J Mol Evol (1995) 40:629–639

JOURNAL OF MOLECULAR EVOLUTION

The Contribution of DNA Slippage to Eukaryotic Nuclear 18S rRNA Evolution

John M. Hancock*

Molecular Evolution and Systematics Group and Bioinformatics Facility, Research School of Biological Sciences, Australian National University, Canberra, ACT 0200, Australia

Received: 16 December 1993 / Accepted: 7 July 1994

Abstract. Six of 204 eukaryotic nuclear small-subunit ribosomal RNA sequences analyzed show a highly significant degree of clustering of short sequence motifs that indicates the fixation of products of replication slippage within them in their recent evolutionary history. A further 72 sequences show weaker indications of sequence repetition. Repetitive sequences in SSU rRNAs are preferentially located in variable regions and in particular in V4 and V7. The conserved region immediately 5′ to V7 (C7) is also consistently repetitive. Whereas variable regions vary in length and appear to have evolved by the fixation of slippage products, C7 shows no indication of length variation. Repetition within C7 is therefore either not a consequence of slippage or reflects very ancient slippage events. The phylogenetic distribution of sequence simplicity in small-subunit rRNAs is patchy, being largely confined to the Mammalia, Apicomplexa, Tetrahymenidae, and Trypanosomatidae. The regions of the molecule associated with sequence simplicity vary with taxonomic grouping as do the sequence motifs undergoing slippage. Comparison of rates of insertion and substitution in a lineage within the genus *Plasmodium* confirms that both rates are higher in variable regions than in conserved regions. The insertion rate in variable regions is substantially lower than the substitution rate, suggesting that selection acts more strongly on slippage products than on point mutations in these regions. Patterns of coevolution between variable regions may reflect the consequences of selection acting on the incorporation of slippage-derived sequences across the gene.

Key words: 18S rRNA evolution — Molecular coevolution — Replication slippage — Variable regions — Compensatory slippage

Introduction

Studies of the evolution of large-subunit nuclear ribosomal RNAs (LSU-rRNAs) have shown that vertebrate and rice (*Oryza sativa*) LSU-rRNAs show statistically significant levels of internal sequence repetition at the level of tri- and tetranucleotides (Hancock and Dover 1988, 1990). A number of lines of evidence indicate that sequence repetition of this kind is primarily the result of the fixation of the products of replication slippage during recent sequence evolution (reviewed in Hancock 1993). An intriguing observation that derived from these initial studies was that sequence simplicity in LSU-rRNAs, like variation in sequence length, was not uniformly distributed in phylogenetic space but appeared to be restricted to particular lineages, particularly the vertebrates (Hancock and Dover 1988, 1990). Subsequent analysis of cDNA sequences encoding the TATA-binding protein TBP (Hancock 1993) and long genomic DNA sequences (Hancock 1994a,b) has shown a similar phylogenetic distribution of simple sequences. This, and an observed correlation between levels of sequence repetition in geno-

**Present address:* MRC Clinical Sciences Centre, Royal Postgraduate Medical School, Hammersmith Hospital, London W12 0NN, UK

mic DNA sequences and SSU-rRNAs (Hancock 1994a), suggests that global levels of selection on genomes have acted to restrict the level of sequence simplicity of both genomes and their component sequences (Hancock 1994a).

Sequence repetition within LSU-rRNAs is confined to expansion segments, which are rapidly evolving regions that do not have homologues in *Escherichia coli* and other prokaryotic LSU-rRNAs (Clark et al. 1984). Dot-matrix analysis of LSU-rRNAs in species with repetitious expansion segments showed that the simple sequence motifs found in different expansion segments of any particular LSU-rRNA gene were similar (Hancock and Dover 1988, 1990) while levels of sequence repetition in expansion segments, as measured by the SIMPLE program (Tautz et al. 1986), increased with increasing expansion segment length. These data suggested a model whereby sequence repetition and expansion segment length increase resulted from the operation of slippage within these regions of the genes and the subsequent fixation of these slippage-generated sequences by molecular drive (Hancock and Dover 1988).

The function of expansion segments within LSU-rRNAs remains unresolved. Experimental replacements of *Saccharomyces cerevisiae* LSU-rRNA expansion segment V9 by a variety of analogous sequences from other species failed to disrupt ribosome function (Musters et al. 1991), but *Tetrahymena thermophila* LSU-rRNA expansion segment D2 only accommodated additional sequence at the tip of a secondary structural stem but not in its body (Sweeney and Yao 1989). This, combined with comparative analysis of secondary structures (Engberg et al. 1990), suggests that evolutionary constraints exist on expansion segment secondary structure. Analyses of the sites of incorporation of slippage-generated sequences into LSU-rRNA expansion segments were consistent with this. These showed a pattern of incorporation of simple sequences into LSU-rRNA genes that resulted in complementary simple sequences (for example, GGG and CCC) lying opposite one another in regions of the rRNA sequence that form complementary strands of secondary structural stems (compensatory slippage; Hancock and Dover 1990). This resulted in the preservation of compact secondary structures during evolution despite the incorporation of slippage-derived sequences.

Analysis of the complete ribosomal DNA (rDNA) repeat of *Drosophila melanogaster* (Tautz et al. 1988) showed that, unlike the LSU-rRNA gene, the small-subunit rRNA (SSU-rRNA) gene showed no evidence of sequence simplicity, suggesting that the fixation of simple sequences within SSU-rRNAs was subject to stronger selection than was the case for LSU-rRNAs. However, as the number of available SSU-rRNA sequences has increased, it has become clear that SSU-rRNA also shows considerable length variability. This provides the opportunity both to study the way in which selection acts to suppress the fixation of simple sequences by analyzing their distributions within SSU-rRNA molecules, and to investigate, in a much larger sample than has been available previously, the phylogenetic distribution of the readiness of genomes to accept simple sequences.

Here I describe sequence simplicity analyses of 204 eukaryotic nuclear SSU-rRNA sequences from 194 species using a modified version of the SIMPLE algorithm (SIMPLE34; Hancock and Armstrong 1994). Six of these sequences (3%) showed significantly high levels ($P < 0.003$) of overall sequence repetition, while a further 72 showed evidence of generalized or localized sequence repetition. Simple sequences in SSU-rRNAs are largely, although not completely, confined to the Mammalia and Protista, especially the Apicomplexa, Tetrahymenidae, and Trypanosomatidae. Two sites within the molecule are especially prone to contain simple motifs, the variable region V4 and a region immediately 5′ to variable region V7 (termed C7 here). These two regions have substantially different evolutionary properties. V4 is highly variable in length and sequence and appears to have undergone evolution by the fixation of products of replication slippage. C7, on the other hand, is relatively strongly conserved at the sequence level. Sequence repetition in C7 may be a molecular fossil of very ancient events in the evolution of SSU-rRNAs, or may be an emergent property of this region reflecting intramolecular interactions within the RNA and/or interactions with ribosomal proteins.

