity region, BoxD100 (69.4%), was much higher than overall. We observed that many species have similar lengths of subrepeats. For example, many species have ~60-bp lengths: *Xenopus* (60 bp), swimming crab (60 bp), tomato (63 bp), mosquito (64 bp), and potato (54 bp). The similarity value obtained from the comparison of *Xenopus* 60-bp subrepeat with swimming crab 60-bp subrepeat was 46.0%. The 54-bp

Table 2 Comparison of similarity values among long rDNA IGS subrepeats from the plant species before and after dropping out poly(N) runs

	Arab295	BVF325	bVH379	bVR340	Car456	Rice254
Arab295	–	36.2 → 58.1	30.3 → 54.4	33.3 → 56.3	31.8* → **76.6***	37.9* → 49.8*
bVF325		–	35.8 → **66.7**	32.2 → 54.5	30.7* → 56.7*	27.5 → 44.5*
bVH379			–	33.7 → 53.6	35.4*→ 56.0*	31.8* → 50.2*
bVR340				–	35.9* → 53.4*	37.8* → 53.7*
Car456					–	38.6* → 55.7*
Rice254						–

Note. Data are similarity values (%) between two subrepeats. Numbers on the left and right side of the arrows indicate similarity values before and after dropping out poly(N) runs, respectively Sequence labels include their lengths: Arab295 (*Arabidopsis*, 295 bp), bVF325 (*Vicia faba*, 325 bp), bVH379 (*Vicia hirsute*, 379 bp), bVR340 (*Vigna radiate*, 174 bp), Car456 (carrot, 456 bp), and Rice254 (rice, 254 bp) rDNA IGS subrepeat. *Subrepeat excluding long gaps generated by size differences at either the 5′ or the 3′ end

subrepeats of potato and the 63-bp subrepeats of tomato also demonstrated similarity. Figure 6 represented the dropout alignment among the 60-bp subrepeats of *Xenopus*, the 54-bp subrepeat from the potato IGS, and the 63-bp subrepeat from the tomato rDNA IGS (Fig. 6D). These alignments showed that the similarity between potato 54-bp and tomato 63-bp subrepeats was increased from 76.8% to 86.4%. The similarity between tomato 63-bp and *Xenopus* 60-bp subrepeats was increased from 29.7% to 59.6%.

Discovery of these conserved nucleotides among different subrepeats from the various species, motivated us to explore similar features in many subrepeats from various species. In this alignment, we used 23 types of relatively short subrepeats from 15 species; 11 subrepeats from 7 plant species, 6 subrepeats from 3 insect species, 2 subrepeats from 2 crustacean species, 3 subrepeats from 2 amphibian species, and 1 subrepeat from a nematode. The multiple alignments for all these subrepeats are shown in Fig. 7. In certain cases, to compensate for length variation, we used a part of these subrepeats: either the 5′ or the 3′ end. We also checked alignments using reverse-complementary sequences to consider a possible gene inversion. We marked three regions, S1 (stem 1), S2 (stem 2), and L1 (loop 1), based on the secondary structure of the *Xenopus* 60-bp subrepeat before comparing them. From this study, we found that most subrepeats shared commonly conserved arrangements of sequences in all three regions.

Use of the Dropout Alignment Method to Find Novel Conserved Sequences Between Species

We also applied the dropout method to whole rDNA IGS sequences. First, we chose two *Drosophila* species, the full rDNA IGS sequences from *D. melanogaster* (4394 bp) and *D. funebris* (4031 bp), aligned with a Dot-Plot matrix program (Sonnhammer and Durbin 1995) (Fig. 8). The Dot-Plot matrix program is useful to find gene duplications or inversions between sequences. After dropping out poly(N) runs, the lengths of the rDNA IGS sequences from *D. melanogaster* and *D. funebris* decreased to 3165 bp (28% reduction in length) and 3654 bp (9.3% reduction in length), respectively. We found that the Dot-Plot matrix revealed many relatively long matched sequences after dropping out poly(N) (Fig. 8B). The clear grid-arrayed diagonals indicate that the rDNA IGS is composed of tandemly reiterated subrepeats that are shared between the species. Although many short diagonal patterns appeared in the dot-plot matrix with original sequences (Fig. 8A), they did not extend sufficiently to ascertain detectable similarity between the sequences. We also applied the dropout method