Methods

Sequences were extracted from the RDP ribosomal RNA database (Larsen et al. 1993) version 3.0 mounted on the Australian National Genomic Information System computer. The sequences analyzed are listed in Table 1. Before being subjected to sequence simplicity analysis, sequences were searched for nonconventional sequence characters (i.e., not A, C, G, or U(T)). Sequences containing more than a minimal number of such characters were eliminated from the analysis as such characters contributed to artefactually high sequence simplicity scores.

Sequence simplicity analysis was carried out using SIMPLE34 (Hancock and Armstrong 1994), a modified version of the SIMPLE program (Tautz et al. 1986). For a detailed description of the algorithm and the modifications introduced see Hancock and Armstrong (1994). Briefly, the program searches for clustering of tri- and tetranucleotide motifs within a sequence by moving a 64-bp window along it and at each position searching within the window for repeats of the tri- and tetranucleotide motifs located at its center. A simplicity score (SS) is generated for each position in the sequence which reflects the degree of repetition within the window centered on it. An overall score (simplicity factor, SF), obtained by summing values of SS for all positions within the sequence, is divided by the mean SF for ten random sequences of the same length and base and dinucleotide composition as the test sequence to generate a simplicity score for the sequence (relative simplicity factor, RSF).

The RSF is 1.000 for a sequence showing the same amount of motif clustering as the random sequences, and significantly greater than 1.000 for "simple" sequences. Statistical significance is judged based on the number of standard deviations of the SFs of the ten random

Table 1. Summary simplicity analysis of eukaryotic SSU-rRNAs

SPECIES	L[a]	%GC[b]	RSF	SIG[c]	N[d]	C[e]
ANIMALIA						
CHORDATA						
VERTEBRATA						
MAMMALIA						
Homo sapiens	1,869	56.1	1.131	+	0	+
Mus musculus	1,869	56.0	1.119	+	0	–
Oryctolagus cuniculus	1,863	55.3	1.103	+	0	–
Oryctolagus cuniculus	1,858	55.3	1.112	++	0	–
Rattus norvegicus	1,874	55.8	1.119	+	5	–
Rattus norvegicus	1,870	55.8	1.123	+	0	–
AMPHIBIA NS[f]: *Xenopus laevis;* OSTEICHTHYES NS: *Fundulus heteroclitus, Sebastolobus altivelis;* CHONDRICHTHYES NS:*Echinorhinus cookei, Squalus acanthias*						
UROCHORDATA						
Herdmania momus	1,803	50.4	1.074	–	3	
ARTHROPODA						
INSECTA NS: *Aedes albopictus, Drosophila melanogaster*						
Acyrthosiphon pisum	2,469	59.4	1.304	+++	23	++
Tenebrio molitor	1,921	51.5	1.002	–	1	
CHELICERATA NS: *Eurypelma californica;* MALACOSTRACA NS: *Artemia salinia*						
NEMATODA						
SERCENENTEA						
Caenorhabditis elegans	1,759	47.0	1.059	–	5	
Strongyloides stercoralis	1,766	38.4	1.245	+++	24	++
PLATYHELMINTHES						
TREMATODA						
Opisthorchis viverrini	1,992	51.1	1.100	–	4	
Schistosoma mansoni	1,992	49.4	1.137	–	2	
Schistosoma mansoni	1,992	49.3	1.133	–	2	
MOLLUSCA						
BIVALVIA						
Placopecten magellanicus	1,814	49.3	1.080	–	1	
CNIDARIA						
ANTHOZOA						
Anemonia sulcata	1,799	47.0	1.073	–	5	
PLANTAE						
MAGNOLIOPHYTA NS: *Oryza sativa*						
LILIOPSIDA						
Zea mays	1,809	51.0	1.127	+	2	–
MAGNOLIOPSIDA NS: *Arabidopsis thaliana, Fragaria ananassa, Glycine max, Lycopersicon esculentum, Sinapis alba*						
PINOPHYTA: CYCADOPSIDA NS: *Zamia pumila*						
FUNGI						
ZYGOMYCOTINA NS: *Endogone pisiformis, Gigaspora margarita, Glomus intraradices, Mucor racemosus*						
ASCOMYCOTINA						
HEMIASCOMYCOTINA NS: *Candida albicans, Candida glabrata, Candida krusei, Candida lusitaniae, Candida tropicalis, Kluyveromyces lactis, Saccharomyces cerevisiae*						
Debaryomyces hansenii	1,810	45.3	1.061	–	2	
Torulaspora delbrueckii	1,796	45.0	1.088	–	3	
EUASCOMYCETES NS: *Blastomyces dermatitidis, Chaetomium elatum, Colletotrichum gloeosporiodes, Cryptococcus neoformans, Eremascus albus, Leucostoma persoonii, Monascus purpureus, Neurospora crassa, Podospora anserina, Talaromyces flavus*						
Ascosphaera apis	1,736	48.0	1.016	–	1	
Aspergillus fumigatus	1,798	49.0	1.036	–	1	
Aureobasidium pullulans	1,800	48.2	1.037	–	7	
Byssochlamys nivea	1,730	49.2	1.013	–	1	
Coccidiodes immitis	1,798	48.1	1.015	–	1	
Ophiostoma stenoceras	1,732	48.9	1.051	–	5	
Pennicilium notatum	1,797	48.4	1.068	–	1	
Sporothrix schenckii	1,730	49.0	1.037	–	5	
Thermoascus crustaceus	1,718	49.1	1.025	–	1	
UNCERTAIN AFFILIATION NS: *Pneumocystis carinii, Schizosaccharomyces pombe*						

Table 1. Continued

SPECIES	L[a]	%GC[b]	RSF	SIG[c]	N[d]	C[e]
BASIDIOMYCOTINA						
USTOMYCETES						
Leucosporidium scottii	1,803	46.8	1.075	–	6	
TRUE BASIDIOMYCETES NS: *Athelia bombacina, Boletus satanas, Coprinus cinereus, Schizophyllum commune, Spongipellis unicolor, Thanatephorus praticola, Xerocomus chrysenteron*						
Cronartium ribicola	1,761	43.5	1.029	–	1	
Peridermium harknessii	1,769	43.4	1.025	–	1	
PROTISTA						
ASSEMBLAGE CHLOROBIONTS NS: *Ankistrodesmus stipitatus, Characium hindakii, Characium perforatum, Characium vacuolatum, Chlamydomonas reinhardtii, Chlorella fusca, Chlorella kessleri, Chlorella lobophora, Chlorella minutissima, Chlorella protothecoides, Chlorella saccharophila, Chlorella sorokiniana, Chlorella vulgaris, Friedmannia israeliensis, Hydrodictyon reticulatum, Neochloris aquatica, Parietochloris pseudoalveolaris, Pediastrum duplex, Prototheca zopfii, Prototheca wickerhamii, Scenedesmus obliquus, Volvox carteri*						
CHLOROPHYTA						
Chlorococcopsis minuta	1,795	49.8	1.115	+	4	–
Dunaliella parva	2,166	48.5	1.119	++	0	–
Nanochlorum eucaryotum	1,796	49.9	1.086	–	1	
ASSEMBLAGE CHROMOBIONTS						
CHRYSOPHYTA NS: *Chromulina chromophila, Hibberdia magna, Mallomonas papillosa, Mallomonas striata, Ochromonas danica, Synura spinosa*						
DINOPHYTA NS: *Emiliana huxleyi, Symbiodinium* sp, *Symbiodinium microadriaticum, Symbiodinium pilosum, Tribonema aequale*						
PHAEOPHYTA NS: *Bacillaria paxillifer, Costaria costata, Cylindrotheca closteriva, Fucus distichus, Nannochloropsis salina, Nitzchia apiculata, Rhizosolenia setigera, Stephanopyxis broschii, Skeletonema costatum*						
ASSEMBLAGE CILIATES						
CILIOPHORA NS: *Blepharisma americanum, Colpoda inflata, Euplotes aediculatus, Onychodromus quadricornutus, Oxytricha nova, Stylonychia pustulata*						
Colpidium campylum	1,754	43.6	1.147	++	0	–
Glaucoma chattoni	1,746	43.4	1.156	++	0	–
Opisthonecta henneguyi	1,730	42.9	1.023	–	1	
Oxytricha granulfera	1,778	46.0	1.080	–	5	
Paramecium tetraurelia	1,753	44.3	1.120	++	4	–
Tetrahymena australis	1,752	42.9	1.134	++	0	–
Tetrahymena borealis	1,752	43.0	1.136	+	0	–
Tetrahymena canadensis	1,752	43.0	1.136	+	0	–
Tetrahymena capricornis	1,752	42.9	1.132	++	0	–
Tetrahymena hegewischi	1,752	42.9	1.130	++	0	–
Tetrahymena hyperangularis	1,752	42.8	1.132	++	0	–
Tetrahymena mallacensis	1,753	42.8	1.147	++	0	–
Tetrahymena nanneyi	1,752	42.8	1.132	++	0	–
Tetrahymena patula	1,752	42.8	1.136	++	0	–
Tetrahymena pigmentosa	1,752	42.8	1.132	++	0	–
Tetrahymena pyriformia	1,752	42.9	1.138	+	0	–
Tetrahymena thermophila	1,753	42.8	1.143	++	2	–
Tetrahymena tropicalis	1,752	42.8	1.117	+	0	–
ASSEMBLAGE CRYPTOMONADS						
CRYPTOPHYTA						
Cryptomonas PHI	1,775	46.6	1.089	–	5	
(neucleomorph)	2,039	48.9	1.014	–	15	
Pyrenomonas salina	1,762	44.7	1.088	–	6	
ASSEMBLAGE DINOFLAGELLATES						
PERIDINEA NS: *Crypthecodinium cohnii, Prorocentrum micans*						
ASSEMBLAGE EUGLENOZOA						
EUGLENOPHYTA						
Euglena gracilis	2,305	57.0	1.005	–	1	
KLINETOPLASTIDEA NS: *Bodo caudatus, Tritrichomonas foetus, Trypanosoma brucei*						
Crithidia fasciculata	2,205	49.8	1.141	+	1	–
Endotrypanum monterogeii	2,174	49.9	1.122	+	0	–
Leishmania amazonensis	2,138	49.9	1.141	++	0	–
Leishmania brasiliensis	2,139	49.9	1.132	+	0	–

Table 1. Continued

SPECIES	L[a]	%GC[b]	RSF	SIG[c]	N[d]	C[e]
Leishmania donovani	2,204	49.7	1.117	+	0	–
Leishmania major	2,137	49.8	1.149	+	0	–
Leishmania tarentolae	2,195	49.7	1.136	+	0	–
Leptomonas sp	2,175	49.6	1.149	+	1	–
Trypanosoma cruzi	2,319	48.5	1.076	+	0	–
ASSEMBLAGE MASTIGOMYCETES						
CHYTRIDIOMYCOTA NS: *Achlya bisexualis, Blastocladiella emersonii, Chytridium confervae, Lagenidium giganteum, Neocallimastix* sp, *Phytophthora megasperma, Spizellomyces acuminatus*						
ASSEMBLAGE MICROSPORIDIA						
MICROSPORIDIA NS: *Vairimorpha necatrix*						
ASSEMBLAGE POLYMASTIGOTES NS: *Giardia Intestinalis (lamblia)*						
ASSEMBLAGE RHIZOPODS						
AMOEBOZOA NS: *Acanthamoeba castellanii, Entamoeba histolytica, Hartmanella vermiformis, Naegleria gruberi, Paratetramitus jugosus, Tetramitus rostratus, Vahlkampfia lobospinosa*						
EUMYCETOZOA NS: *Physarum polycephalum*						
Dictyostelium discoideum	1,871	42.4	1.057	–	5	
ASSEMBLAGE RHODOPHYTES						
RHODOPHYTA						
Gracilaria lemaneiformis	1,771	49.6	1.061	–	4	
Gracilaria tikvahiae	1,771	49.2	1.061	–	1	
Gracilaria verrucosa	1,771	48.8	1.038	–	1	
Gracilariopsis sp	1,782	49.8	1.099	–	5	
Porphyra umbilicalis	1,770	50.8	1.195	+	13	–
ASSEMBLAGE SPOROZOA						
APICOMPLEXA NS: *Plasmodium berghei* A, *Plasmodium berghei* C, *Plasmodium fragile* A, *Sarcocystis muris*						
Babesia bigemina (3)	1,694	45.9	1.175	+	15	–
Plasmodium falciparum A	2,090	35.6	1.166	+++	11	++
Plasmodium falciparum C	2,146	33.0	1.189	+++	8	++
Plasmodium gallinaceum A	2,102	34.6	1.174	+	18	++
Plasmodium lophurae	2,118	34.5	1.185	+++	3	++
Plasmodium malariae	2,147	34.5	1.176	+++	4	++
Theileria annulata	1,744	45.1	1.128	–	1	

[a] Sequence length

[b] Base composition expressed as percentage G+C

[c] Level of significance achieved by RSF. + reached $P < 0.05$; ++ $P < 0.01$; +++ $P < 0.003$

[d] Number of significantly simple motifs (SSMs) at 90% confidence

[e] Presence or absence of pattern of variable region coevolution on dot-matrix analysis at 19/35 stringency. ++ strong pattern; + weak pattern

[f] Sequences in which no indications of sequence simplicity were found

sequences separating the test sequence SF from 1.000. Three confidence limits, 99.7% ($P < 0.003$), 99.9% ($P < 0.01$), and 95.0% ($P < 0.05$), are returned by the program.