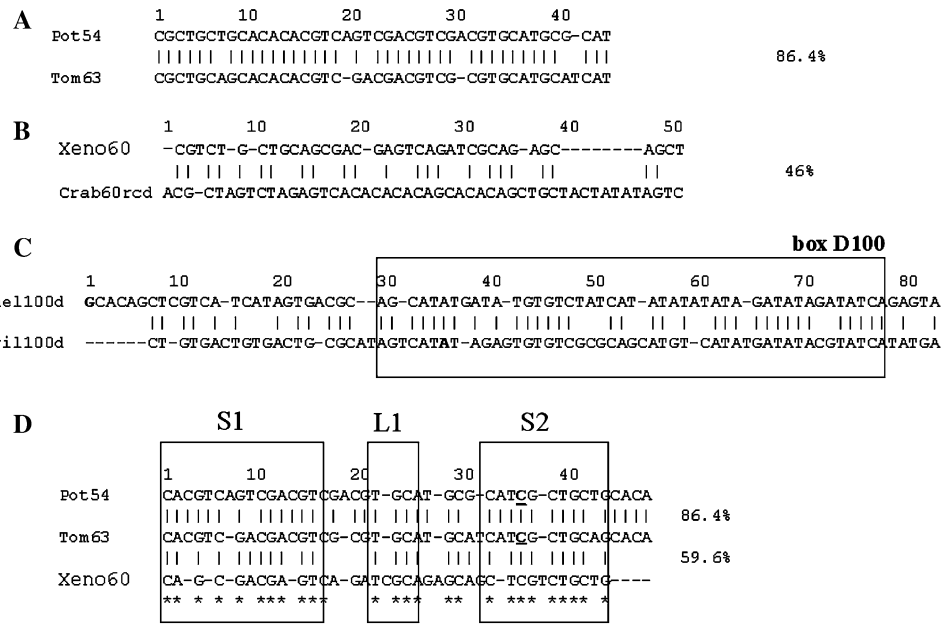

Fig. 6 Multiple alignments among different subrepeats from various species after dropping out poly(N)'s: alignments (**A**) between the 54-bp subrepeat of potato and the 63-bp subrepeat of tomato; (**B**) between the 60-bp subrepeat of *Xenopus* and the 60-bp subrepeat of swimming crab; (**C**) between the 100-bp subrepeat of *D. melanogaster* and the 100-bp subrepeat of *D. virilis*; (**D**) among the 54-bp subrepeat of potato, the 63-bp subrepeat of tomato, and the 60-bp subrepeat of *Xenopus*. Labels refer to the following subrepeats after dropping out poly(N) runs: Pot54 (potato, 54 bp), Tom63 (tomato, 63 bp), Xeno60 (*Xenopus*, 60 bp), Crab60rcd (crab, 60 bp reverse-complemented), Dmel100d (*D. melanogaster*, 100 bp), and Dvil100d (*D. virilis*, 100 bp). Vertical bars indicate identical nucleotides. Highly conserved regions in the alignment are marked by a box and labeled in the same manner as in Fig. 3. At the right margin, pairwise similarities (%) are indicated. Sequence alignments were carried out using the Needleman-Wunsch global alignment algorithm applied with a match score of 2, a mismatch score of −1, and a gap penalty of −2