In addition to an overall sequence simplicity measure, SIMPLE34 also identifies sites in the sequence reaching significantly high simplicity scores by analyzing the distribution of simplicity scores in the test and random sequences and identifying scores that are 90% likely not to have occurred by chance in the test sequence. (See Hancock and Armstrong 1994, for the method of calculation of this probability). This allows identification of motifs with significantly high SSs. Such motifs are termed significantly simple motifs (SSMs). The number of SSMs within a sequence is counted and their positions are identified.

Comparisons of sequence locations of variable vs conserved regions were carried out by extracting sequences in aligned format from the RDP database. This allowed comparison of localities in different molecules using the same coordinate system. The definition of variable regions and stem numbering of Neefs et al. (1993) was used. Positions of variable and conserved regions corresponded to nucleotides 1–61 (C1), 62–86 (V1), 87–105 (C2), 106–355 (V2), 356–518 (C3), 519–584 (V3), 585–684 (C4), 685–917 (V4), 918–1,099 (C5), 1,100–1,169 (V5), 1,170–1,277 (C6), 1,278–1,322 (V6), 1,323–1,392 (C7), 1,393–1,452 (V7), 1,453–1,540 (C8), 1,541–1,594 (V8), 1,595–1,720 (C9), 1,721–1,815 (V9), and 1,816–1,870 (C10) of the human 18S rRNA sequence published by Gonzales and Schmickel (1986).

Dot-matrix analysis of coevolutionary patterns among variable regions was carried out using SIP, a modification of DIAGON (Staden 1981). Dot matrices were plotted at a stringency of 19/35 proportional match (Hancock and Dover 1988).

Results

RSF Scores of SSU-rRNA Sequences

Results of SIMPLE34 analysis of SSU-rRNA sequences are presented in Table 1. RSF scores and numbers of SSMs are presented where either the RSF of the sequence reached a likelihood (significance level) of at least $P < 0.05$ of being greater than 1 or the sequence contained at least one SSM. Sequences are organized according to the classification of Neefs et al. (1993). Six

Table 2. Numbers of species in which different triplet motifs are associated with sequence simplicity

Positions 1–3	4th position				Total
	A	C	G	T	
AAG	0	0	3	0	3
AAT	3	0	0	1	4
AGT	0	0	0	1	1
ATA	1	0	0	3	4
CAG	0	1	0	0	1
CCG	0	1	3	0	4
CCT	0	0	0	2	2
CGC	0	1	1	0	2
CGG	0	1	1	0	2
CGT	0	2	0	0	2
CTG	0	1	7	0	8
GCC	0	0	3	0	3
GCG	0	1	1	1	3
GCT	0	0	1	0	1
GGC	0	1	2	0	3
GGG	0	1	0	0	1
GGT	0	0	4	2	6
GTC	0	0	2	0	2
GTG	0	4	3	0	7
GTT	1	0	0	0	1
TAA	0	1	0	1	2
TAT	2	0	0	2	4
TCG	0	0	1	0	1
TCT	0	0	1	0	1
TGG	0	0	1	16	17
TTA	1	0	0	1	2
TTG	0	0	2	0	2
TTT	0	0	1	7	8

sequences (3%) reached an RSF that was significant at the 99.7% level—those from *Acyrthosiphon pisum* (pea aphid), *Strongyloides stercoralis* (a nematode), and three *Plasmodium species: P. falciparum* (A and C genes; see Discussion for an explanation of A and C genes in *Plasmodium* species), *P. lophurae,* and *P. malariae.* All sequences reaching this degree of significance contained at least one SSM. A further 15 sequences reached scores significant at the 99.0% level, only two of them containing SSMs. At the 95.0% level an additional 22 sequences reached a significant RSF, ten containing a SSM. Thirty-five of the sequences whose RSF was not significant at the 95% level also contained at least one SSM. Thus in total 78 sequences either had a significantly high RSF or contained at least one SSM or both. Full details of RSF values for all the sequences analyzed can be obtained from the author on request.

Frequencies of Significantly Simple Motifs

Twenty-eight tri- and 44 tetranucleotide motifs occurred as a SSM at least once in this panel of sequences. Table 2 shows the numbers of species in which each of these motifs appeared at least once. Three tetranucleotide motifs were associated with simplicity in at least five species: TGGT (16 species), CTGG and TTTT (seven species each). Table 3 shows a breakdown of the types and locations of SSMs within the SSU-rRNA sequence subdivided into conserved (C1 to C10) and variable (V1 to V9) regions. (See above.) Frequencies of occurrence of

Table 3. Motifs associated with sequence simplicity in different sequence regions of SSU-rRNAs

Species	Region: SSM(No.[a])
R. norvegicus	V4: gcgg(4); V9: ggcc
H. momus	V4: gtcg(3)
A. pisum	V2: gccg (3); V4: gtcg(3), cgtc(2), cggg, cgcg, cgcc (3), ccgc(2), gggc; V7: cggc(5), ggcg(2)
T. molitor	C7: ggtg
C. elegans	V4: ggtt; V6: tggt[b]; C7: tggt[b](3)
S. stercoralis	V4: ttat(5), tatt; V7: ttat(2), tatt(2), tatt[b](5), ttaa[b](2), aata(3), atat, ataa; C8: taat[b], ttaa[b]
O. viverrini	C7: tggt; V7: gtgc, ggtg(2)
S. mansoni	V3: aatt; C7: ggtg
P. magellanicus	C7: ggtg
A. sulcata	V4: gccg(4); C7: tggt
Z. mays	C7: tggt(2)
D. hansenii	V4: ttgg; cctt
T. delbrueckii	V4: cctt(3)
A. apis	V4: ctgg
A. fumigatus	V4: ctgg
A. pullulans	V4: ccgg(2), gccg; C7: gtgg(3); C9: cgtc
B. nivea	V4: ctgg
C. immitis	V4: ctgg
O. stenoceras	V4: ccgg(2); C7: gtgg(3)
P. notatum	V4: ctgg
S. schenckii	V4: ccgg(2); C7: gtgg(3)
T. crustaceus	V4: ctgg
L. scottii	C5: aagg; V6: tggt[b]; C7: tggt[b](3); V7: gctg
C. ribicola	C7: tggt
P. harknessii	C7: tggt
C. minuta	V4: ctgg(3), tggg
N. eucaryotum	V4: gccg
O. henneguyi	C7: tggt
O. granulifera	V2: tatt(2); C7: tggt; V8: cagc(2)
P. tetraurelia	C4: agtt[b]; V4: agtt(2); gtta
T. thermophila	V7: taac(2)
C. PHI	C4: tcgg[b]; V4: tcgg[b](4)
C. PHI (NM)	V2: tttt(12); V4: tctg(3)
P. salina	C4: tcgg[b]; V4: tcgg[b](5)
E. gracilis	V4: ctgc
C. fasciculata	V4: gtgc
Leptomonas sp	V4: gtgc
D. discoideum	C3: aata[b]; V3: aata[b](2); C7: tggt(2)
G. lemaneiformis	V6: tggt[b]; C7: tggt(3)
G. tikvahiae	C7: tggt
G. verrucosa	C7: tggt
Gracilariopsis sp	V4: gtgc: V6: tggt[b]; C7: tggt(3)
P. umbilicalis	V4: gcgc(5), gcgt, ggcg(3); V6: tggt[b]; C7: tggt(3)
B. bigemina	V4: tttt(11); ggtt; ttgg; tggt; C10: aagg
P. falciparum A	V4: tttt(5), tttg(5); V8: atat
P. falciparum B	V2: ataa(4); V7: tttt(4)
P. gallinaceum A	V7: tttt(8); V8: atat(2), tata(8)
P. lophurae	V7: tttt; V8: tata(2)
P. malariae	V7: tttt(4)
T. annulata	C10: aagg

[a] Numbers in parentheses are the numbers (greater than one) of each motif identified

[b] Motifs from arrays that overlap region boundaries

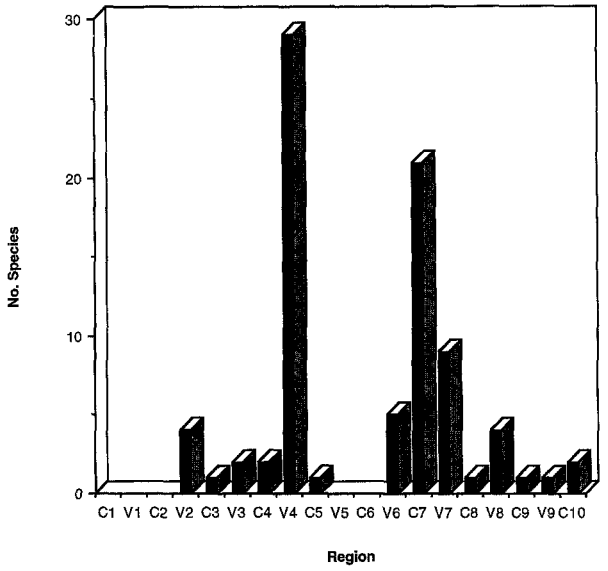

Fig. 1. Histogram showing for each region within the SSU-rRNA gene the number of species containing at least one significantly simple motif (SSM) as calculated by SIMPLE34 using the criterion that the score obtained by a motif must have a 90% chance or less of not having occurred by chance.

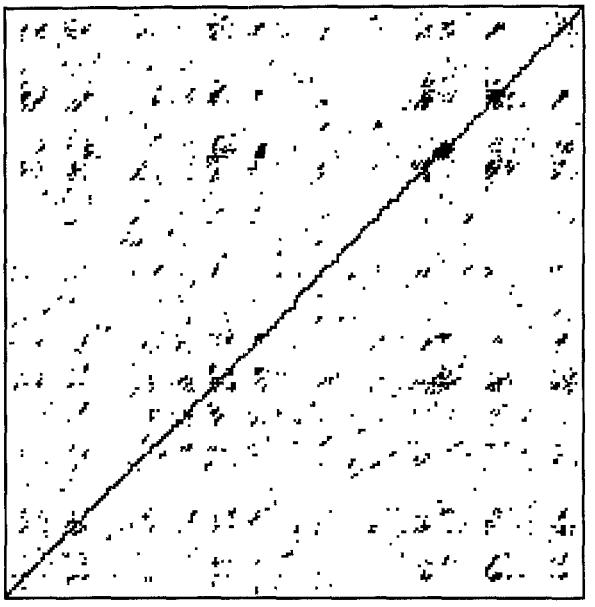

Fig. 2. Dot-matrix self-comparison of the *P. gallinaceum* SSU-rRNA sequence. Generated using the SIP program at a stringency of 19 matches out of 35 (proportional match) (see Methods).

significant motifs in different segments of the molecule are displayed graphically in Fig. 1. Three regions within the molecule (variable regions V4 and V7 and conserved region C7) contained 86% of occurrences of distinct SSMs.

Variable-Region Coevolution

Dot-matrix analysis was carried out for all sequences whose RSFs reached 95% or higher significance. Of these 43 sequences, eight showed visible patterns of sequence similarity between variable regions at intermediate stringency, corresponding to variable-region coevolution (Hancock and Dover 1988). Six of these were those reaching the $P < 0.003$ level of significance, while the other two were *Plasmodium gallinaceum* (RSF = 1.175, $P < 0.05$) and *Homo sapiens* (RSF = 1.131, $P < 0.05$). The pattern in *H. sapiens* was weak. Figure 2 illustrates the pattern in the *P. gallinaceum* gene.

Relationship Between Sequence Base Composition and RSF

Table 4 shows the distribution of RSF score significance compared to the degree of sequence base compositional bias in 201 SSU-rRNA sequences (excluding duplicates). The χ^2 value for this contingency table is highly significant ($\chi^2 = 124.74$, $df = 12$, $P < 0.001$), i.e., the distribution of score significance is nonrandom with respect to base compositional bias. Median base compositional biases for the four significance classes were 2.3% for $P > 0.05$, 4.1% for $P < 0.05$, 7.1% for $P < 0.01$, and 14.95% for $P < 0.003$. All sequences with RSFs significant at the $P < 0.003$ level had base compositional biases $\geq 9.4\%$.