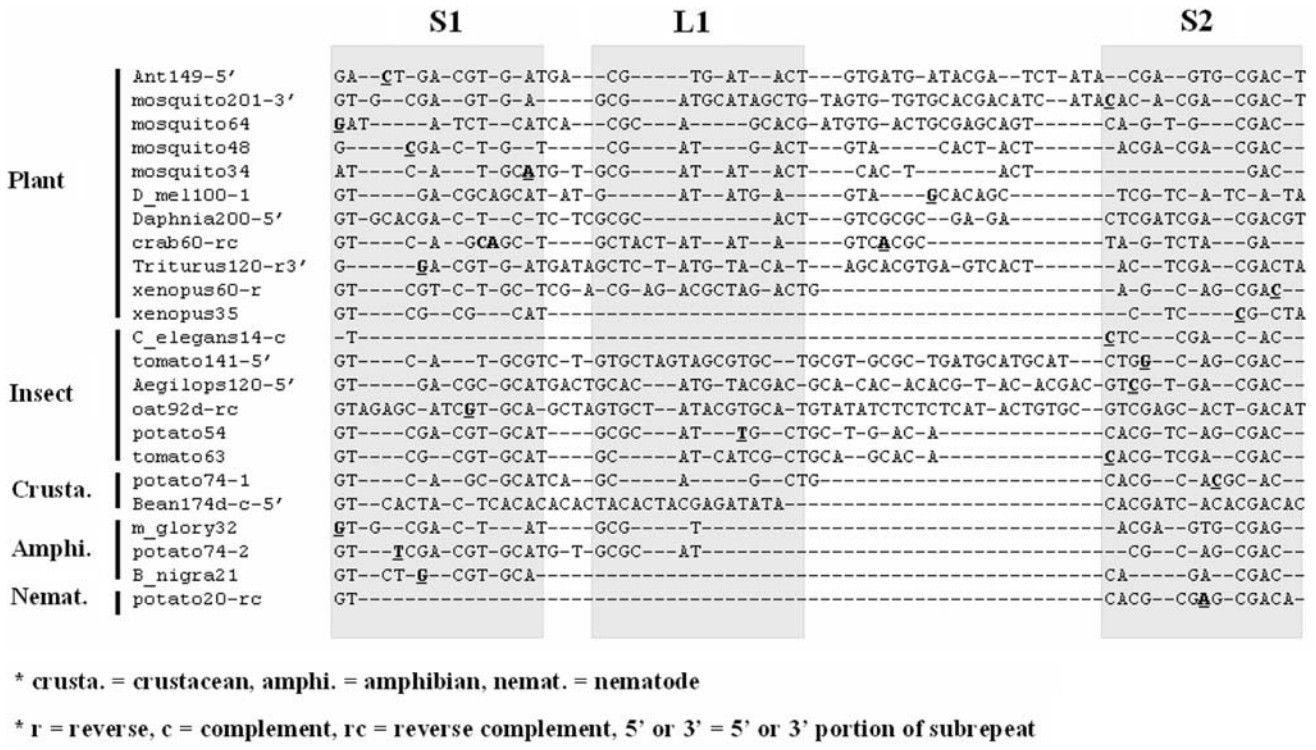

Fig. 7 Multiple alignments among many subrepeats from various species. Total 23 types of subrepeats from 15 species; 11 subrepeats from 7 plant species, 6 subrepeats from 3 insect species, 2 subrepeats from 2 crustacean species, 3 subrepeats from 2 amphibian species, and 1 subrepeats from 1 nematode species. All or portion of subrepeats, starting from either 5′ or 3′ end, were used. Labels refer to the following subrepeats after dropping out poly(N) runs: tomato141 (tomato, 141 bp), Aegilops120 (*A. umbellulata*, 120 bp), Bean174d (adzuki bean, 174 bp), oat92d (oat, 92 bp), potato54 (potato, 54 bp), tomato63 (tomato, 63 bp), potato74 (potato, 74 bp), m_glory32 (morning glory, 32 bp), B_nigra21 (mustard, 21 bp), potato20 (potato, 20 bp), Ant149 (ant, 149 bp), mosquito64 (mosquito, 64 bp), mosquito48 (mosquito, 48 bp), mosquito34 (mosquito, 34 bp), D_mel100 (*D. melanogaster*, 100 bp), Daphnia200 (*D. pulex*, 200 bp), crab60 (crab, 60 bp), Triturus120 (salamander, 120 bp), xenopus60 (*Xenopus*, 60 bp), xenopus35 (*Xenopus*, 35 bp) and C-elegans14 (*C. elegans*,14 bp). In certain cases, we have used suitably oriented [e.g., regular, reversed (r), complementary (c), and reverse-complementary (rc)] sequences to consider a possible gene inversion or duplication. 5′ or 3′ indicates 5′ or 3′ portion of subrepeat. The gray boxes marked by S1, S2, and L1 indicate the sequences that have the conserved secondary structure of the *Xenopus* 60-bp subrepeat. Alignments were generated by the ClustalW in the MegAlign (version 5.07, DNASTAR Inc.) with a gap initiation penalty of 15 and a gap extension penalty of 6

to find matching sequences by using the BLAST searching program (Windows-based BLAST program is available on the NCBI web site). We selected 100-bp subrepeats from *D. melanogaster* and searched possible matching sequences in full rDNA IGS sequences from *D. funebris*. After dropping out poly(N) runs, we obtained significantly extended matching sequences (Fig. 8D). Based on these matching sequences, we reconstructed possible 100-bp *D. funebris* rDNA IGS subrepeat. The similarity value between 100-bp *D. melanogaster* and *D. funebris* rDNA IGS 100-bp subrepeats was relatively high, 71.4% (Fig. 8E).

Discussion

The IGS of many species is known to contain many tandemly reiterating subrepeats. Different copy numbers of repeating elements (or subrepeats) account for most of the length variations of rDNA IGSs among closely related species (King et al. 1993; Tautz et al. 1987). Furthermore, the sequences of the subrepeats themselves differ across species. The analysis described in this paper suggests that most of this variation occurs by a repetition at the nucleotide level, manifested by the occurrence of runs of the same base, or poly(N) runs. These heterogeneities in size and sequence of subrepeats have made it difficult to compare them directly and to discover common motifs, which may have been conserved during evolution. Nevertheless, the subrepeats may be important in that they are likely to have transcriptional enhancers and promoters for the RNA polymerase I machinery (Labhart and Reeder 1984; Reeder 1990).

One of the most important characteristics of rDNA IGS sequences, or occurrence of many short poly(N)'s, is not unique to rDNA IGS subrepeats. In fact, it is a universal character that can be expected theoretically in any given

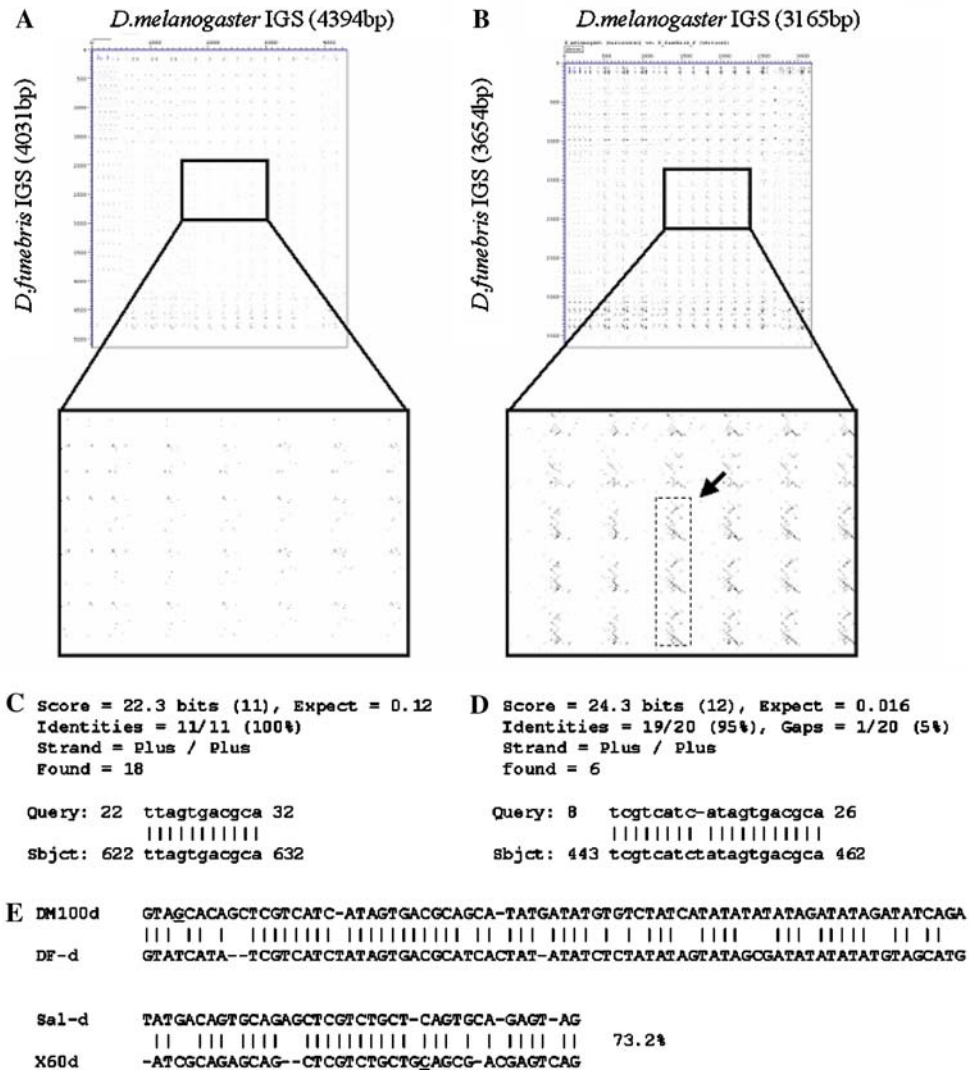

Fig. 8 Dot-plot comparisons of whole rDNA IGS sequences from two *Drosophila* species, *D. melanogaster* (4394 bp) and *D. funebris* (4031 bp). The internal repeat patterns were determined by comparing each sequence with itself in a WinDotter (a dot-matrix program developed by Sonnhammer and Durbin [1996]) with the default window width of 25 residues, score threshold of 40, and stringency of 10 nucleotides perfect match; the same parameters were used for all comparisons. Left panel: the alignment between two unmodified IGS sequences (**A**). Right panel: the alignment between two IGS sequences (**B**) after dropping out poly(N). The small boxes show a magnified window from A and B, respectively. Arrow within the rectangular region indicates grid-arrayed matching sequences between two species. Blast search: query, 100-bp *D. melanogaster* rDNA IGS subrepeats; target, full rDNA IGS from the *D. funebris* before (**C**) and after (**D**) dropping out poly(N) runs (gap opening, 5; gap extension, 2). (**E**)Reconstructed alignment between the 100-bp subrepeat of *D. melanogaster* (Dm100d) and the possible subrepeat of *D. funebris* (Df-d) (top) and between the 60-bp subrepeat of *Xenopus* (X60d) and the possible subrepeat of salamander (Sal-d) (bottom). Sequence alignments were carried out using the Needleman-Wunsch global alignment algorithm applied with a match score of 2, a mismatch score of −1, and a gap penalty of −2