Table 4. Frequencies of sequences with different base compositions having RSFs reaching different levels of significance

Base comp. bias	>0.05	<0.05	<0.01	<0.003	Total
20.1–25	2	0	0	0	2
15.1–20	0	1	0	3	4
10.1–15	5	0	0	2	7
5.1–10	29	9	13	1	52
0.1–5	122	12	2	0	136
Total	158	22	15	6	201

Discussion

These analyses show that a small proportion of SSU-rRNA sequences contains a significant amount of internal repetition in a way similar to some LSU-rRNA genes, while a larger proportion contains lower levels of internal repetition. These repetitive regions are concentrated in regions that show high sequence variability (Neefs et al. 1993) in both SSU-rRNAs and LSU-rRNAs. Despite this similarity, simple sequences in eukaryotic SSU-rRNAs show a much more restricted distribution than they do in LSU-rRNAs (Hancock and Dover 1988, 1990), reflecting the more stringent selection acting on SSU-rRNAs (Tautz et al. 1988). Only three regions, variable regions V4 and V7 and conserved region C7, account for 86% of the occurrences of SSMs in SSU-rRNAs, again reflecting more stringent selection on simple sequences in SSU-rRNAs than in LSU-rRNAs.

Comparison of the evolution of these sequence regions shows them to be very different, leading to different conclusions about the origin of sequence simplicity in the different regions.

Variable-region V4 is highly variable in sequence, so despite the substantial database now available it has not been possible to construct an unambiguous secondary-structure model for part of it (Gutell 1993). This region therefore seems likely to have undergone periods of expansion by the incorporation of slippage-generated sequences in a similar way to some eukaryotic LSU-rRNA expansion segments. The identification of very different SSMs in V4 in different groups of organisms (for example, GTCG/CGTC/CGGG/CGCG/CGCC/CCGC/GGGC in *Acyrthosiphon pisum* compared to TTTT/TTTG in the *Plasmodium falciparum* A gene) suggests a number of independent slippage/fixation events in its evolutionary history in different evolutionary lineages, resulting in nonhomology of at least part of it in different taxa. A similar evolutionary scenario appears to apply to V7, which contains the SSM TTTT in *Plasmodium* species and CGGC/GGCG in *A. pisum.* Other variable regions within SSU-rRNA, especially those at the 5′ end of the molecule, appear to be more refractory to the fixation of simple sequences generated by slippage, although in four species variable-region V2 contains SSMs.

The second-most-common site for detection of SSMs in SSU-rRNAs was in the conserved region C7 immediately 5′ to variable region V7. This had a higher frequency of SSMs than V7 itself. C7 makes up parts of four secondary structural elements, stems 37–40, that are highly conserved in evolution, being present in eubacteria, archaebacteria, and eukaryotes. Sequence repetition in C7 is based on the motif TGGT and the two related motifs GTGG and GGTG (Table 3). Comparison of C7 between the analyzed sequences using aligned output from the RDP database showed that few C7 regions differ in length, and those that do do so by single base insertion/deletions. It therefore seems unlikely that the sequence simplicity detected in this part of SSU-rRNAs reflects the recent action of slippage.

If C7 has not incorporated products of slippage at least since the divergence of eukaryotes, and possibly longer, the sequence simplicity detected in this region must have other origins. One possible explanation is that it is a molecular fossil of a very ancient series of slippage events that generated stems 38–40. These events would have predated the genesis of the three taxonomic domains. Since that time, these stems may have adopted an important function in the ribosome, eliminating the possibility of incorporating further slippage-generated sequences or point mutations that would erase the original pattern of repetition. C7 may then have been originally part of a larger variable region encompassing V6, C7, and V7 as well as part of C8. While C7 remained evolutionarily stable, the neighboring V6 region continued to diverge after the separation of eubacteria from archaebacteria and eukaryotes, resulting in the observed divergence in secondary structural pattern between the two groupings (Neefs et al. 1993). V7 has preserved a compact secondary structure throughout evolution, but has been enlarged by the incorporation of slippage-derived sequences into its single stem (JMH, in preparation).

An alternative explanation for the repetitiveness of C7 is that it is an essential feature of the region that evolved by point mutation from a nonrepetitive precursor. Its repetition may then reflect some functional aspect of this region of the rRNA, such as action as a conformational switch. It might also reflect intra- or intermolecular interactions, involving RNA and/or proteins, in which it participates. C7 may then represent an example of sequence simplicity being an emergent property of a sequence undergoing point mutation.

Factors Affecting the Fixation of Simple Sequences in SSU-rRNAs

Significant sequence simplicity ($P < 0.05$) is highly unevenly distributed between different taxonomic groups, with high concentrations in the eutherian mammals and in certain protist groups. This patchy distribution is consistent with previous observations on the distribution of sequence simplicity in LSU-rRNA, where only vertebrates and *O. sativa* showed elevated simplicity (Hancock and Dover 1988, 1990), and in the TATA-binding protein TBP and long genomic sequences, which show a similar pattern of phylogenetic distribution to SSU- and LSU-rRNAs (Hancock 1993, 1994a). If simple sequences in these molecules have arisen predominantly from the action of slippage, as seems likely, this points to an uneven pattern of incorporation of slippage-derived sequences in different lineages during evolution. This could have resulted either from a propensity of genomes in particular lineages at particular times to undergo slippage or from different degrees of selection acting on slippage-derived sequences in different lineages at different times or a combination of the two. However, analyses of the overall level of sequence simplicity of long genomic sequences (Hancock 1994a), which have shown significant correlations between genomic sequence simplicity and both genome size and SSU-rRNA RSF, indicate that the phylogenetic pattern of simple sequence content of SSU-rRNA genes is part of a pattern of change of genome-wide selection pressure against the incorporation of simple sequences. This may be related to the evolution of genome size (Hancock 1994a,b; and see Cavalier-Smith 1985).

Cases in which distinct SSU-rRNA sequences derive from single species allow analysis of the extent to which selection on simple sequences within SSU-rRNAs is genome specific. Two examples occur in this dataset: the nuclear and nucleomorph sequences from *Cryptomonas* PHI and the asexual and sexual sequences from *Plasmodium falciparum* and *Plasmodium berghei.*

The two different SSU-rRNA genes to be found in *Plasmodium* species, known as A and C genes, are expressed at different stages in the organisms' life cycles (Dame et al. 1984). *Plasmodium* species can be classified into two groups depending on whether or not their SSU-rRNAs are repetitive. While *P. berghei* shows low sequence simplicity and no SSMs in either of its SSU-rRNA genes, *P. falciparum* has highly repetitive genes that differ in the distribution of simple sequence motifs: SSMs in the *P. falciparum* A gene are confined mostly to variable region V4, whereas in its C gene variable regions V2 and V7 contain simple sequences but V4 does not. Although they are located at different sites, simple sequences in the two *P. falciparum* genes are based on A/T-rich tracts, consistent with the predominance of A + T-rich simple sequences in eukaryotic genomes (Hancock 1994b).

rRNA genes are often tandemly arranged and undergo concerted evolution by unequal crossing-over and gene conversion (Dover 1982). The lack of concerted evolution indicated by the different distributions of simple sequences in these two genes may therefore reflect their nontandem arrangement (Unnasch and Wirth 1983). The differences in types and locations of simple sequences in the two genes may reflect differential selective pressures on the two genes or chance seeding of slippage at different sites in the two gene lineages. However, the sharing of elevated levels of sequence simplicity between the differently expressed genes in *P. falciparum* is consistent with the proposition that the tendency to contain simple sequences is common to the organism as a whole rather than being a feature of individual genes.