noncoding spacer region or gene as well as in any randomly selected sequence, as estimated by Eqs. 1–3. However, it is important to point out that rDNA IGS subrepeats have certain specific patterns of short poly(N) runs in some species. These different patterns lead us to believe that these different poly(N) run patterns might drive the subrepeats to various evolutionary pathways, resulting in diversity both in size and in the primary structure. In comparing poly(N) percentage and length among various subrepeats, we did not find any specific range of the percentage of poly(N) for certain type of taxa. Although we found that two amphibian species have an unusually higher percentage of poly(N) compared to other taxa, we still need more genomic data from other amphibian and related nonamphibian species to conclusively validate this property as a characteristic of amphibians.

Interestingly, we found that many species have their own specific reiterating pattern of poly(N) runs. The most distinct patterns involve reiterations of guanine/cytosine (G/C) or adenine/thymine (A/T). Moreover, they often

share similar frequencies of poly(G)'s and poly(C)'s of the same length: for instance, five incidences of CCC would often match with five incidences of GGG. Such patterns could correlate with base pairing in "stems" during the formation of a hypothetical secondary structure or could be a result of the same mutational drive on opposite strands. However, some species show an unequal number of poly(G/C) or poly(A/T). Also, some species reiterate predominantly one base. Two hypotheses could explain this phenomenon. First, most polynucleotides in these specific reiterations might be located in loop regions of a presumed secondary structure. Loop regions are more often variable than stem regions. Such variation could be accumulated in regions that are not subject to strong purifying selection. The second hypothesis postulates the addition of the same nucleotide by an active mechanism, most likely a process such as "slippage during replication" (Tautz et al. 1986).

We also used a dropout alignment method to reveal similarities among subrepeats within certain species. We extensively studied the *Xenopus* IGS, which is composed of four subrepeats. Although the 60- and the 81-bp subrepeats are arranged in an alternating pattern and are highly similar to each other (Moss et al. 1980), the other subrepeats, 35 and 100 bp, are not (Fig. 3A). The *Xenopus* subrepeats were discovered to have an unusually higher percentage of poly(N). This high poly(N) content at different locations contributes significantly to the differences between these subrepeats, as is clearly revealed when the dropout method is applied (Fig. 3B) to align them. The similarity value after dropping out the poly(N) from two *Xenopus* subrepeats, 81 and 100 bp, was significantly increased and revealed apparent similarity among all four subrepeat types. We also found high similarity regions between two *Drosophila* subrepeats, 100 and 240 bp, after dropping out poly(N) runs (Fig. 4). As another example, the 330- and 200-bp rDNA IGS subrepeats of *Daphnia pulex* are very different from each other not only in size but also in sequence (Crease 1993), owing to the variability in percentage of the poly(N). We obtained a high similarity value between the two *D. pulex* subrepeats after dropping out poly(N) runs (data not shown). Similar values of sequence similarities were identified in plant species, tomato and potato rDNA IGS subrepeats.

Perhaps even more dramatically, we discovered that relationships between the subrepeats of different species could be detected using dropout alignment. Initially, we focused on sequences around 60 bp in length, because most species have small subrepeats around 60 -bp long, and also because the 60-bp subrepeat of *Xenopus* is well known for its function as a transcriptional enhancer (Reeder 1989). As would be expected, we observed a high similarity between the *Xenopus* 60-bp subrepeat and the swimming crab 60-bp subrepeat as well as between the 54-bp subrepeats of potato and the 63-bp subrepeats of tomato. Interestingly, the intraspecies similarity value between the 81- and the 100-bp subrepeats of *Xenopus* was lower than the interspecies similarity between *Xenopus* and swimming crab 60-bp subrepeats. Similarly, the degree of identity between the 54-bp subrepeats and the 74-bp subrepeats of potato is lower than that between the 54-bp subrepeats of potato and the 63-bp subrepeats of tomato. Considering that there is functional conservation of intergenic spacer elements across distantly related species (Reeder 1990), and that not all the subrepeats in rDNA IGS play the same role (Robinett et al. 1997), we hypothesized that the subrepeats which have similar functions are more conserved even in distantly related species than other subrepeats with different functions in the same species. Our results also indicate that regions that can form stem-loop structures (S1, S2, and L1 in the *Xenopus* 60-bp subrepeat) are conserved in their arrangement. Therefore, the dropout method could be useful in the systematic detection of secondary structure patterns.

One potential problem with the dropout method is that it reduces the length of nucleotide sequences by about 25%. This effect could introduce an undesirable bias, since shorter sequences have a better chance to match each other than longer sequences. Using randomized sequences selected from the same genomic sequences, we showed that the longer sequences do display lower similarity values, but not significantly. Therefore, it is reasonable to conclude that significantly increased similarity values obtained by dropout alignment reveal otherwise hidden evolutionary homology among rDNA IGS subrepeats.

The dropout alignment method is also effective in finding conserved sequences in whole IGS sequence comparisons, such as two *Drosophila* rDNA IGS sequences (Fig. 8). Although we could not find any significant conservation using unmodified rDNA IGS sequences in a Dot-Plot alignment, application of the dropout method clearly revealed many short matching sequences (Fig. 8B). Somewhat serendipitously, we also discovered another possible use for the dropout method in identifying biologically important regions through a search for matching sequences corresponding to known subrepeats. For *D. funebris*, there exist no reported subrepeats appearing in previous studies. However, we were able to reconstruct possible rDNA IGS subrepeats in *D. funebris* by using the BLAST search program coupled with the dropout method. Furthermore, the ability to construct a good multiple alignment of dropped-out IGS subrepeats from different species (Fig. 7) also suggests that single-nucleotide poly(N) runs are the primary reason for the apparent incongruities among these sequences. Thus, we propose that the reiterating nucleotides, resulting in poly(N) runs, occur at a higher rate than other types of mutations and,

thus, may have played a greater role in evolutionary changes. This hypothesis is consistent with what has been found through interspecies comparisons of other genes, such as developmental genes in different breeds of dogs (Fondon and Garner 2004). We believe that the dropout method is a useful general tool for searching for deep similarities that may be concealed by distantly or rapidly diverging sequences with fast poly(N) insertion rates. Other possible applications, for example, would include pretreatment of query and library sequences, using shrinkage of poly(N)'s, before standard BLAST searches. We are further exploring various generalizations of the dropout method to reiterating patterns at other levels, such as short reiterating dinucleotides. We thus believe that further studies with the dropout algorithm and its variants will provide us with important clues about the evolutionary mechanisms responsible for the diversification of IGS and other sequences.

References

Bach R, Allet B, Crippa M (1981) Sequence organization of the spacer in the ribosomal genes of *Xenopus clivii* and *Xenopus borealis*. Nucleic Acids Res 9:5311–5330

Baldridge GD, Fallon AM (1992) Primary structure of the ribosomal DNA intergenic spacer from the mosquito, *Aedes albopictus*. DNA Cell Biol 11:51–59

Baldridge GD, Dalton MW, Fallon AM (1992) Is higher-order structure conserved in eukaryotic ribosomal DNA intergenic spacers? J Mol Evol 35:514–523

Barker RF, Harberd NP, Jarvis MG, Flavell RB (1988) Structure and evolution of the intergenic region in a ribosomal DNA repeat unit of wheat. J Mol Biol 201:1–17

Bhatia S, Singh Negi M, Lakshmikumaran M (1996) Structural analysis of the rDNA intergenic spacer of *Brassica nigra*: evolutionary divergence of the spacers of the three diploid Brassica species. J Mol Evol 43:460–468

Black WC, McLain DK, Rai KS (1989) Patterns of variation in the rDNA cistron within and among world populations of a mosquito, *Aedes albopictus* (Skuse). Genetics 121:539–550

Borisjuk N, Hemleben V (1993) Nucleotide sequence of the potato rDNA intergenic spacer. Plant Mol Biol 21:381–384

Cordesse F, Cooke R, Tremousaygue D, Grellet F, Delseny M (1993) Fine structure and evolution of the rDNA intergenic spacer in rice and other cereals. J Mol Evol 36:369–379

Crease TJ (1993) Sequence of the intergenic spacer between the 28S and 18S rRNA-encoding genes of the crustacean, *Daphnia pulex*. Gene 134:245–249

Cross NC, Dover GA (1987) Tsetse fly rDNA: an analysis of structure and sequence. Nucleic Acids Res 15:15–30

Cunningham PR, Weitzmann CJ, Ofengand J (1991) SP6 RNA polymerase stutters when initiating from an AAA... sequence. Nucleic Acids Res 19:4669–4673

Da Rocha PS, Bertrand H (1995) Structure and comparative analysis of the rDNA intergenic spacer of *Brassica rapa*. Implications for the function and evolution of the *Cruciferae spacer*. Eur J Biochem 229:550–557

Degnan BM, Yan J, Hawkins CJ, Lavin MF (1990) rRNA genes from the lower chordate *Herdmania momus*: structural similarity with higher eukaryotes. Nucleic Acids Res 18:7063–7070