In *Cryptomonas,* the nuclear and nucleomorph sequences differ in both the number and distribution of SSMs. The nuclear sequence shows a similar pattern of distribution of motifs to that of *Pyrenomonas salina,* its closest relative in this dataset, with SSMs lying straddling the boundary of variable region V4, while the nucleomorph gene contains a simple sequence array in variable region V2 as well as V4 and is much longer (2,039 compared to 1,775 nucleotides). As the nucleomorph of *Cryptomonas* may be an eukaryotic endosymbiont with a residual genome (see Cavalier-Smith 1993 and references therein), this appears to reflect lower selective pressure on the nucleomorph than on the nuclear gene.

As well as different levels of selection acting at the genome level, the onset of slippage in a gene lineage might be affected by the level of base compositional bias (i.e., the difference between its base composition and 50%) in the gene. Genes with more biased base compositions would be inherently likely to contain a higher frequency of clustered motifs that could act as substrates for slippage (Hancock and Dover 1990; Dover and Tautz 1986). The SSU-rRNA dataset generally supports this (Results). The main exceptions are the three sequences from *Giardia intestinalis* (*G. lamblia,* two genes) and *Vairimorpha necatrix.* SSU-rRNAs from both of these species exhibit extreme short length and base compositional bias (Boothroyd et al. 1987; Vossbrinck et al. 1987), which appears to be related to their very early divergence from other eukaryotes (Sogin et al. 1989; but see Leipe et al. 1993). Their short lengths, which are less than that of *E. coli* 16S rRNA, appear to reflect extreme selective pressure that would preclude length increase by the incorporation of slippage-derived sequences. Indeed, it is possible that such extremely biased base compositions could only evolve in SSU-rRNAs under extreme selective pressure on length, as these genes might otherwise have been subject to catastrophic expansion.

As well as providing insight into the role of genome-wide selection in influencing levels of sequence simplicity in SSU-rRNAs, the genus *Plasmodium* also provides the best opportunity to investigate the emergence of simple sequences within SSU-rRNAs. This is because, of the 11 genera represented by more than one species in this dataset, *Plasmodium* is the only genus whose SSU-rRNAs contain simple sequences and that shows a higher variation in RSF scores (as measured by standard deviation) than the higher-order taxon of which it is a part, the Apicomplexa. This higher level of variation suggests recent accumulation of sequence simplicity within the genus. Phylogenetic analysis of *Plasmodium* based on SSU-rRNA sequences divides the genus into "avian" and "mammalian" groups (Waters et al. 1991, 1993). Members of the avian group contain simple sequences while members of the mammalian group do not. As the other apicomplexan sequences in the database contain little or no sequence simplicity (*T. annulata, S. muris*), or contain simple sequences in different positions than those in *Plasmodium* species (*B. bigemina*), this partition suggests that simple sequences in *Plasmodium* SSU-rRNAs originated after the separation of the avian and the mammalian groups. This was presumably about 300 million years before present, the approximate date of divergence of the bird and mammal lineages (Nei 1987). The presence in the high-simplicity group of *P. falciparum* and *P. malariae,* which infect humans, is consistent with the suggestion of lateral transfer from the avian to the human lineage (Waters et al. 1991).

Rates of substitution and insertion within SSU-rRNA-variable regions are not trivial to estimate in the presence of replication slippage because nucleotides that are opposed in alignments may not be related by descent. To some extent this problem can be ameliorated by using secondary-structure information to aid alignment (Van de Peer et al. 1993). However, secondary structures may also be misleading where selection acts to constrain the incorporation of simple sequences to complementary strands of stems (Hancock and Dover 1990). Because of this, estimates of insertion and substitution rate are best made in closely related species where the homology of individual sites is least questionable. *P. gallinaceum* and *P. lophurae* represent the best opportunity for such analysis among the *Plasmodium* species as they parasitize

two bird species (*Gallus sonnerattii* and *Lophura ignitii*, respectively; see Waters et al. 1993) which diverged relatively recently (approximately 8×10^6 years ago; Kornegay et al. 1993).

Differences between the *P. gallinaceum* and *P. lophurae* SSU-rRNA sequences comprise 22 insertion/deletions (five in conserved regions and 17 in variable regions) and 177 substitutions (five in conserved regions and 112 in variable regions). Assuming that all insertions/deletions represent insertions, and using Kimura and Ohta's (1972) method for calculating genetic distance, these differences provide crude estimates of insertion rates of 2.6×10^{-10} insertions site^{-1} year^{-1} in conserved regions and 9.1×10^{-10} insertions site^{-1} year^{-1} in variable regions. This latter rate is similar to rates estimated for expansion segments of *Drosophila* LSU-rRNAs (Ruiz Linares et al. 1991; Rousset et al. 1991; Hancock et al. 1988). Estimated mean substitution rates are 6.5×10^{-10} site^{-1} year^{-1} in conserved regions and 1.2×10^{-8} site^{-1} year^{-1} in variable regions. Substitution rates are therefore higher than insertion rates in both conserved and variable regions in this lineage. The mean substitution rate in variable regions, which appears to be much higher than rates of insertion, approaches the neutral substitution rate in *Drosophila* (Sharp and Li 1989) and may therefore be close to the mutation rate in *Plasmodium*. By contrast the insertion rate due to slippage remains well below microsatellite mutation rates (Dallas 1992). This difference in rates is observed even in variable regions undergoing little replication slippage (data not shown), suggesting that substitution rates are not confounded significantly by the effects of replication slippage in this comparison. Therefore, in *Plasmodium* variable regions at least, the products of replication slippage appear to be under more selection than are point mutations. This may reflect selection to preserve secondary structure (Hancock and Dover 1990).

Although most SSU-rRNA molecules do not show evidence of recent slippagelike processes, all the sequences reaching a significance level of $P < 0.003$ showed a pattern of sequence similarity between variable regions comparable to that seen between expansion segments in vertebrate LSU-rRNAs (Hancock and Dover 1988, 1990) (Fig. 2). Such a pattern of expansion segment/variable region coevolution is therefore a common feature of rRNA genes that contain repetitive variable regions/expansion segments. Coevolutionary patterns potentially reflect functional coevolution, slippage acting on common substrate motifs initially present in different parts of the same molecule, or micro-gene conversion between variable regions (Hancock and Dover 1988). However if selection on slippage-derived mutations is higher than on point mutations in variable regions (see above), and as the incorporation of simple sequences is constrained by secondary structural considerations (Hancock and Dover 1990), patterns of coevolution may have functional significance.