De Lucchini S, Andronico F, Nardi I (1997) Molecular structure of the rDNA intergenic spacer (IGS) in Triturus: implications for the hypervariability of rDNA loci. Chromosoma 106:315–326

Denver DR, Morris K, Kewalramani A, Harris KE, Chow A, Estes S, Lynch M, Thomas WK (2004) Abundance, distribution, and mutation rates of homopolymeric nucleotide runs in the genome of *Caenorhabditis elegans*. J Mol Evol 58:584–595

Dover GA, Tautz D (1986) Conservation and divergence in multigene families: alternatives to selection and drift. Philos Trans R Soc Lond B Biol Sci 312:275–289

Ellis RE, Sulston JE, Coulson AR (1986) The rDNA of *C. elegans*: sequence and structure. Nucleic Acids Res 14:2345–2364

Fondon JW 3rd, Garner HR (2004) Molecular origins of rapid and continuous morphological evolution. Proc Natl Acad Sci USA 101:18058–18063

Fujiwara H, Ishikawa H (1987) Structure of the *Bombyx mori* rDNA: initiation site for its transcription. Nucleic Acids Res 15:1245–1258

Ganal M, Torres R, Hemleben V (1988) Complex structure of the ribosomal DNA spacer of *Cucumis sativus* (cucumber). Mol Gen Genet 212:548–554

Gonzalez IL, Wu S, Li WM, Kuo BA, Sylvester JE (1992) Human ribosomal RNA intergenic spacer sequence. Nucleic Acids Res 20:5846

Grellet F, Delcasso D, Panabieres F, Delseny M (1986) Organization and evolution of a higher plant alphoid-like satellite DNA sequence. J Mol Biol 187:495–507

Gruendler P, Unfried I, Pointner R, Schweizer D (1989) Nucleotide sequence of the 25S–18S ribosomal gene spacer from *Arabidopsis thaliana*. Nucleic Acids Res 17:6395–6396

Higgins DG, Bleasby AJ, Fuchs R (1992) CLUSTAL V: improved software for multiple sequence alignment. Comput Appl Biosci 8:189–191

Jacques JP, Hausmann S, Kolakofsky D (1994) Paramyxovirus mRNA editing leads to G deletions as well as insertions. EMBO J 13:5496–5503

Kahl G (1988) Architecture of eukaryotic genes. VCH Verlagsgesellschaft, Weinheim, Germany/New York

Kato A, Yakura K, Tanifuji S (1984) Sequence analysis of *Vicia faba* repeated DNA, the FokI repeat element. Nucleic Acids Res 12:6415–6426

King K, Torres RA, Zentgraf U, Hemleben V (1993) Molecular evolution of the intergenic spacer in the nuclear ribosomal RNA genes of cucurbitaceae. J Mol Evol 36:144–152

Kohorn BD, Rae PM (1982) Nontranscribed spacer sequences promote in vitro transcription of Drosophila ribosomal DNA. Nucleic Acids Res 10:6879–6886

Koller HT, Frondorf KA, Maschner PD, Vaughn JC (1987) In vivo transcription from multiple spacer rRNA gene promoters during early development and evolution of the intergenic spacer in the brine shrimp Artemia. Nucleic Acids Res 15:5391–5411

Kuehn M, Arnheim N (1983) Nucleotide sequence of the genetically labile repeated elements 5′ to the origin of mouse rRNA transcription. Nucleic Acids Res 11:211–224

Kwon OY, Ishikawa H (1992) Unique structure in the intergenic and 5′ external transcribed spacer of the ribosomal RNA gene from the pea aphid *Acyrthosiphon pisum*. Eur J Biochem 206:935–940

Labhart P, Reeder RH (1984) Enhancer-like properties of the 60/81 bp elements in the ribosomal gene spacer of *Xenopus laevis*. Cell 37:285–289

Lakshmikumaran M, Negi MS (1994) Structural analysis of two length variants of the rDNA intergenic spacer from *Eruca sativa*. Plant Mol Biol 24:915–927

MacIntyre RJ (1985) Molecular evolutionary genetics. Plenum Press, New York

Mandal RK (1984) The organization and transcription of eukaryotic ribosomal RNA genes. Prog Nucleic Acid Res Mol Biol 31:115–160

Morgan GT, Middleton KM (1988) Organization and sequence of the compact rDNA spacer of the tailed frog, *Ascaphus truei*. Nucleic Acids Res 16:10917

Moss T, Boseley PG, Birnstiel ML (1980) More ribosomal spacer sequences from *Xenopus laevis*. Nucleic Acids Res 8:467–485

Murtif VL, Rae PM (1985) In vivo transcription of rDNA spacers in Drosophila. Nucleic Acids Res 13:3221–3239