References

Boothroyd JC, Wang A, Campbell DA, Wang CC (1987) An unusually compact ribosomal DNA repeat in the protozoan *Giardia lamblia*. Nucleic Acids Res 15:4065–4084

Cavalier-Smith T (1985) Introduction: the evolutionary significance of genome size. In: Cavalier-Smith T (ed) The evolution of genome size. John Wiley, New York, pp 1–36

Cavalier-Smith T (1993) Evolution of the eukaryotic genome. In: Broda PMA, Oliver SG, Sims PFG (eds) Society for general microbiology symposium 50: the eukaryotic genome: organization and regulation. Cambridge University Press, Cambridge, pp 333–385

Clark CG, Tague BW, Ware VC, Gerbi SA (1984) *Xenopus laevis* 28S ribosomal RNA: a secondary structure model and functional implications. Nucleic Acids Res 12:6197–6220

Dallas JF (1992) Estimation of microsatellite mutation rates in recombinant inbred strains of mouse. Mammal Genome 3:452–456

Dame JB, Sullivan M, McCutchan TF (1984) Two major sequence classes of ribosomal RNA genes in *Plasmodium berghei*. Nucleic Acids Res 12:5943–5952

Dover GA (1982) Molecular drive: a cohesive mode of species evolution. Nature 299:111–117

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

Engberg J, Nielsen H, Lenaers G, Murayama O, Fujitani H, Higashinakagawa T (1990) Comparison of primary and secondary 26S rRNA structures in two *Tetrahymena* species: evidence for a strong evolutionary and structural constraint in expansion segments. J Mol Evol 30:514–521

Gonzalez IL, Schmickel RD (1986) The human 18S ribosomal RNA gene: evolution and stability. Am J Hum Genet 38:419–427

Gutell RR (1993) Collection of small subunit (16S- and 16S-like) ribosomal RNA structures. Nucleic Acids Res 21:3051–3054

Hancock JM (1993) Evolution of sequence repetition and gene duplications in the TATA-binding protein TBP (TFIID). Nucleic Acids Res 21:2823–2830

Hancock JM (1994a) Genomic tectonics: slippage-driven tectonic processes in genome evolution. Submitted

Hancock JM (1994b) Consistent patterns of sequence bias in slippage-derived sequences: their origins and consequences for gene and genome evolution. Submitted

Hancock JM, Armstrong JS (1994) SIMPLE34: an improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences. Comput Appl Biosci 10:67–70

Hancock JM, Dover GA (1988) Molecular coevolution among cryptically simple expansion segments of eukaryotic 26S/28S rRNAs. Mol Biol Evol 5:377–391

Hancock JM, Dover GA (1990) 'Compensatory slippage' in the evolution of ribosomal RNA genes. Nucleic Acids Res 18:5949–5954

Hancock JM, Tautz D, Dover GA (1988) Evolution of the secondary structures and compensatory mutations of the ribosomal RNAs of *Drosophila melanogaster*. Mol Biol Evol 5:393–414

Kimura M, Ohta T (1972) On the stochastic model for estimation of mutational distance between homologous proteins. J Mol Evol 2: 87–90

Kornegay JR, Kocher TD, Williams LA, Wilson AC (1993) Pathways of lysozyme evolution inferred from the sequences of cytochrome b in birds. J Mol Evol 37:367–379

Leipe DD, Gunderson JH, Nerad TA, Sogin ML (1993) Small subunit ribosomal RNA+ of *Hexamita inflata* and the quest for the first branch in the eukaryotic tree. Mol Biochem Parasitol 59:41–48

Larsen N, Olsen GJ, Maidak BL, McCaughey MJ, Overbeek R, Macke TJ, Woese CR (1993) The ribosomal database project. Nucleic Acids Res 21:3021–3023

Musters W, Gonçalves PM, Boon K, Raué HA, van Heerikhuizen H, Planta RJ (1991) The conserved GTPase center and variable region V9 from *Saccharomyces cerevisiae* 26S rRNA can be replaced by their equivalents from other prokaryotes or eukaryotes without detectable loss of ribosomal function. Proc Natl Acad Sci USA 88: 1469–1473

Neefs J-M, Van de Peer Y, De Rijk P, Chapelle S, De Wachter R (1993) Compilation of small ribosomal subunit RNA structures. Nucleic Acids Res 21:3025–3049

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Rousset F, Pélandakis M, Solignac M (1991) Evolution of compensatory substitutions through G.U intermediate state in *Drosophila* rRNA. Proc Natl Acad Sci USA 88:10032–10036

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

Sharp PM, Li W-H (1989) On the rate of DNA sequence evolution in *Drosophila*. J Mol Evol 28:398–402

Sogin ML, Gunderson JH, Elwood HJ, Alonso RA, Peattie DA (1989) Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from *Giardia lamblia*. Science 243:75–77

Staden R (1982) An interactive graphic program for comparing and aligning nucleic acid and amino acid sequences. Nucleic Acids Res 10:2951–2961

Sweeney R, Yao M-C (1989) Identifying functional regions of rRNA by insertion mutagenesis and complete gene replacement in *Tetrahymena thermophila*. EMBO J 8:933–938

Tautz D, Hancock JM, Webb DA, Tautz C, Dover GA (1988) Complete sequences of the rRNA genes of *Drosophila melanogaster*. Mol Biol Evol 5:366–376

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

Unnasch TR, Wirth DF (1983) The avian malaria *Plasmodium lophurae* has a small number of heterogeneous ribosomal RNA genes. Nucleic Acids Res 11:8443–8459

Van de Peer Y, Neefs J-M, De Rijk P, De Wachter R (1993) Reconstructing evolution from eukaryotic small-ribosomal-subunit RNA sequences: calibration of the molecular clock. J Mol Evol 37:221–232

Vossbrinck CR, Maddox JV, Friedman S, Debrunner-Vossbrinck BA, Woese CA (1987) Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes. Nature 326:411–414

Waters AP, Higgins DG, McCutchan TF (1991) *Plasmodium falciparum* appears to have arisen as a result of lateral transfer between avian and human hosts. Proc Natl Acad Sci USA 88:3140–3144

Waters AP, Higgins DG, McCutchan TF (1993) Evolutionary relatedness of some primate models of *Plasmodium*. Mol Biol Evol 10: 914–923