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453

Ohnishi H, Yamamoto MT (2004) The structure of a single unit of ribosomal RNA gene (rDNA) including intergenic subrepeats in the Australian bulldog ant *Myrmecia croslandi* (Hymenoptera: Formicidae). Zool Sci 21:139–146

Peyretaillade E, Biderre C, Peyret P, Duffieux F, Metenier G, Gouy M, Michot B, Vivares CP (1998) Microsporidian Encephalitozoon cuniculi, a unicellular eukaryote with an unusual chromosomal dispersion of ribosomal genes and a LSU rRNA reduced to the universal core. Nucleic Acids Res 26:3513–3520

Pikaard CS, Reeder RH (1988) Sequence elements essential for function of the *Xenopus laevis* ribosomal DNA enhancers. Mol Cell Biol 8:4282–4288

Polanco C, Perez de la Vega M (1994) The structure of the rDNA intergenic spacer of *Avena sativa* L.: a comparative study. Plant Mol Biol 25:751–756

Reeder RH (1989) Regulatory elements of the generic ribosomal gene. Curr Opin Cell Biol 1:466–474

Reeder RH (1990) rRNA synthesis in the nucleolus. Trends Genet 6:390–395

Robinett CC, O'Connor A, Dunaway M (1997) The repeat organizer, a specialized insulator element within the intergenic spacer of the Xenopus rRNA genes. Mol Cell Biol 17:2866–2875

Rogers SO, Beaulieu GC, Bendich AJ (1993) Comparative studies of gene copy number. Methods Enzymol 224:243–251

Ruiz Linares A, Hancock JM, Dover GA (1991) Secondary structure constraints on the evolution of Drosophila 28 S ribosomal RNA expansion segments. J Mol Biol 219:381–390

Ryu SH, Do YK, Hwang UW, Choe CP, Kim W (1999) Ribosomal DNA intergenic spacer of the swimming crab, *Charybdis japonica*. J Mol Evol 49:806–809

Schmidt-Puchta W, Gunther I, Sanger HL (1989) Nucleotide sequence of the intergenic spacer (IGS) of the tomato ribosomal DNA. Plant Mol Biol 13:251–253

Schnare MN, Gray MW (1982) Nucleotide sequence of an exceptionally long 5.8S ribosomal RNA from *Crithidia fasciculata*. Nucleic Acids Res 10:2085–2092

Simeone A, La Volpe A, Boncinelli E (1985) Nucleotide sequence of a complete ribosomal spacer of *D. melanogaster*. Nucleic Acids Res 13:1089–1101

Sollner-Webb B, Tower J (1986) Transcription of cloned eukaryotic ribosomal RNA genes. Annu Rev Biochem 55:801–830

Sonnhammer EL, Durbin R (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:GC1–GC10

Stark GR, Debatisse M, Giulotto E, Wahl GM (1989) Recent progress in understanding mechanisms of mammalian DNA amplification. Cell 57:901–908

Suzuki A, Tanifuji S, Komeda Y, Kato A (1996) Structural and functional characterization of the intergenic spacer region of the rDNA in *Daucus carota*. Plant Cell Physiol 37:233–238

Takaiwa F, Kikuchi S, Oono K (1990) The complete nucleotide sequence of the intergenic spacer between 25S and 17S rDNAs in rice. Plant Mol Biol 15:933–935

Tautz D, Trick M, Dover GA (1986) Cryptic simplicity in DNA is a major source of genetic variation. Nature 322:652–656

Tautz D, Tautz C, Webb D, Dover GA (1987) Evolutionary divergence of promoters and spacers in the rDNA family of four Drosophila species. Implications for molecular coevolution in multigene families. J Mol Biol 195:525–542

Toth G, Gaspari Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 10:967–981

Tower J, Henderson SL, Dougherty KM, Wejksnora PJ, Sollner-Webb B (1989) An RNA polymerase I promoter located in the CHO and mouse ribosomal DNA spacers: functional analysis and factor and sequence requirements. Mol Cell Biol 9:1513–1525

Unfried K, Schiebel K, Hemleben V (1991) Subrepeats of rDNA intergenic spacer present as prominent independent satellite DNA in *Vigna radiata* but not in *Vigna angularis*. Gene 99:63–68

Wu CC, Fallon AM (1998) Analysis of a ribosomal DNA intergenic spacer region from the yellow fever mosquito, *Aedes aegypti*. Insect Mol Biol 7:19–29

Yavachev LP, Georgiev OI, Braga EA, Avdonina TA, Bogomolova AE, Zhurkin VB, Nosikov VV, Hadjiolov AA (1986) Nucleotide sequence analysis of the spacer regions flanking the rat rRNA transcription unit and identification of repetitive elements. Nucleic Acids Res 14:2799–810

Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31:3406–3415